# **Generative Adversarial Networks**

# Dougal J. Sutherland Gatsby, UCL $\rightarrow$ TTIC $\rightarrow$ UBC



(from thispersondoesnotexist.com)

MLCC 2019

## **Generative models**

- Start with a bunch of examples:  $X_1, \ldots, X_n \sim \mathbb{P}$
- Want a model for the data:  $\mathbb{Q} \approx \mathbb{P}$

# **Generative models**

- Start with a bunch of examples:  $X_1, \ldots, X_n \sim \mathbb{P}$
- Want a model for the data:  $\mathbb{Q} \approx \mathbb{P}$
- Might want to do different things with the model:
  - Find most representative data points / modes
  - Find outliers, anomalies, ...
  - Discover underlying structure of the data
  - Impute missing values
  - Use as prior (semi-supervised, machine translation, ...)
  - Produce "more samples"



# **Generative models**

- Start with a bunch of examples:  $X_1, \ldots, X_n \sim \mathbb{P}$
- Want a model for the data:  $\mathbb{Q} \approx \mathbb{P}$
- Might want to do different things with the model:
  - Find most representative data points / modes
  - Find outliers, anomalies, ...
  - Discover underlying structure of the data
  - Impute missing values
  - Use as prior (semi-supervised, machine translation, ...)
  - Produce "more samples"



### Why produce samples?



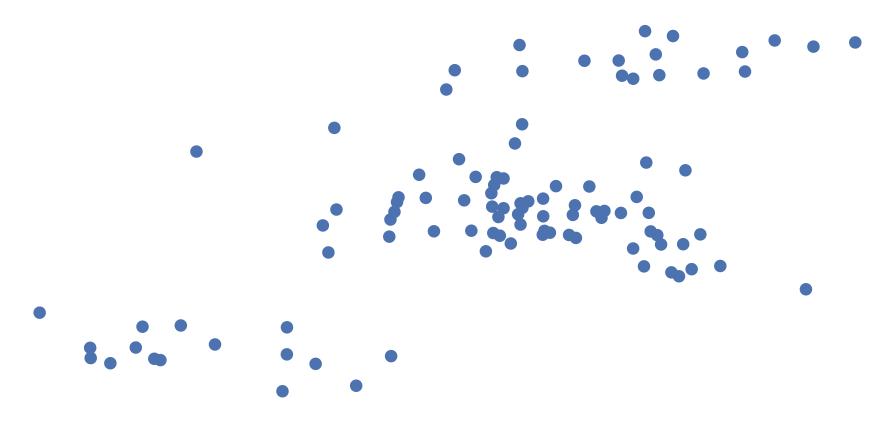
# Is artificial intelligence set to become art's next medium?

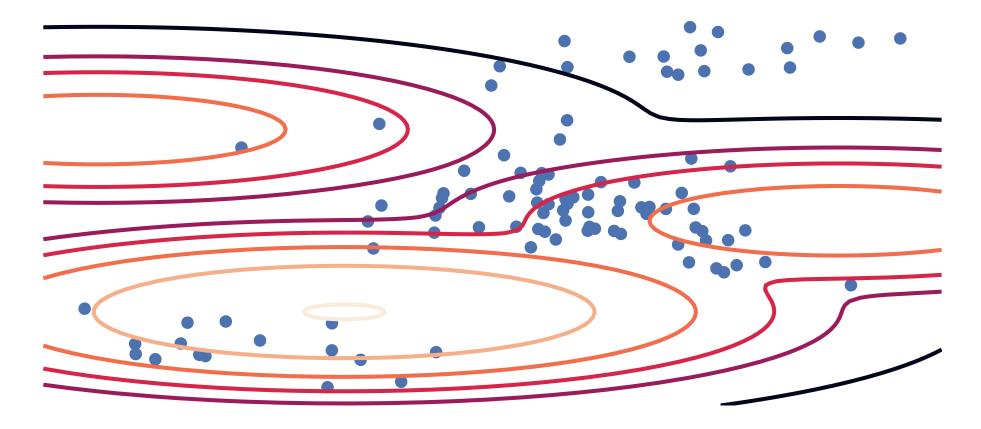
Al artwork sells for \$432,500 — nearly 45 times its high estimate — as Christie's becomes the first auction house to offer a work of art created by an algorithm

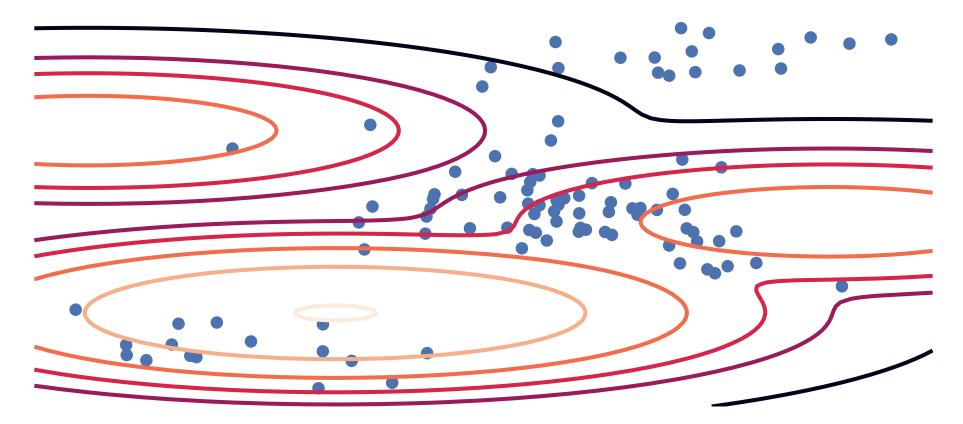
The portrait in its gilt frame depicts a portly gentleman, possibly French and — to judge by his dark frockcoat and plain white collar — a man of the church. The work appears unfinished: the facial features are somewhat indistinct and there are blank areas of canvas. Oddly, the whole composition is displaced slightly to the north-west. A label on the wall states that the sitter is a man named Edmond Belamy, but the giveaway clue as to the origins of the work is the artist's signature at the bottom right. In cursive Gallic script it reads:

 $\min_{G} \max_{D} \mathbb{E}_{x}[\log(D(x))] + \mathbb{E}_{z}[\log(1 - D(G(z)))]$ 

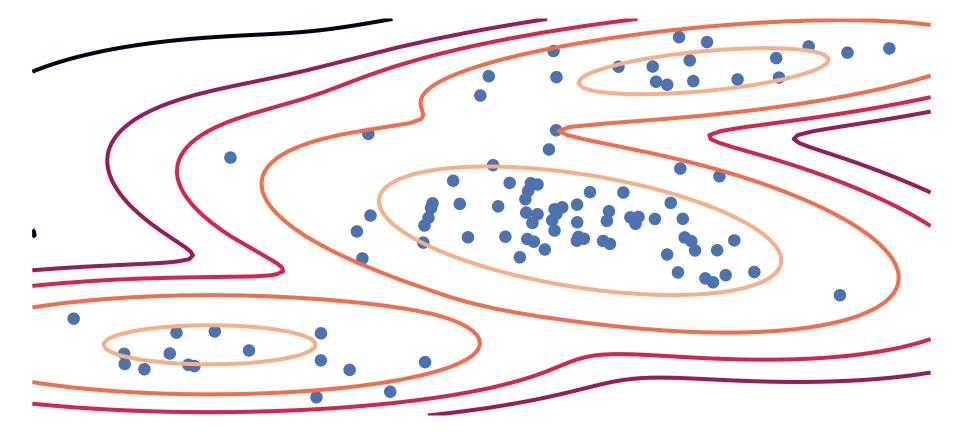
Image © Obvious



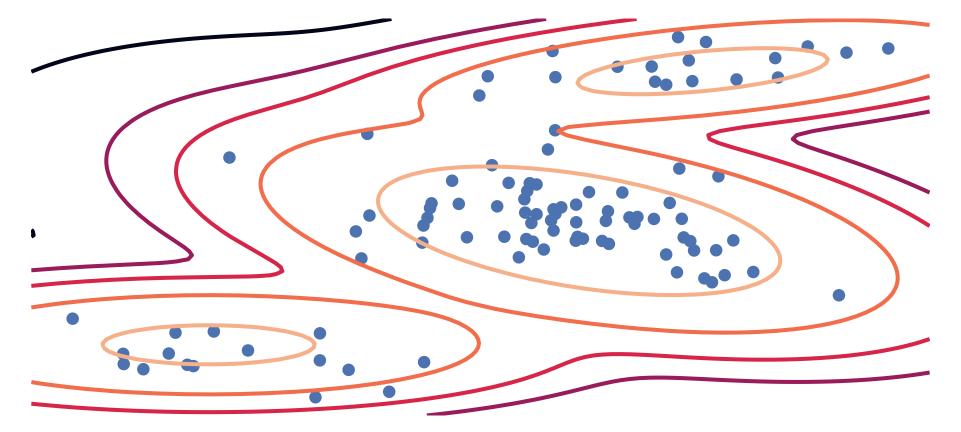




• Maximum likelihood:  $\max_{ heta} \mathbb{E}_{X \sim \mathbb{P}}[\log q_{ heta}(X)]$ 



• Maximum likelihood:  $\max_{\theta} \mathbb{E}_{X \sim \mathbb{P}}[\log q_{\theta}(X)]$ 



- Maximum likelihood:  $\max_{ heta} \mathbb{E}_{X \sim \mathbb{P}}[\log q_{ heta}(X)]$
- Equivalent:  $\min_{\theta} \operatorname{KL}(\mathbb{P} \| \mathbb{Q}_{\theta}) = \min_{\theta} \int p(x) \log \frac{p(x)}{q_{\theta}(x)} \mathrm{d}x$

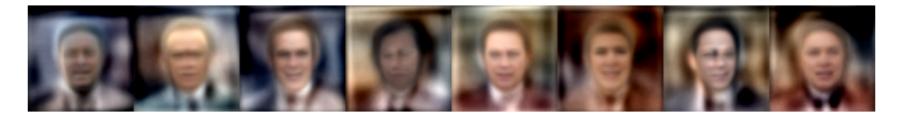
# Traditional models for images

• 1987-style generative model of faces (Eigenface via Alex Egg)



# Traditional models for images

• 1987-style generative model of faces (Eigenface via Alex Egg)



• Can do fancier versions, of course...

# Traditional models for images

• 1987-style generative model of faces (Eigenface via Alex Egg)



- Can do fancier versions, of course...
- Usually based on Gaussian noise  $pprox L_2$  loss

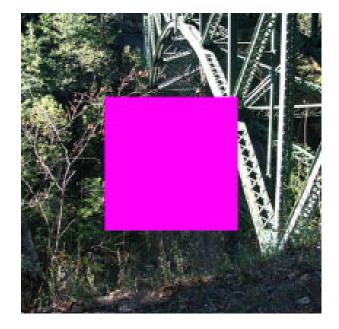
• One use case of generative models is inpainting [Harry Yang]:



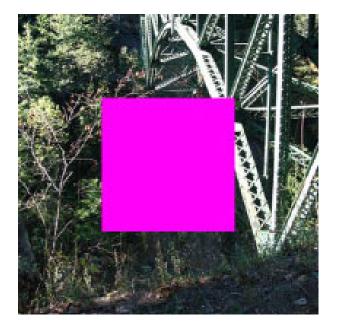


• One use case of generative models is inpainting [Harry Yang]:

• One use case of generative models is inpainting [Harry Yang]:



• One use case of generative models is inpainting [Harry Yang]:



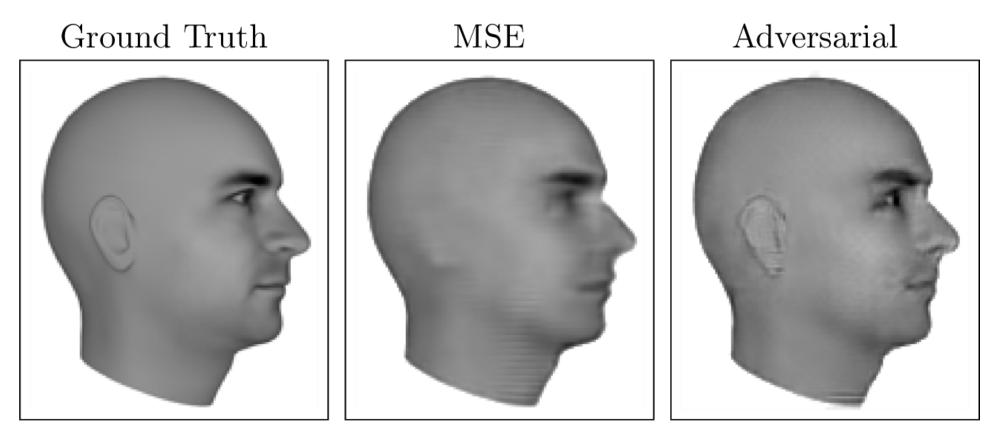
•  $L_2$  loss / Gaussians will pick the *mean* of possibilities

• One use case of generative models is inpainting [Harry Yang]:



•  $L_2$  loss / Gaussians will pick the *mean* of possibilities

### **Next-frame video prediction**



#### [Lotter+ 2016]

#### Generator ( $\mathbb{Q}_{\theta}$ )



#### Discriminator



#### Generator ( $\mathbb{Q}_{\theta}$ )



#### Discriminator



Target (P)





#### Discriminator





Target (P)







No way!  $\Pr(\text{real}) = 0.03$ 

Target (P)











No way!  $\Pr(\text{real}) = 0.03$ 

Target (P)











No way! Pr(real) = 0.03

Target (P)





Is this real?









No way! Pr(real) = 0.03

Target (P)





Is this real?





No way! Pr(real) = 0.03

 $\text{Umm...} \Pr(\text{real}) = 0.48$ 

• MLCC so far: models  $f(x) = w^\mathsf{T} \Phi(x) + b$ ,  $f: \mathcal{X} o \mathbb{R}$ 

• MLCC so far: models  $f(x) = w^\mathsf{T} \Phi(x) + b$ ,  $f: \mathcal{X} o \mathbb{R}$ 

$$egin{aligned} &f_0(x)=x\ &f_\ell(x)=\sigma_\ell\left(W_\ell f_{\ell-1}(x)+b_\ell
ight),f_\ell:\mathbb{R}^{d_{\ell-1}} o\mathbb{R}^{d_\ell} \end{aligned}$$

• MLCC so far: models  $f(x) = w^\mathsf{T} \Phi(x) + b$ ,  $f: \mathcal{X} o \mathbb{R}$ 

$$egin{aligned} &f_0(x)=x\ &f_\ell(x)=\sigma_\ell\left(W_\ell f_{\ell-1}(x)+b_\ell
ight), f_\ell:\mathbb{R}^{d_{\ell-1}} o\mathbb{R}^{d_\ell} \end{aligned}$$

•  $\sigma_{\ell}$  is an *activation function*:

$$\max(x,0) \qquad rac{1}{1+e^{-x}} \qquad rac{e^x-e^{-x}}{e^x+e^{-x}} \qquad \cdots$$

• MLCC so far: models  $f(x) = w^{\mathsf{T}} \Phi(x) + b$ ,  $f: \mathcal{X} o \mathbb{R}$ 

$$egin{aligned} &f_0(x)=x\ &f_\ell(x)=\sigma_\ell\left(W_\ell f_{\ell-1}(x)+b_\ell
ight), f_\ell:\mathbb{R}^{d_{\ell-1}} o\mathbb{R}^{d_\ell} \end{aligned}$$

- $\sigma_\ell$  is an activation function:  $\max(x,0) \quad rac{1}{1+e^{-x}} \quad rac{e^x-e^{-x}}{e^x+e^{-x}} \quad \cdots$
- Classification usually uses log loss (cross-entropy):

 $\ell(y,t) = 1(t=1)\log f(x) + 1(t=0)\log(1-f(x))$ 

• MLCC so far: models  $f(x) = w^{\mathsf{T}} \Phi(x) + b$ ,  $f: \mathcal{X} o \mathbb{R}$ 

$$egin{aligned} &f_0(x)=x\ &f_\ell(x)=\sigma_\ell\left(W_\ell f_{\ell-1}(x)+b_\ell
ight), f_\ell:\mathbb{R}^{d_{\ell-1}} o\mathbb{R}^{d_\ell} \end{aligned}$$

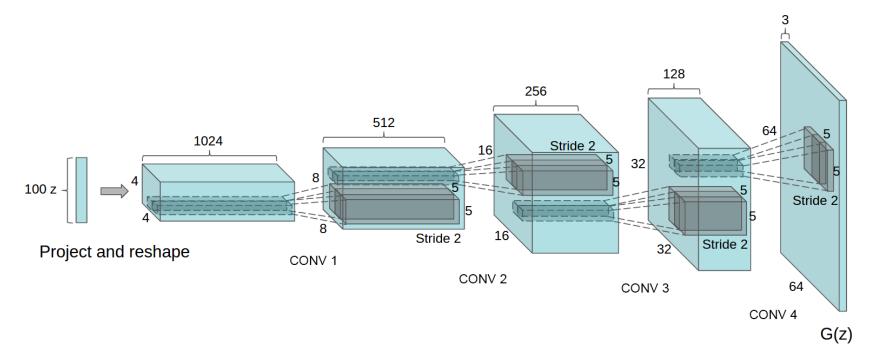
- $\sigma_\ell$  is an activation function:  $\max(x,0) \qquad rac{1}{1+e^{-x}} \qquad rac{e^x-e^{-x}}{e^x+e^{-x}} \qquad \cdot$
- Classification usually uses log loss (cross-entropy):

 $\ell(y,t) = 1(t=1)\log f(x) + 1(t=0)\log(1-f(x))$ 

• Optimize with gradient descent

### **Generator networks**

• How to specify  $\mathbb{Q}_{\theta}$ ?



#### [Radford+ ICLR-16]

- $Z \sim \mathbb{Z} = \text{Uniform}\left([-1,1]^{100}
  ight)$
- $G_{m heta}: [-1,1]^{100} o \mathcal{X}$  ,  $G_{m heta}(Z) \sim \mathbb{Q}_{m heta}$

# **GANs in equations**

• Tricking the discriminator:

$$\min_{\theta} \max_{\psi} \frac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}} [\log D_{\psi}(X)] + \frac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{\theta}} [\log(1 - D_{\psi}(Y))]$$

## **GANs in equations**

• Tricking the discriminator:

$$\min_{\theta} \max_{\psi} \frac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}} [\log D_{\psi}(X)] + \frac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{\theta}} [\log(1 - D_{\psi}(Y))]$$

• Using the generator network for  $\mathbb{Q}_{\theta}$ :

$$\min_{oldsymbol{ heta}} \max_{\psi} rac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\log D_{\psi}(X)] + rac{1}{2} \mathop{\mathbb{E}}_{Z \sim \mathbb{Z}}[\log(1 - D_{\psi}(G_{oldsymbol{ heta}}(Z))))$$

## **GANs in equations**

• Tricking the discriminator:

$$\min_{\theta} \max_{\psi} \frac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}} [\log D_{\psi}(X)] + \frac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{\theta}} [\log(1 - D_{\psi}(Y))]$$

• Using the generator network for  $\mathbb{Q}_{\theta}$ :

$$\min_{oldsymbol{ heta}} \max_{\psi} rac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\log D_{\psi}(X)] + rac{1}{2} \mathop{\mathbb{E}}_{Z \sim \mathbb{Z}}[\log(1 - D_{\psi}(G_{oldsymbol{ heta}}(Z)))]$$

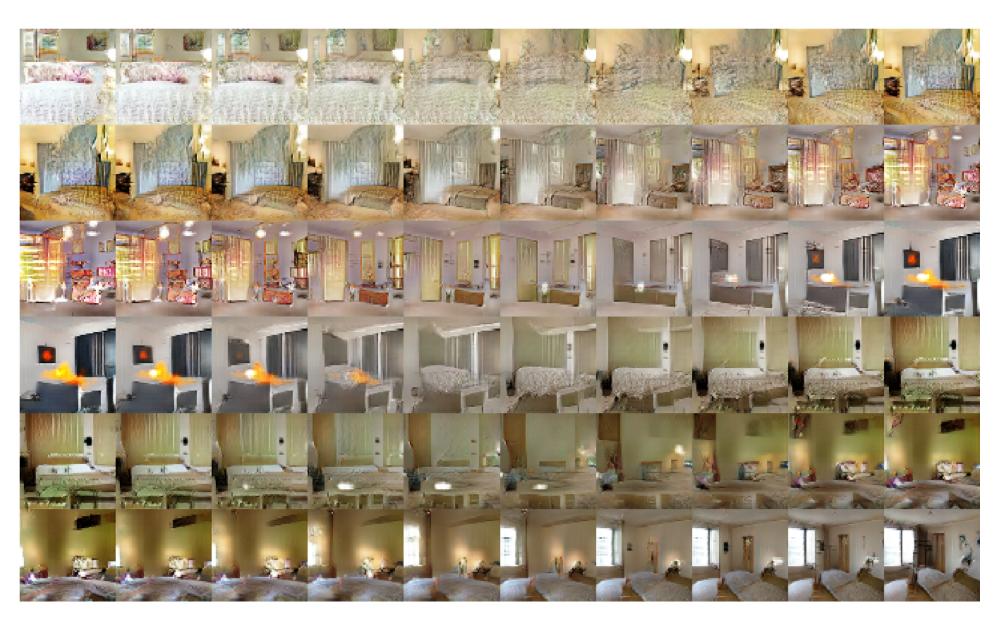
• Can do alternating gradient descent!

#### Original paper's results [Goodfellow+ NeurIPS-14]





#### DCGAN results [Radford+ ICLR-16]



Running code from [Salimans+ NeurIPS-16]:

14 C 1 C 1 C	Contraction of	10000	100000	10000	10000	CONTRACTOR OF	CONTRACTOR OF		States and States
222222	1000000		1203020	1202020	100000-00-				
12121	100	1000	Statements.	Color Service Service	Contraction of the	1000	1000	100	1996
	100000	*******		121212	1000000		100000000	10000000	1000000
inte feite feite feite	100000	59595951	1151515	59595955	19696951	1464646	1000000	59595951	59595951
	100000	100000000	10000	Contraction of the	10000	Contractory of the	100000		100000
one space space									
<b>MARKED</b>	Statistics.	CHARGE STREET,	Subman.	Charles Charles	Statistics.	Contraction of the local division of the loc	1464646	100000000	COLUMN TWO IS NOT
1222222		122222	CONTRACTOR OF	1000000	100000	100000	1000000	121212	121212
		365	222222		200	1000	120202	1999	12222
			Sector of the se		Contraction of the	and the second s	Sector set	Concerned to	Sectors:
1999	1999990	14444	1000	222222	10202	1000	1000	1996	19999
and the state of the	10000000	100000000	10000000	1000000	10000000	Statistics."	ALC: NO POST	And the state of the	And the Advertise of the
114141	12020	izizia.	20202	122222	1262625	122222	1202022	12626251	1262625
	Sector of the	Sectors:	Concerned of the local distribution of the l	Sectors:	Concerned of			Concerned in	
885		12222	and the second s	200		and Martin	1222222	1202020	1222222
	10000000	10000000	100000000	10000000	10000000	And the second second	And the second second	1999	100000
222222	10000	161616	1000	202022	100001	20202	101010	1222222	1252525
	Contraction of the local distance of the loc	Contraction of the	Colorado Carlo Colorado	State of the local diversion of the local div	Contraction of the local division of the loc	the local data in the	Contraction of the local distribution of the	one water water to	take bade bade to
a service s	A DESCRIPTION OF	A DESCRIPTION OF	A DESCRIPTION OF	A DESCRIPTION OF THE OWNER.	A DESCRIPTION OF THE OWNER.	of the local division of the local divisiono	a track the second	A DESCRIPTION OF	a logiture in the
100000	1000	1000000	1069695	111111	10000	1141411	1060600	1000	1000
	State State State St.	1000	100000	And Property lies	And Address of the		100000	And in case of the local division of the loc	And Property in
	200	100000	12222	100,00,00	129.29.20	1999		100000	1202020
onto party party limit	sale and the sale	sale will belle the	Call and a Co	table balls findly family	sale with large line	only party party in	outs with both	sale with fails	sale with their sale
29295	25252	1000	262621	1000	-		10202	222	2022
title had a lost a lost	tota letta letta luri	TALL AND ADD TO	the local data in the	tota loda loda tor	tions have been to	tions from the local diversion of	tota betta betta tor	ting body body in	TAXABLE AND A
A DEPARTMENT OF	10000000000	COLUMN T	distantion of the	10000000000	A REAL PROPERTY AND A REAL PROPERTY.	And Address of the owner owner owner owner own	100000000000000000000000000000000000000	A STATE OF STATE	A DEPENDENCE OF
101010	20202	122222	1000	22222	22222	12222	1202020	122222	202020
	10000000	1000000	1000000	10000000	100000000	and the second second	10000000	10000000	And the second second
	22222	1000	122222	10000	292925	200		202020	1000
and the local division of the	14444	1996	19993	1000	1000	State of the local division of the	Statements for	1000	14441
	1000000	1000000	1000000	Contraction of the local distribution of the	1000000	100000	10000000	10000000	10000000
and a stand of the stand of the	and a local diversion of	and a starting starting store		and the state of the	and the set of the set	THE REAL PROPERTY.	the state of the second	and the stands share to be	and a stand a stand a stand
NAME:	112121	Sugara and	Sugara and	1000000	STATISTICS.	<b>MANNE</b>	1000000	STATISTICS.	THE R. LEWIS CO.
						120202-		1292921	and the
inter Serte Serte Serte	Statements in	Cardia Cardia Cardia Cardia	Statements of	Same and	100000	1000	10000000	STATE:	10000
222		122224	*******		222	100	10000000	-000	
10 10 10 10 10	1989895	ininini .	2007	14141	1000005	199999	19696951	10000000	1000
		1000	10000		1000	and the second	and the last	100000000	and the second second
	1000000	CHARGE CO.	10000000	1000000	10000000		10000000	10000	10000
1222244	12222	12222	121211	121212	10000	121211	121212	101010	122224
5959555	59595955	59595955	59595955	59595955	59595955	110101	10606061	1000000	122222
1000		Contraction of the local distribution of the		Contraction of the local division of the loc	-	and the second se			and the second s
1000000	10000	1000	14444	1262625	1262625	169610	1000	10161	1000
Statistics of the second	and the state of	Standards.	Constraints of	Standard Street	and the state of	Standard Street	Standage States	STREET, STREET	COLUMN TWO IS NOT
11010		14646461	1144444	111111	1000000	100000	1000000	100000	120202
Contract of the	Concession of the local division of the loca	10000	Contraction of the local division of the loc	Concession of the local division of the loca			Concernance of the local diversion of the loc	Concession of the	Concession of the local division of the loca
1919	100000	14444	2222222	1202020	10000	1262625	1222222	1202020	1262625
and the	A DESCRIPTION OF THE OWNER.	THE REAL PROPERTY.	100000	And the state of the	A DECISION OF A DECISIONO OF A D	10000	A REAL PROPERTY.	100000	100000
222222	1999	2222222	100.00.00	1000	10000	10,20,20	1202025	222222	10.00.00
the local division in which the	tide hade being the	tide here bet a	Table Sold Suffrage Color	tide herts herts to	only living the local	tide and a local division	tide hads been the	only field here a	TAXABLE AND A
a sea sea sea -	A Property line .	A REAL PROPERTY.	A Real Property lies of	A REAL PROPERTY.	A REAL PROPERTY.	A REAL PROPERTY.	A PROPERTY AND	A Real Property lies, in	A Real Property lies.
	1969696	19696961	1969696	120202-0	1969695	1969696	19696961	19696961	1202020
200	200		200	and the second se				and the second se	a second s
A SA SA SA	1222222	12626261	3292923	1262625	120202	120202	1000	12626251	1262625
	10000	1000	CONTRACTOR OF	10000	CONTRACTOR OF STREET, ST.	Contraction of the local division of the loc	100000	CONTRACTOR OF	CONTRACTOR OF STREET, ST.
102020	1222223	1222222	12222.22	122223	19393	123333	1282822	123 23 23	1292923
take further family family	take first a local data	take fasts fasts for	take bade bade ton	take bady bady for	only loss in the	and is not in factor that	other Real of Contractory	take fight fight for	and it lights fights family
a second second	and the second second	and the second	a state state of	and the second second	A REAL PROPERTY.		and the second second	a state state of	a traction to a local state
19696969	1969695	19696951	1909095	1969695	1969696	1202020	1202020	1000000	1202000
	STATISTICS.				10000			1000	
	122222	3222222	1202020	1202020	1222222	1222222	1222222	1202020	12222220
Non Property in	THERE IS NOT	Concentration of the local division of the l	THERE IS NOT	STREET.	COLUMN T	COLUMN TWO IS NOT	CONTRACTOR OF STREET,	CONTRACTOR OF STREET, ST	COLUMN TWO IS NOT
3999	102021	1222322	1222511	1222512	102020	12222	26262	1000010	1000
title fields fields for	Total Anda Anda Ar	Total Anda Anda Col	TAXABLE AND A	Cold Serie Serie Cold	TAXABLE AND A	tella fiella fiella fert	ting the local of the	Testa Antis Antis Art	lines from the local
		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1							

Running code from [Salimans+ NeurIPS-16]:

	-	-	-	-		-	-	-	
	-		-	-	-	-	-	-	Contraction of the
-	-	-	-	-	-	-	-	-	-
Contractor of the	10000000	1000000000	1000001	100000000	-	-	-	10000000000	100000000
	100000000	-	1000001	100000000	THE OWNER OF	-	-	10000000000	100000000
-	1000001	1000001	040401	2000001	10000001	-	10000001	2000001	10000001
and the second s				100000	-		10000000		
	The second s	100000000	-	100000000	10404041	-	-	100000000	10404040
	-		-	200000	2000001	-	-	-	1000000
-	200001	2000001	-	000001	0808051		-	-	10000001
and a second second	100000000	100000000	10000000			100000	10000000	1000000000	1000000000
	1000000		1000000	-	20000000	000000	100000000	-	10000001
	COLORED IN COLORED	-	CHOICE !!	-	-	000000	10000000	-	000000
Contraction of	1000000	-	1000001		-	0404040	1000000	-	2000001
and a second sec	1000000000		100000000	1000000000	100000000	100000000	1000000000	1000000000	CONTRACTOR OF STREET, ST
100000000	10000000		10000000	-	20000000	-	-	10000000	20000000
1000000000	10000000	2000000	1000000	-	-	-	-	1000000	20000000
	2000000	0404051		-		-	200000	2000000	0000000
	100000000	10000000000						10000000000	
	-		1000000	-		-	-	-	
	-		1000000	-	2000000	-	1000000	-	000000
	0404040		100000		000000	000000	1000000		
Contraction of the									
	-	-	-	1000000	000000	-	-	-	
	-		COLORIDA C	100000	CHORES !!	-	Concession of the local division of the loca	Concession of the local division of the loca	COLORIDAC:
000000				040401	000000	000001		000001	
Cardina de Cardon de		Contraction of the		-	Contraction of the	100000000	-		Cardenand Cardenand
	-		Concerned in the	080808	-	-	Concession of the local division of the loca	-	-
	-	-	-	-	-	-	1000	-	-
		-			2000001		3404041	280805	
and a second second	- Contraction of the	Cardena de la composition de l	Contraction of the			Contraction of the			Contraction of the local division of the loc
	-		Concession of the local division of the loca	Contraction of the local division of the loc	Concession of the local division of the loca	Statement of the local division of the local	Statistics.	Statement of	Contraction of the local division of the loc
	and the second s	Concession of the local division of the loca	Concession of the local division of the loca	Concession of the local division of the loca	Contraction of the local division of the loc	Statement of the local division of the local	Statistics.	Statistics.	Contraction of the local division of the loc
Contraction of the local division of the loc		Constant of		Concession of	Conception of	Concession.	Second Second	Conception of	
and a second sec	and the second s		Cardena and Cardena			Contraction of the	Contraction of the	- Contraction of the	
Contraction of the	Concession.	Concession of the local division of the loca	Concession of the local division of the loca	Concession of the local division of the loca	Concession of the local division of the loca	Concession of the local division of the loca	Concession of the local division of the loca	Concession of the local division of the loca	Contraction of the local division of the loc
-	Concession in the local division in the loca	-	Contraction of the local division of the loc	and the second second	Contraction of the local division of the loc	-	and the second second	Concession of the local division of the loca	Contraction of the local division of the loc
Contraction of the local division of the loc	COMPANY OF L	Concession of the local division of the loca	Concession of the local division of the loca	Contraction of the local division of the loc	-	-	Conception of the local division of the loca	Concession of the local division of the loca	Contraction of the local division of the loc

Running code from [Salimans+ NeurIPS-16]:

17	7th	2	naș	ŦŰ,	The	de	1	12	the
reci,	14	The	110		the	44	nde,		44
140	τά,	74	44	til.	4	44	114	10	114
100		14	ile	-4	ú4	-44		4	
								12	
14	Ύμ	49	14		23	111	#	4	4
								44	
TA	-	τų		44	eq.	LUL.	44	7%	44
110 100 100 100						and the second second		44	

Running code from [Salimans+ NeurIPS-16]:



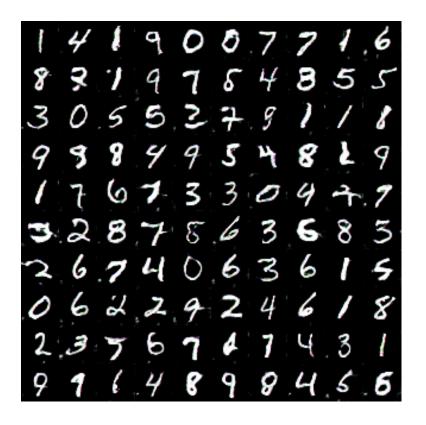
Running code from [Salimans+ NeurIPS-16]:



Running code from [Salimans+ NeurIPS-16]:



Running code from [Salimans+ NeurIPS-16]:



Running code from [Salimans+ NeurIPS-16]:

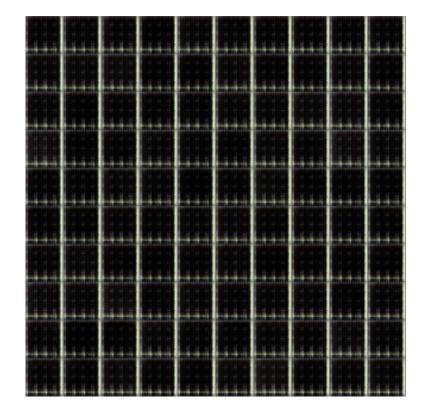


Running code from [Salimans+ NeurIPS-16]:



Running code from [Salimans+ NeurIPS-16]:

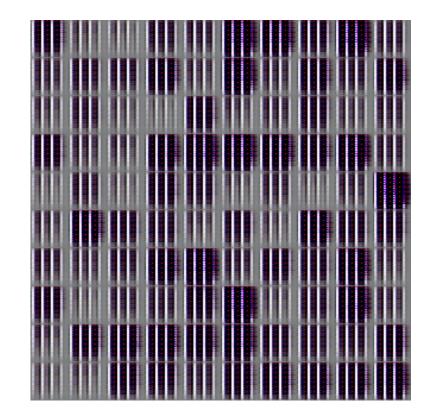




Run 1, epoch 900

Running code from [Salimans+ NeurIPS-16]:

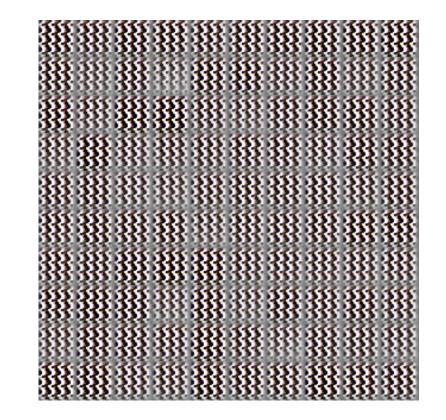




Run 1, epoch 900

Running code from [Salimans+ NeurIPS-16]:

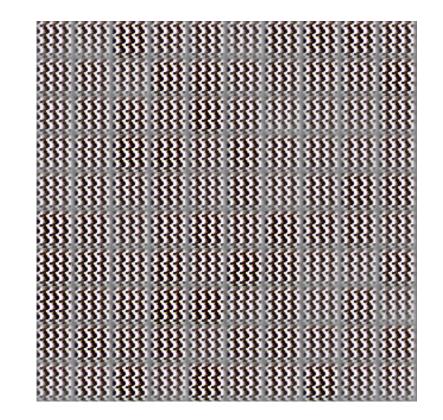




Run 1, epoch 900

Running code from [Salimans+ NeurIPS-16]:

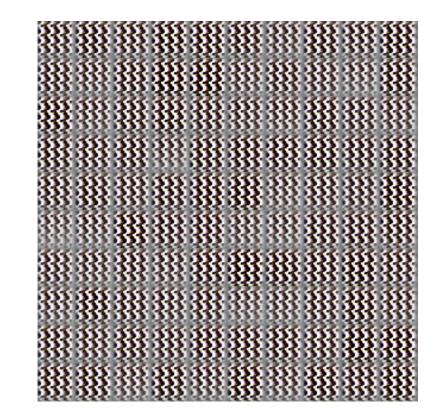




Run 1, epoch 900

Running code from [Salimans+ NeurIPS-16]:





Run 1, epoch 900

## One view: distances between distributions

• What happens when  $D_\psi$  is at its optimum?

#### **One view: distances between distributions**

- What happens when  $D_\psi$  is at its optimum?
- If distributions have densities,  $D^*_\psi(x) = rac{p(x)}{p(x)+q_ heta(x)}$

#### **One view: distances between distributions**

- What happens when  $D_\psi$  is at its optimum?
- If distributions have densities,  $D^*_\psi(x) = rac{p(x)}{p(x)+q_ heta(x)}$
- If  $D_\psi$  stays optimal throughout, heta tries to minimize

$$rac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}} igg[ \log rac{p(X)}{p(X) + q_{ heta}(X)} igg] + rac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{ heta}} igg[ \log rac{q_{ heta}(X)}{p(X) + q_{ heta}(X)} igg]$$

which is  $\mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) - \log 2$ 

#### Jensen-Shannon divergence

$$egin{aligned} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) &= rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \ &+ rac{1}{2}\int q_{ heta}(x)\lograc{q_{ heta}(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \end{aligned}$$

#### Jensen-Shannon divergence

$$egin{aligned} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) &= rac{1}{2} \int p(x) \log rac{p(x)}{rac{1}{2} p(x) + rac{1}{2} q_{ heta}(x)} \mathrm{d}x \ &+ rac{1}{2} \int q_{ heta}(x) \log rac{q_{ heta}(x)}{rac{1}{2} p(x) + rac{1}{2} q_{ heta}(x)} \mathrm{d}x \ &= rac{1}{2} \mathrm{KL}igg(\mathbb{P} \,igg\| rac{\mathbb{P} + \mathbb{Q}_{ heta}}{2}igg) + rac{1}{2} \mathrm{KL}igg(\mathbb{Q}_{ heta} \,igg\| rac{\mathbb{P} + \mathbb{Q}_{ heta}}{2}igg) \end{aligned}$$

#### Jensen-Shannon divergence

$$\begin{split} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{\theta}) &= \frac{1}{2} \int p(x) \log \frac{p(x)}{\frac{1}{2}p(x) + \frac{1}{2}q_{\theta}(x)} \mathrm{d}x \\ &+ \frac{1}{2} \int q_{\theta}(x) \log \frac{q_{\theta}(x)}{\frac{1}{2}p(x) + \frac{1}{2}q_{\theta}(x)} \mathrm{d}x \\ &= \frac{1}{2} \mathrm{KL} \left( \mathbb{P} \left\| \frac{\mathbb{P} + \mathbb{Q}_{\theta}}{2} \right) + \frac{1}{2} \mathrm{KL} \left( \mathbb{Q}_{\theta} \right\| \frac{\mathbb{P} + \mathbb{Q}_{\theta}}{2} \right) \\ &= \mathrm{H} \left[ \frac{\mathbb{P} + \mathbb{Q}_{\theta}}{2} \right] - \frac{\mathrm{H}[\mathbb{P}] + \mathrm{H}[\mathbb{Q}_{\theta}]}{2} \end{split}$$

$$egin{aligned} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) &= rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \ &+ rac{1}{2}\int q_{ heta}(x)\lograc{q_{ heta}(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \end{aligned}$$

• If  $\mathbb{P}$  and  $\mathbb{Q}_{\theta}$  have (almost) disjoint support

$$rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)}\mathrm{d}x$$

$$egin{aligned} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) &= rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \ &+ rac{1}{2}\int q_{ heta}(x)\lograc{q_{ heta}(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \end{aligned}$$

• If  $\mathbb{P}$  and  $\mathbb{Q}_{\theta}$  have (almost) disjoint support

$$rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)}\mathrm{d}x=rac{1}{2}\int p(x)\log(2)\mathrm{d}x$$

$$egin{aligned} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) &= rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \ &+ rac{1}{2}\int q_{ heta}(x)\lograc{q_{ heta}(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \end{aligned}$$

• If  $\mathbb{P}$  and  $\mathbb{Q}_{\theta}$  have (almost) disjoint support

$$rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)}\mathrm{d}x = rac{1}{2}\int p(x)\log(2)\mathrm{d}x = rac{1}{2}\log 2$$

$$egin{aligned} \mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) &= rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \ &+ rac{1}{2}\int q_{ heta}(x)\lograc{q_{ heta}(x)}{rac{1}{2}p(x)+rac{1}{2}q_{ heta}(x)}\mathrm{d}x \end{aligned}$$

• If  $\mathbb{P}$  and  $\mathbb{Q}_{\theta}$  have (almost) disjoint support

$$rac{1}{2}\int p(x)\lograc{p(x)}{rac{1}{2}p(x)}\mathrm{d}x=rac{1}{2}\int p(x)\log(2)\mathrm{d}x=rac{1}{2}\log 2$$

so  $\mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) = \log 2$ 

Generator ( $\mathbb{Q}_{\theta}$ )



#### Discriminator

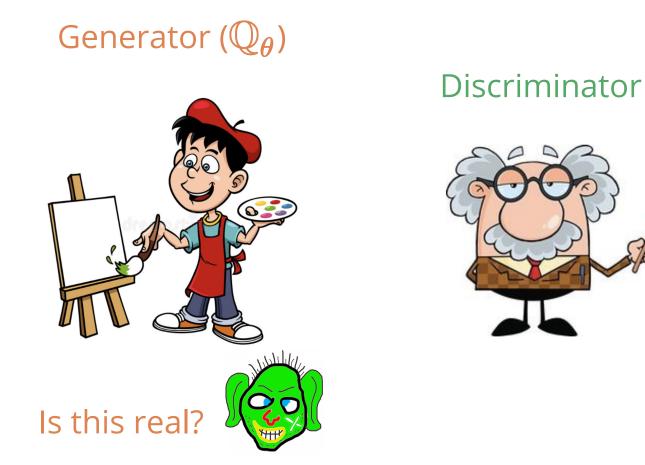


Generator ( $\mathbb{Q}_{\theta}$ )



#### Discriminator







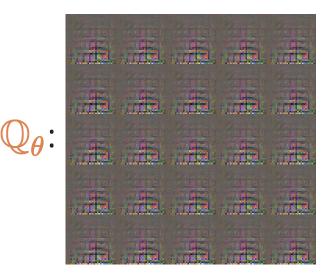
Target ( $\mathbb{P}$ )





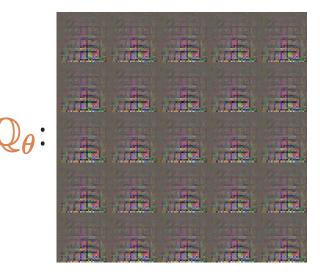
:(I don't know how to do any better...





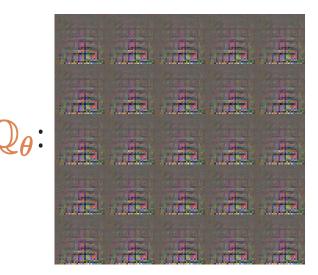
• At initialization, pretty reasonable:





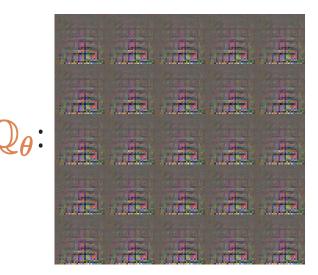
• Remember we might have  $G_ heta: \mathbb{R}^{100} o \mathbb{R}^{64 imes 64 imes 3}$ 





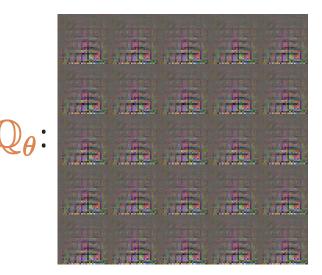
- Remember we might have  $G_ heta: \mathbb{R}^{100} o \mathbb{R}^{64 imes 64 imes 3}$
- For usual  $G_ heta$ ,  $\mathbb{Q}_ heta$  is supported on a countable union of manifolds with dim  $\leq 100$





- Remember we might have  $G_ heta: \mathbb{R}^{100} o \mathbb{R}^{64 imes 64 imes 3}$
- For usual  $G_ heta$ ,  $\mathbb{Q}_ heta$  is supported on a countable union of manifolds with dim  $\leq 100$
- "Natural image manifold" usually considered low-dim





- Remember we might have  $G_ heta: \mathbb{R}^{100} o \mathbb{R}^{64 imes 64 imes 3}$
- For usual  $G_ heta$ ,  $\mathbb{Q}_ heta$  is supported on a countable union of manifolds with dim  $\leq 100$
- "Natural image manifold" usually considered low-dim
- No chance that they'd align at init, so  $\mathrm{JS}(\mathbb{P},\mathbb{Q}_{ heta}) = \log 2$

# A heuristic partial workaround

• Original GANs almost never use the minimax game

 $\min_{\theta} \max_{\psi} \frac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}} [\log D_{\psi}(X)] + \frac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{\theta}} [\log(1 - D_{\psi}(Y))]$ 

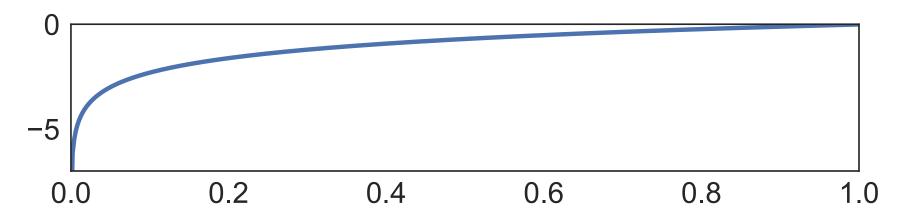
•  $\max_{\theta} \log D_{\psi}(G_{\theta}(Z))$ , not  $\min_{\theta} \log(1 - D_{\psi}(G_{\theta}(Z)))$ 

# A heuristic partial workaround

• Original GANs almost never use the minimax game

$$\min_{oldsymbol{ heta}} \max_{\psi} rac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\log D_{\psi}(X)] + rac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{oldsymbol{ heta}}}[\log(1 - D_{\psi}(Y))]$$

- $\max_{\theta} \log D_{\psi}(G_{\theta}(Z))$ , not  $\min_{\theta} \log(1 D_{\psi}(G_{\theta}(Z)))$
- If  $D_\psi$  is near-perfect, near  $\log 0$  instead of  $\log 1$

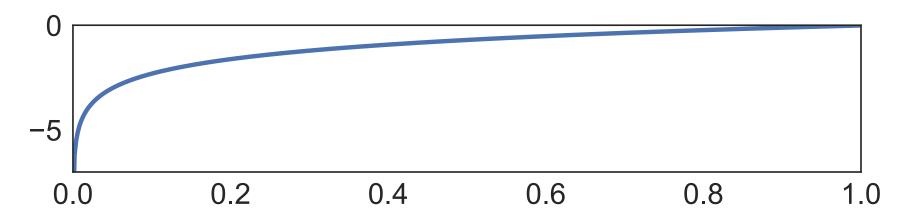


# A heuristic partial workaround

• Original GANs almost never use the minimax game

$$\min_{oldsymbol{ heta}} \max_{\psi} rac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathbb{P}}[\log D_{\psi}(X)] + rac{1}{2} \mathop{\mathbb{E}}_{Y \sim \mathbb{Q}_{oldsymbol{ heta}}}[\log(1 - D_{\psi}(Y))]$$

- $\max_{\theta} \log D_{\psi}(G_{\theta}(Z))$ , not  $\min_{\theta} \log(1 D_{\psi}(G_{\theta}(Z)))$
- If  $D_\psi$  is near-perfect, near  $\log 0$  instead of  $\log 1$



- When  $D_\psi$  is near-perfect, makes it unstable instead of stuck

#### **Solution 1: the Wasserstein distance**

$$\mathcal{W}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

 $f:\mathcal{X} o\mathbb{R}$  is a  $1 ext{-Lipschitz}$  critic function $\|f\|_{ ext{Lip}}=\sup_{x,y\in\mathcal{X}}rac{|f(x)-f(y)|}{\|x-y\|}=\sup_{x\in\mathcal{X}}\|
abla f(x)\|$ 

Turns out  $\mathcal W$  is *continuous*: if  $\mathbb Q_ heta o \mathbb P$ , then  $\mathcal W(\mathbb Q_ heta,\mathbb P) o 0$ 

#### Solution 1: the Wasserstein distance

$$\mathcal{W}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

 $f:\mathcal{X} o\mathbb{R}$  is a  $1 ext{-Lipschitz}$  critic function $\|f\|_{ ext{Lip}}=\sup_{x,y\in\mathcal{X}}rac{|f(x)-f(y)|}{\|x-y\|}=\sup_{x\in\mathcal{X}}\|
abla f(x)\|$ 

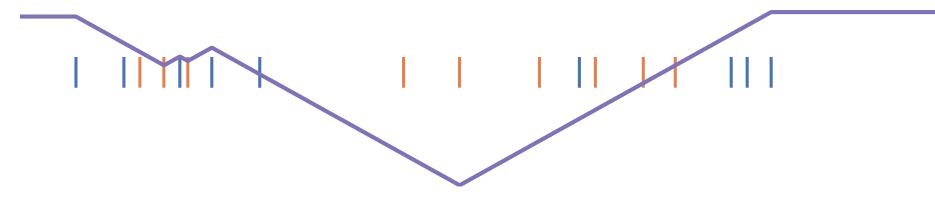
## 

Turns out  $\mathcal W$  is *continuous*: if  $\mathbb Q_ heta o \mathbb P$ , then  $\mathcal W(\mathbb Q_ heta,\mathbb P) o 0$ 

### **Solution 1: the Wasserstein distance**

$$\mathcal{W}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{ ext{Lip}} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

 $f:\mathcal{X} o\mathbb{R}$  is a  $1 ext{-Lipschitz}$  critic function $\|f\|_{ ext{Lip}}=\sup_{x,y\in\mathcal{X}}rac{|f(x)-f(y)|}{\|x-y\|}=\sup_{x\in\mathcal{X}}\|
abla f(x)\|$ 



Turns out  $\mathcal W$  is *continuous*: if  $\mathbb Q_ heta o \mathbb P$ , then  $\mathcal W(\mathbb Q_ heta,\mathbb P) o 0$ 

- Idea: turn discriminator  $D_\psi$  into a critic  $f_\psi$
- Need to enforce  $\|f_\psi\|_{ ext{Lip}} \leq 1$

- Idea: turn discriminator  $D_\psi$  into a critic  $f_\psi$
- Need to enforce  $\|f_\psi\|_{ ext{Lip}} \leq 1$
- Easy ways to do this are way too stringent

- Idea: turn discriminator  $D_\psi$  into a critic  $f_\psi$
- Need to enforce  $\|f_\psi\|_{ ext{Lip}} \leq 1$
- Easy ways to do this are way too stringent
- Instead, control  $\|
  abla f( ilde X)\|$  on average, near the data

- Idea: turn discriminator  $D_\psi$  into a critic  $f_\psi$
- Need to enforce  $\|f_\psi\|_{ ext{Lip}} \leq 1$
- Easy ways to do this are way too stringent
- Instead, control  $\|
  abla f( ilde X)\|$  on average, near the data

$$\mathop{\mathbb{E}}_{ ilde{X}\sim\mathbb{S}}ig(\|
abla_{ ilde{X}}f_\psi( ilde{X})\|-1ig)^2,\quad \mathbb{S} ext{ between }\mathbb{P} ext{ and }\mathbb{Q}_{ heta}$$

- Idea: turn discriminator  $D_\psi$  into a critic  $f_\psi$
- Need to enforce  $\|f_\psi\|_{ ext{Lip}} \leq 1$
- Easy ways to do this are way too stringent
- Instead, control  $\|
  abla f( ilde X)\|$  on average, near the data

$$\mathop{\mathbb{E}}_{ ilde{X}\sim\mathbb{S}}ig(\|
abla_{ ilde{X}}f_\psi( ilde{X})\|-1ig)^2,\quad \mathbb{S} ext{ between }\mathbb{P} ext{ and }\mathbb{Q}_{ heta}$$

• Specifically:  $ilde{X} = heta X + (1- heta) Y$  ,  $heta \sim \mathrm{Uniform}([0,1])$ 

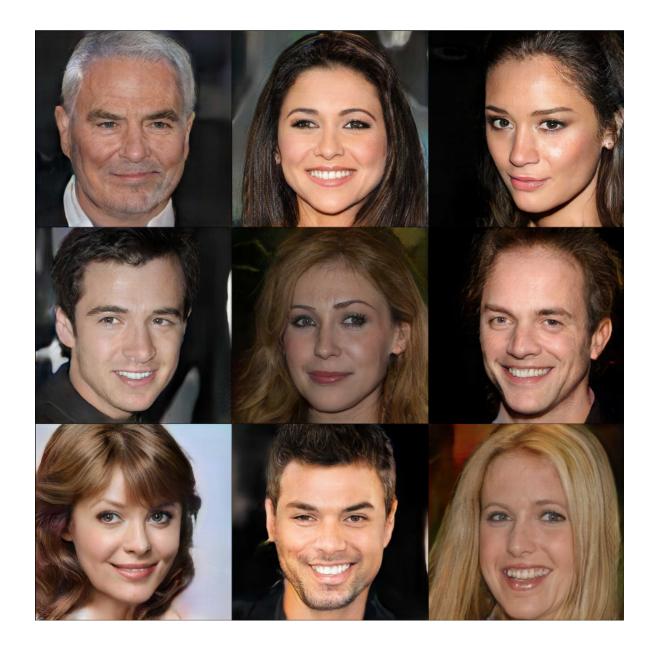
#### Solution 3: Spectral normalization [Miyato+ ICLR-18]

- Regular deep nets:  $f_\ell = \sigma \left( W_\ell f_{\ell-1}(x) + b_\ell 
  ight)$
- Spectral normalization:  $f_\ell = \sigma\left(rac{1}{\|W_\ell\|_2} W_\ell f_{\ell-1}(x) + b_\ell
  ight)$
- $\|W\|_2:= \sup_{x
  eq 0} rac{\|Wx\|_2}{\|x\|_2} = \sigma_{\max}(W)$  is the spectral norm
- Guarantees  $\|f\|_{ ext{Lip}} \leq 1$
- Faster to evaluate than gradient penalties
- Not as well understood yet

#### Solution 3: Spectral normalization [Miyato+ ICLR-18]

- Regular deep nets:  $f_\ell = \sigma \left( W_\ell f_{\ell-1}(x) + b_\ell 
  ight)$
- Spectral normalization:  $f_\ell = \sigma\left(rac{1}{\|W_\ell\|_2} W_\ell f_{\ell-1}(x) + b_\ell
  ight)$
- $\|W\|_2:= \sup_{x
  eq 0} rac{\|Wx\|_2}{\|x\|_2} = \sigma_{\max}(W)$  is the spectral norm
- Guarantees\*  $\|f\|_{ ext{Lip}} \leq 1$
- Faster to evaluate than gradient penalties
- Not as well understood yet

#### New samples [Mescheder+ ICML-18]



#### How to evaluate?



- Consider distance between distributions of image features
- Features  $\phi(x)$  from a pretrained ImageNet classifier

- Consider distance between distributions of image features
- Features  $\phi(x)$  from a pretrained ImageNet classifier

• FID: 
$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}_{\theta}}\|^2 + \operatorname{Tr}\left(\Sigma_{\mathbb{P}} + \Sigma_{\mathbb{Q}_{\theta}} - 2(\Sigma_{\mathbb{P}}\Sigma_{\mathbb{Q}_{\theta}})^{\frac{1}{2}}\right)$$

- Consider distance between distributions of image features
- Features  $\phi(x)$  from a pretrained ImageNet classifier

• FID: 
$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}_{\theta}}\|^2 + \operatorname{Tr}\left(\Sigma_{\mathbb{P}} + \Sigma_{\mathbb{Q}_{\theta}} - 2(\Sigma_{\mathbb{P}}\Sigma_{\mathbb{Q}_{\theta}})^{\frac{1}{2}}\right)$$

Estimator very biased, small variance

- Consider distance between distributions of image features
- Features  $\phi(x)$  from a pretrained ImageNet classifier

• FID: 
$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}_{\theta}}\|^2 + \operatorname{Tr}\left(\Sigma_{\mathbb{P}} + \Sigma_{\mathbb{Q}_{\theta}} - 2(\Sigma_{\mathbb{P}}\Sigma_{\mathbb{Q}_{\theta}})^{\frac{1}{2}}\right)$$

Estimator very biased, small variance

• KID: use Maximum Mean Discrepancy instead

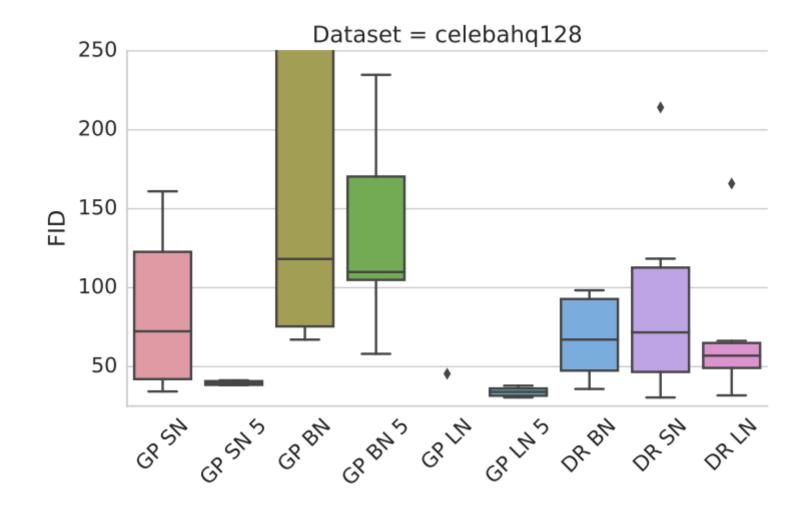
- Consider distance between distributions of image features
- Features  $\phi(x)$  from a pretrained ImageNet classifier

• FID: 
$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}_{\theta}}\|^2 + \operatorname{Tr}\left(\Sigma_{\mathbb{P}} + \Sigma_{\mathbb{Q}_{\theta}} - 2(\Sigma_{\mathbb{P}}\Sigma_{\mathbb{Q}_{\theta}})^{\frac{1}{2}}\right)$$

Estimator very biased, small variance

- KID: use Maximum Mean Discrepancy instead
  - Similar distance with unbiased, ~normal estimator!

## **Comparing approaches [Kurach+ ICML-19]**

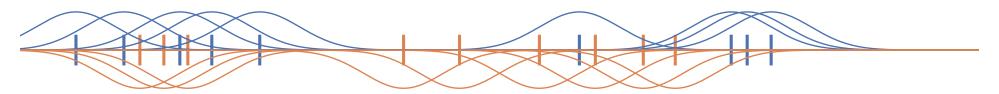


$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

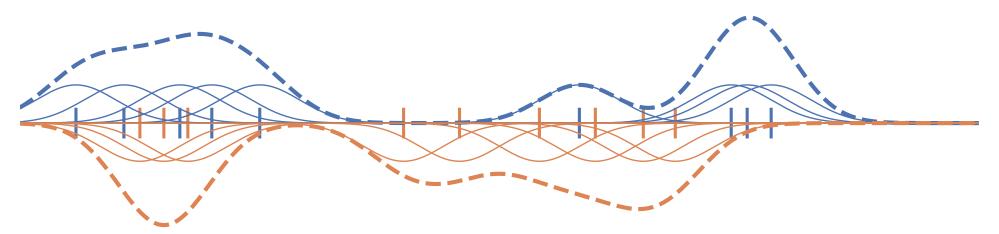
$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



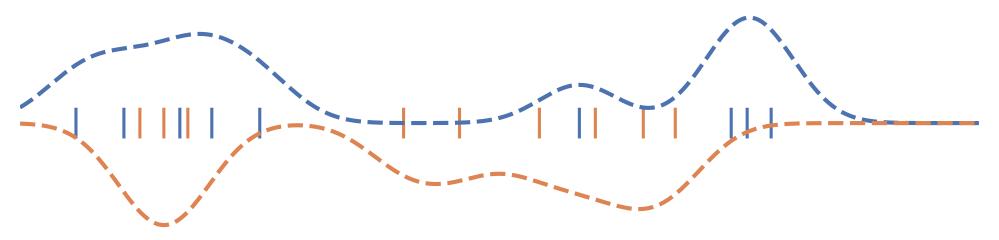
$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



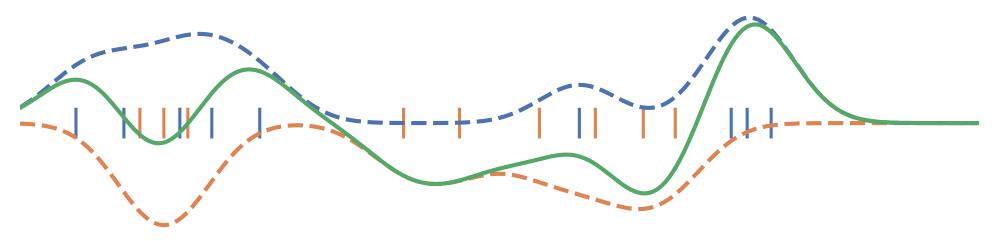
$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



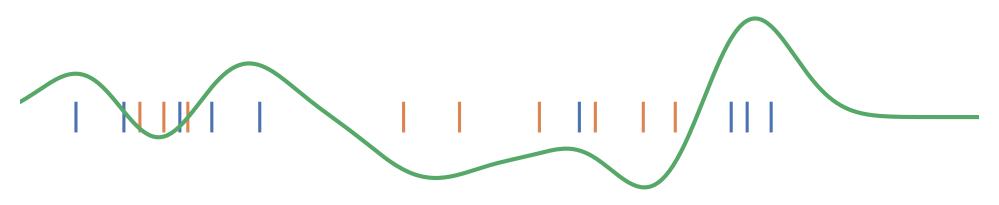
$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



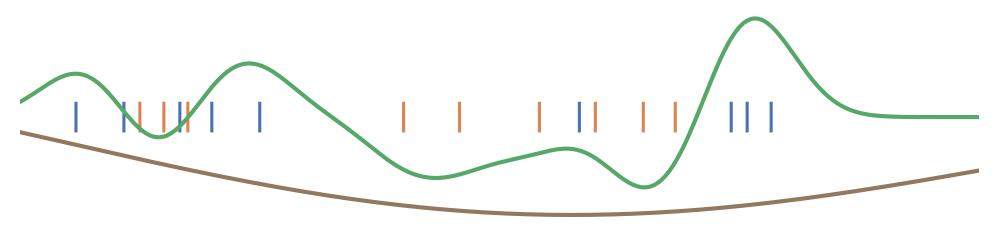
$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



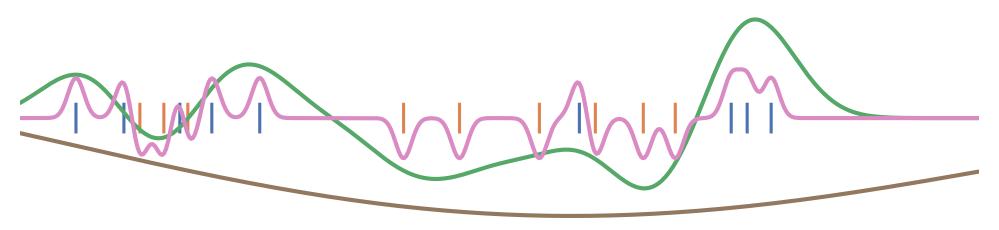
$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$



$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]$$

 $\|f\|_{\mathcal{H}_k}$  is smoothness induced by kernel  $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ Optimal f analytically:  $f^*(t) \propto \mathbb{E}_{X \sim \mathbb{P}} k(t, X) - \mathbb{E}_{Y \sim \mathbb{Q}} k(t, Y)$ 

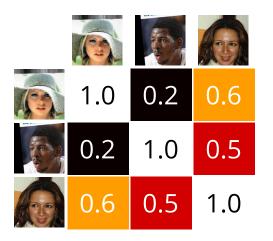
 $\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) = \mathop{\mathbb{E}}_{X,X'\sim\mathbb{P}}[k(X,X')] + \mathop{\mathbb{E}}_{Y,Y'\sim\mathbb{Q}}[k(Y,Y')] - 2\mathop{\mathbb{E}}_{\substack{X\sim\mathbb{P}} Y\sim\mathbb{Q}}[k(X,Y)]$ 

$$\begin{split} \mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \mathop{\mathbb{E}}_{X,X'\sim\mathbb{P}}[k(X,X')] + \mathop{\mathbb{E}}_{Y,Y'\sim\mathbb{Q}}[k(Y,Y')] - 2 \mathop{\mathbb{E}}_{\substack{X\sim\mathbb{P}}\\Y\sim\mathbb{Q}}}[k(X,Y)] \\ & \widehat{\mathrm{MMD}}^2(X,Y) = \mathrm{MAD}(K,Y) + \mathrm{MAD}(K,Y) \\ \end{split}$$

 $\widehat{\mathrm{MMD}}_k^2(X,Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2 \mathrm{mean}(K_{XY})$ 

$$egin{aligned} \mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \mathop{\mathbb{E}}_{X,X'\sim\mathbb{P}}[k(X,X')] + \mathop{\mathbb{E}}_{Y,Y'\sim\mathbb{Q}}[k(Y,Y')] - 2 \mathop{\mathbb{E}}_{\substack{X\sim\mathbb{P}\\Y\sim\mathbb{Q}}}[k(X,Y)] \ & \widehat{\mathrm{MMD}}_k^2(X,Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2 \operatorname{mean}(K_{XY}) \end{aligned}$$

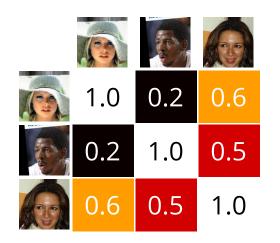


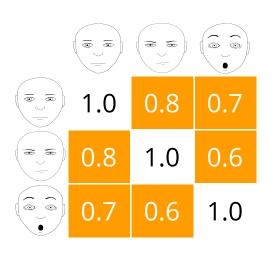


$$egin{aligned} \mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \mathop{\mathbb{E}}_{X,X'\sim\mathbb{P}}[k(X,X')] + \mathop{\mathbb{E}}_{Y,Y'\sim\mathbb{Q}}[k(Y,Y')] - 2 \mathop{\mathbb{E}}_{\substack{X\sim\mathbb{P}\\Y\sim\mathbb{Q}}}[k(X,Y)] \ & \widehat{\mathrm{MMD}}_k^2(X,Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2 \operatorname{mean}(K_{XY}) \end{aligned}$$



 $K_{YY}$ 



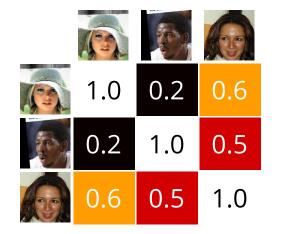


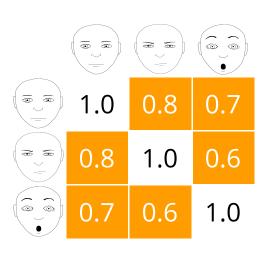
$$egin{aligned} \mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \mathop{\mathbb{E}}_{X,X'\sim\mathbb{P}}[k(X,X')] + \mathop{\mathbb{E}}_{Y,Y'\sim\mathbb{Q}}[k(Y,Y')] - 2 \mathop{\mathbb{E}}_{\substack{X\sim\mathbb{P}\\Y\sim\mathbb{Q}}}[k(X,Y)] \ & \widehat{\mathrm{MMD}}_k^2(X,Y) = \mathrm{mean}(K_{XX}) + \mathrm{mean}(K_{YY}) - 2 \mathop{\mathrm{mean}}(K_{XY}) \end{aligned}$$

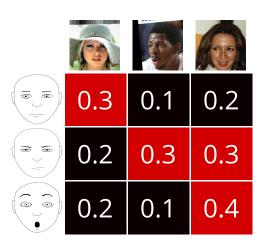




 $K_{XY}$ 







- No need for a discriminator just minimize  $\widehat{\mathrm{MMD}}_k!$
- Continuous loss

#### Critic



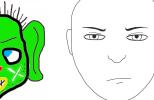
Generator ( $\mathbb{Q}_{\theta}$ )



- No need for a discriminator just minimize  $\widehat{\mathrm{MMD}}_k!$
- Continuous loss





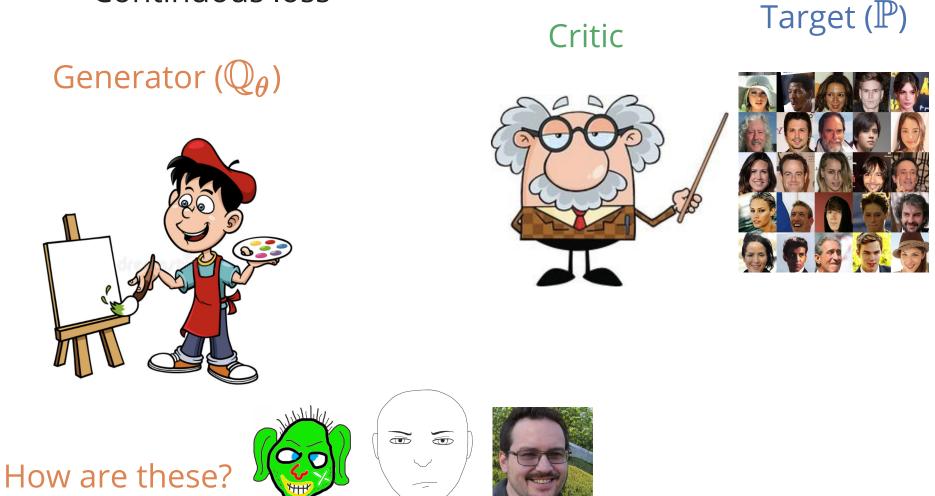




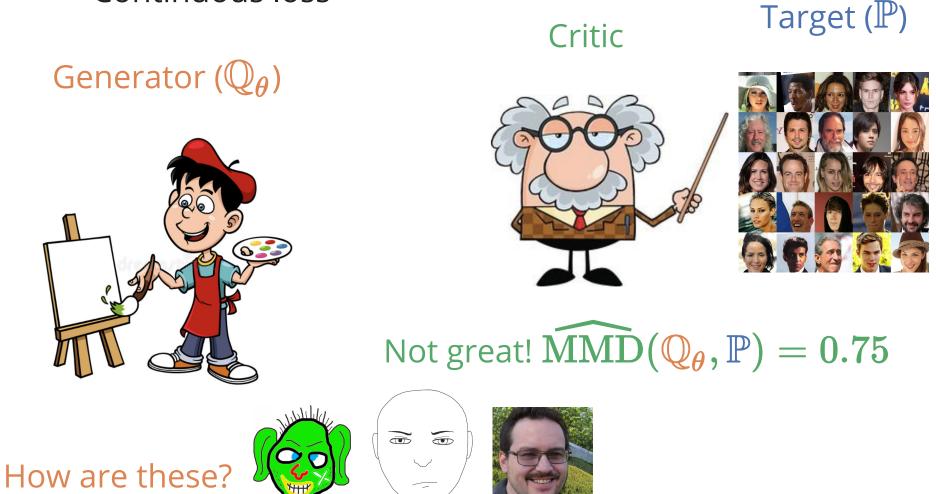
### Critic



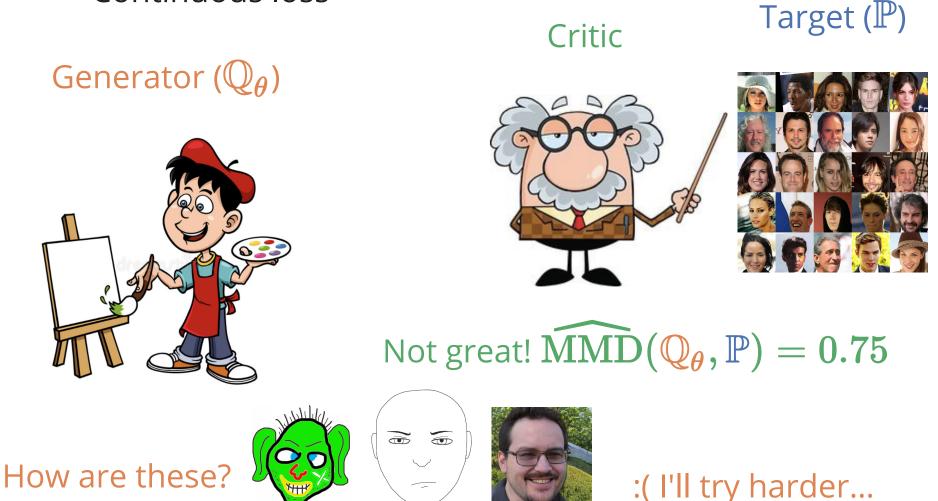
- No need for a discriminator just minimize  $\widehat{\mathrm{MMD}}_k!$
- Continuous loss



- No need for a discriminator just minimize  $\widehat{\mathrm{MMD}}_k!$
- Continuous loss



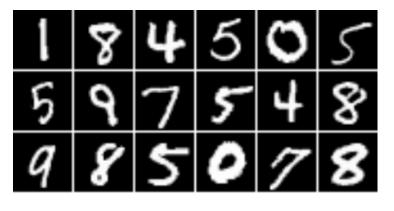
- No need for a discriminator just minimize  $\widehat{\mathrm{MMD}}_k!$
- Continuous loss



- No need for a discriminator just minimize  $\widehat{\mathrm{MMD}}_k!$
- Continuous loss

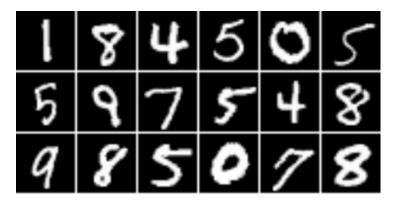


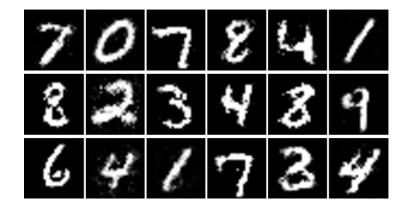
#### MNIST, mix of Gaussian kernels



 $\mathbb{P}$ 

#### MNIST, mix of Gaussian kernels

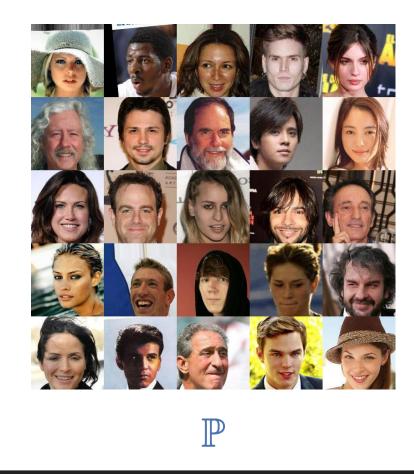




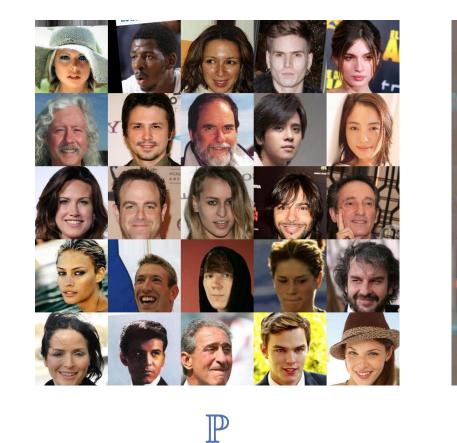
 $\mathbb{P}$ 

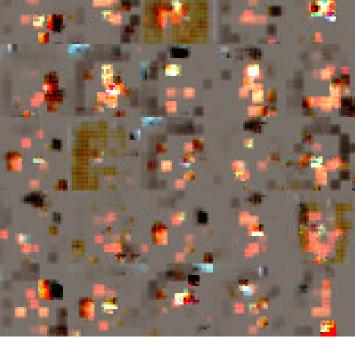
 $\mathbb{Q}_{\theta}$ 

#### Celeb-A, mix of rational quadratic + linear kernels



#### Celeb-A, mix of rational quadratic + linear kernels





 $\mathbb{Q}_{oldsymbol{ heta}}$ 

$$k(x,y) = k_{ ext{top}}(\phi(x),\phi(y))$$

- $\phi: \mathcal{X} 
  ightarrow \mathbb{R}^{2048}$  from pretrained Inception net
- $k_{
  m top}$  simple: exponentiated quadratic or polynomial

$$k(x,y) = k_{ ext{top}}(\phi(x),\phi(y))$$

- $\phi: \mathcal{X} 
  ightarrow \mathbb{R}^{2048}$  from pretrained Inception net
- $k_{
  m top}$  simple: exponentiated quadratic or polynomial



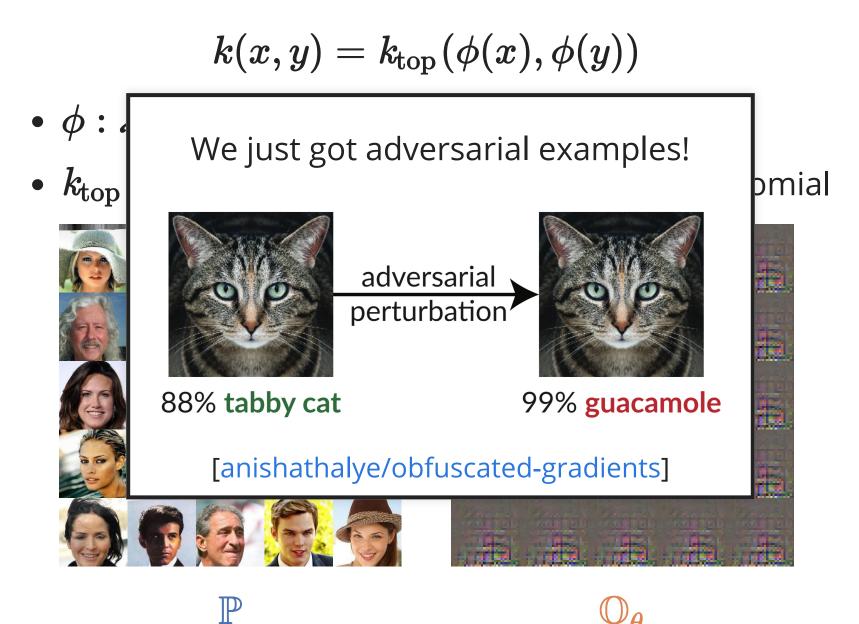
$$k(x,y) = k_{ ext{top}}(\phi(x),\phi(y))$$

- $\phi: \mathcal{X} 
  ightarrow \mathbb{R}^{2048}$  from pretrained Inception net
- $k_{
  m top}$  simple: exponentiated quadratic or polynomial



		æ.		
			. <b>1</b> 1.	
niffi.		Æ.	Æ.	nita.
ιĒ.	.a		Æ.	
	Æ.		. Al	dia.





• Don't just use one kernel, use a *class* parameterized by  $\psi$ :

$$k_\psi(x,y) = k_{
m top}(\phi_\psi(x),\phi_\psi(y))$$

• Don't just use one kernel, use a *class* parameterized by  $\psi$ :

$$k_\psi(x,y) = k_{
m top}(\phi_\psi(x),\phi_\psi(y))$$

• New distance based on *all* these kernels:

$$\mathcal{D}_{ ext{MMD}}(\mathbb{P},\mathbb{Q}) = \sup_{\psi\in\Psi} ext{MMD}_{\psi}(\mathbb{P},\mathbb{Q})$$

• Don't just use one kernel, use a *class* parameterized by  $\psi$ :

$$k_\psi(x,y) = k_{
m top}(\phi_\psi(x),\phi_\psi(y))$$

• New distance based on *all* these kernels:

$$\mathcal{D}_{ ext{MMD}}(\mathbb{P},\mathbb{Q}) = \sup_{\psi\in\Psi} ext{MMD}_{\psi}(\mathbb{P},\mathbb{Q})$$

• Turns out that  $\mathcal{D}_{\mathrm{MMD}}$  *isn't* continuous: have  $\mathbb{Q}_{\theta} \to \mathbb{P}$  but  $\mathcal{D}_{\mathrm{MMD}}(\mathbb{Q}_{\theta}, \mathbb{P}) \not\rightarrow 0$ 

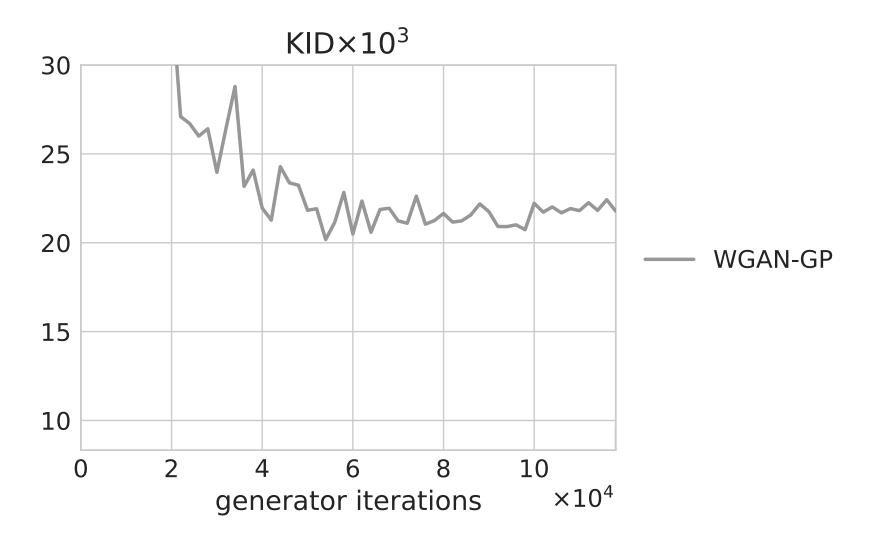
• Don't just use one kernel, use a *class* parameterized by  $\psi$ :

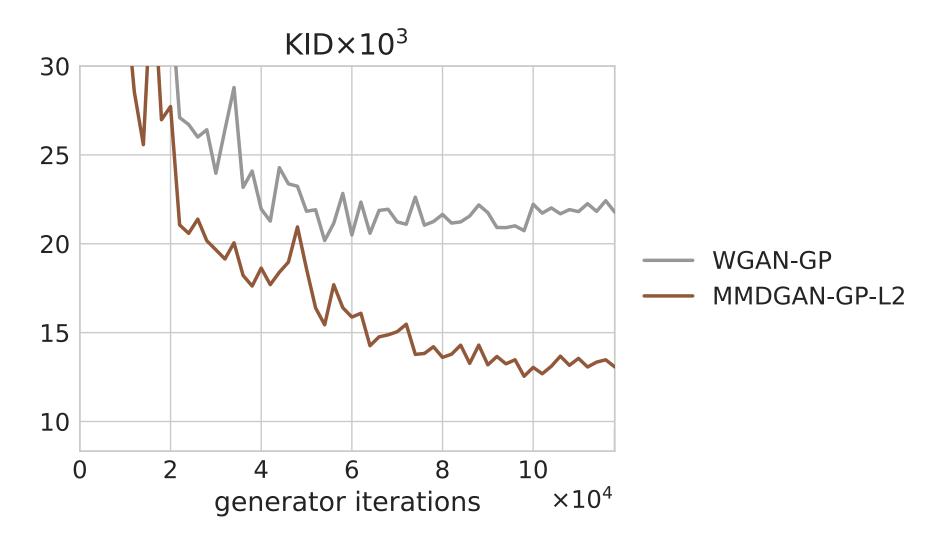
$$k_\psi(x,y) = k_{ ext{top}}(\phi_\psi(x),\phi_\psi(y))$$

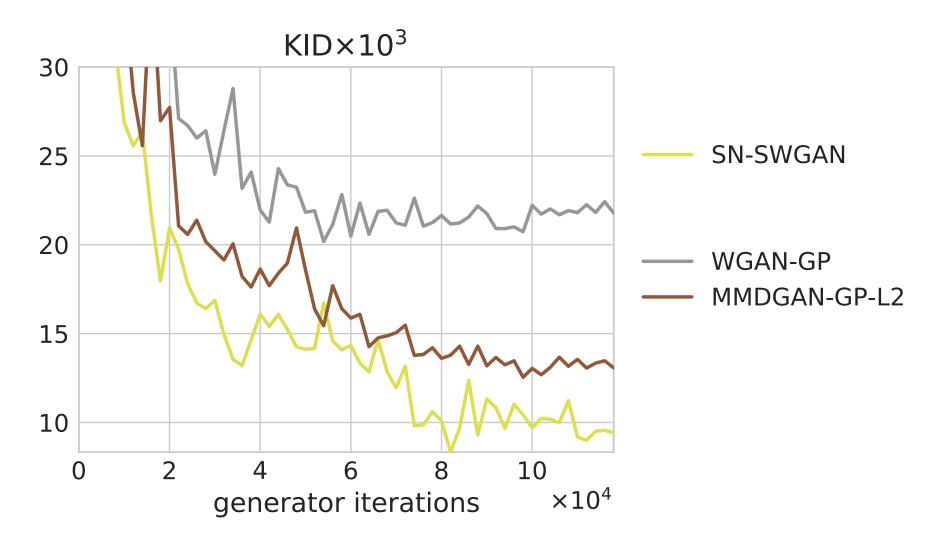
• New distance based on *all* these kernels:

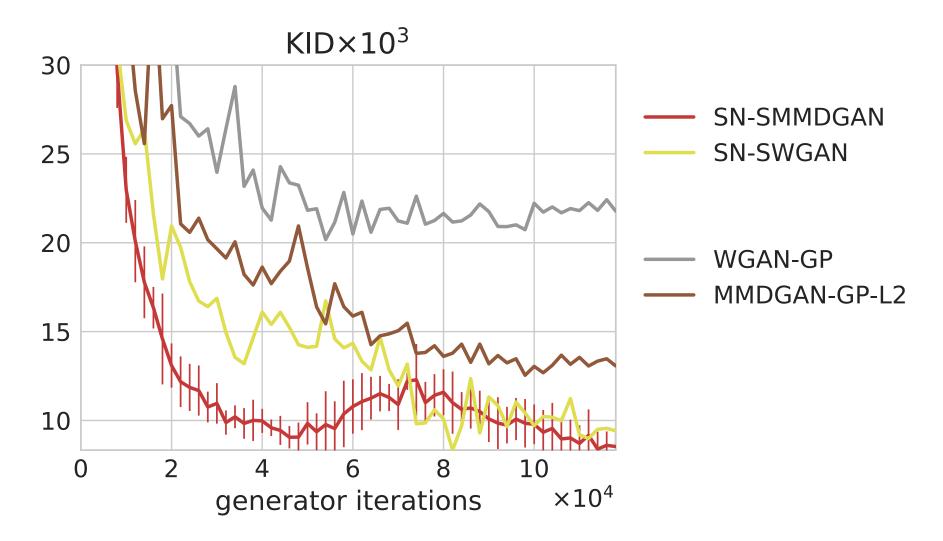
$$\mathcal{D}_{ ext{MMD}}(\mathbb{P},\mathbb{Q}) = \sup_{\psi\in\Psi} ext{MMD}_{\psi}(\mathbb{P},\mathbb{Q})$$

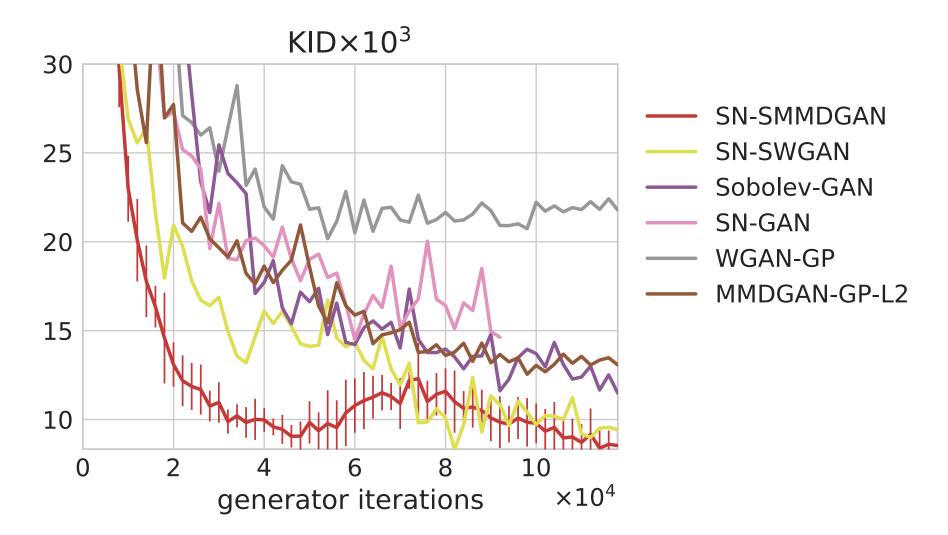
- Turns out that  $\mathcal{D}_{\mathrm{MMD}}$  isn't continuous: have  $\mathbb{Q}_{\theta} \to \mathbb{P}$  but  $\mathcal{D}_{\mathrm{MMD}}(\mathbb{Q}_{\theta}, \mathbb{P}) \not\rightarrow 0$
- Scaled MMD GANs [Arbel+ NeurIPS-18] correct  $\mathcal{D}_{MMD}$  with a gradient penalty to make it continuous



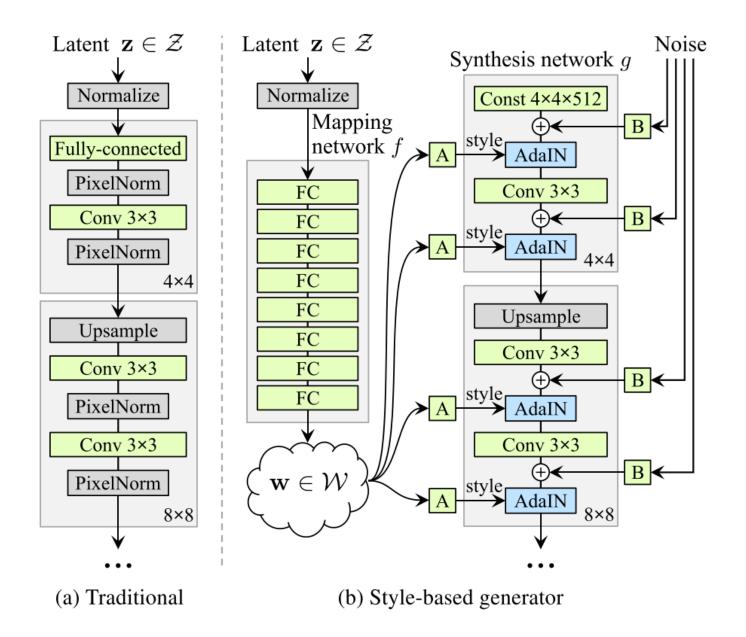








## StyleGANs [Karras+ 2018]



### **StyleGAN: latent structure**

## **StyleGAN: local noise**



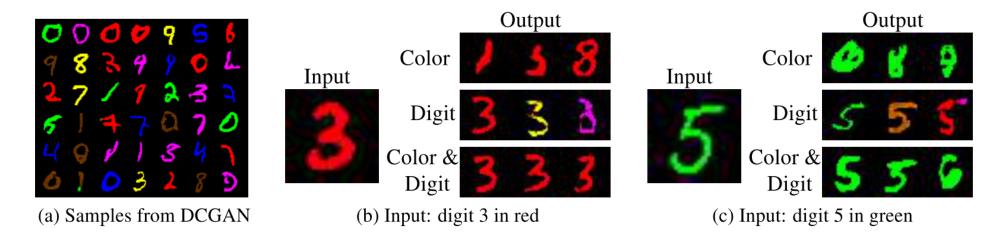
(a) Generated image (b) Stochastic variation (c) Standard deviation

## StyleGANs on a different domain [@roadrunning01]

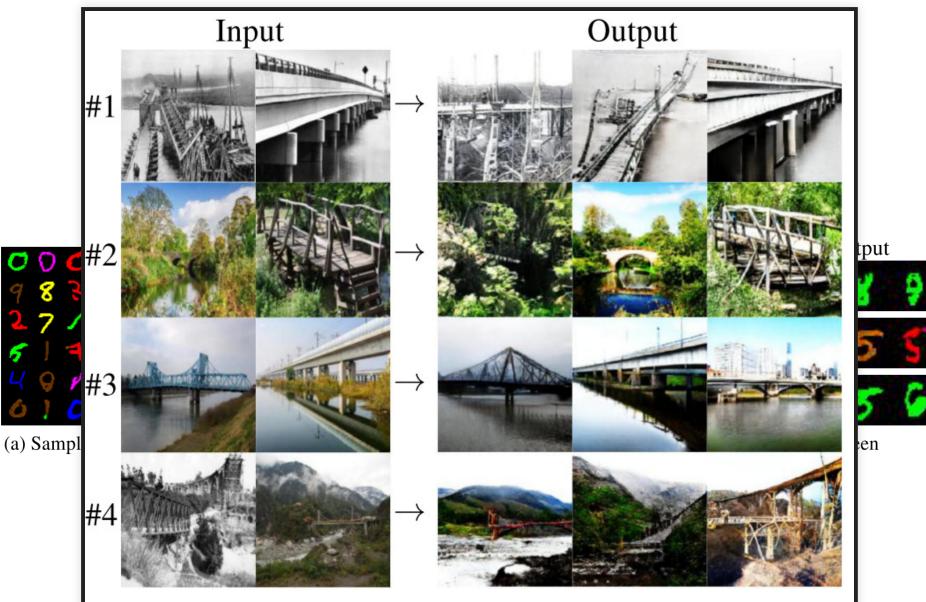
# Finding samples you want [Jitkrittum+ ICML-19]

If we want to find "more samples like  $\{X\}$ ":

$$\min_{\{Z_1,\ldots,Z_n\}} \widehat{\mathrm{MMD}}_k^2ig(\{X_i\}_{i=1}^m,\{G_{oldsymbol{ heta}}(Z_i)\}_{i=1}^nig)$$



## Finding samples you want [Jitkrittum+ ICML-19]



( ) I CIINI 1. . 1.

# **Conditional GANs and BigGAN**

- Conditional GANs: [Mirza+ 2014]
  - ullet Just add a class label as input to  $G_ heta$  and  $D_\psi$
- BigGAN [Brock+ ICLR-19]: a bunch of tricks to make it huge



## Image-to-image translation [Isola+ CVPR-17]

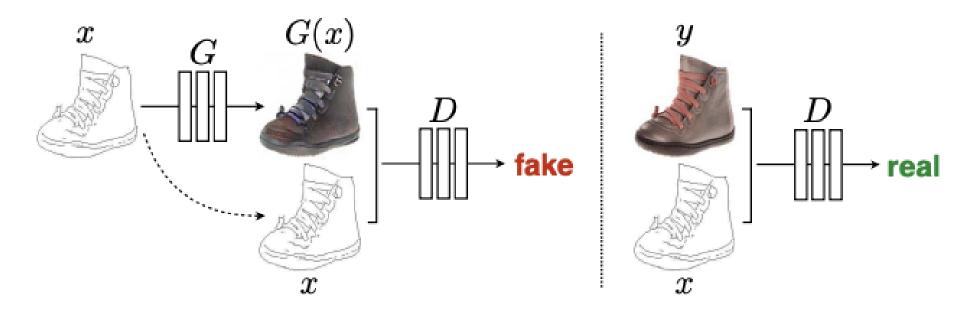
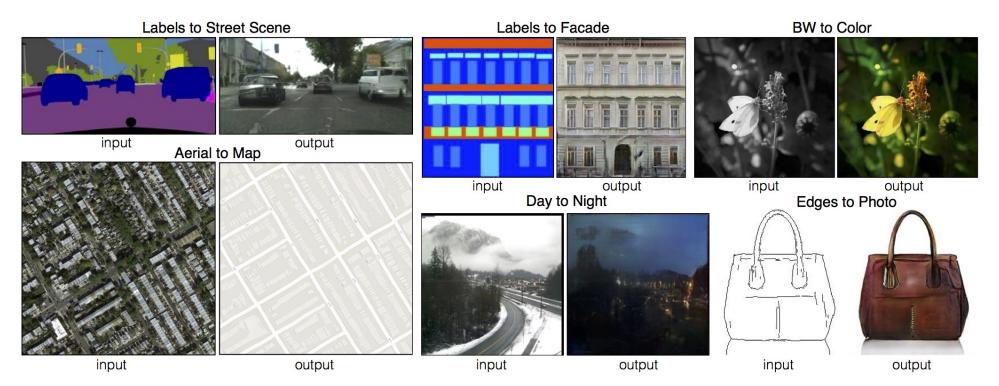
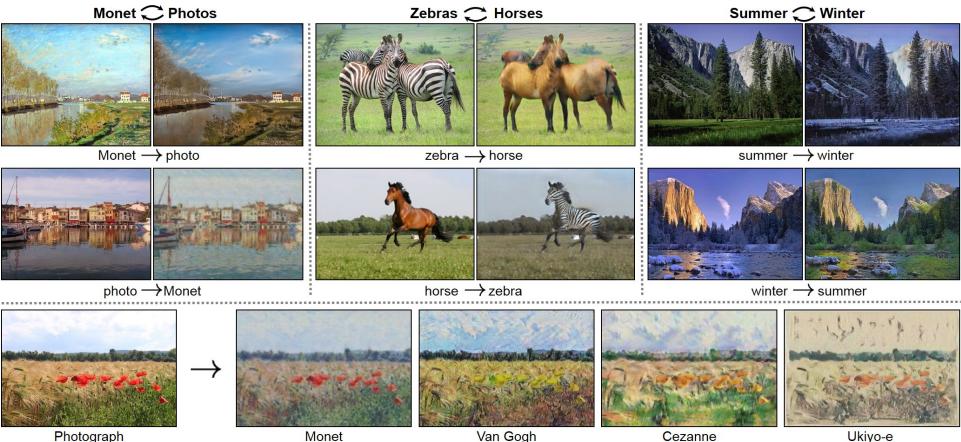


Figure 2: Training a conditional GAN to map edges $\rightarrow$ photo. The discriminator, D, learns to classify between fake (synthesized by the generator) and real {edge, photo} tuples. The generator, G, learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe the input edge map.

## Image-to-image translation [Isola+ CVPR-17]



### CycleGAN [Zhu+ ICCV-17]



Photograph

Monet

Van Gogh

Ukiyo-e

### Pose-to-image translation [Chan+ 2018]

### DeepFakes

## Use your new knowledge for good!

Slides (including links to papers) are online: dougal.me/slides/gans-mlcc

## Use your new knowledge for good!

#### Slides (including links to papers) are online: dougal.me/slides/gans-mlcc

