

# **TTIC 31230, Fundamentals of Deep Learning**

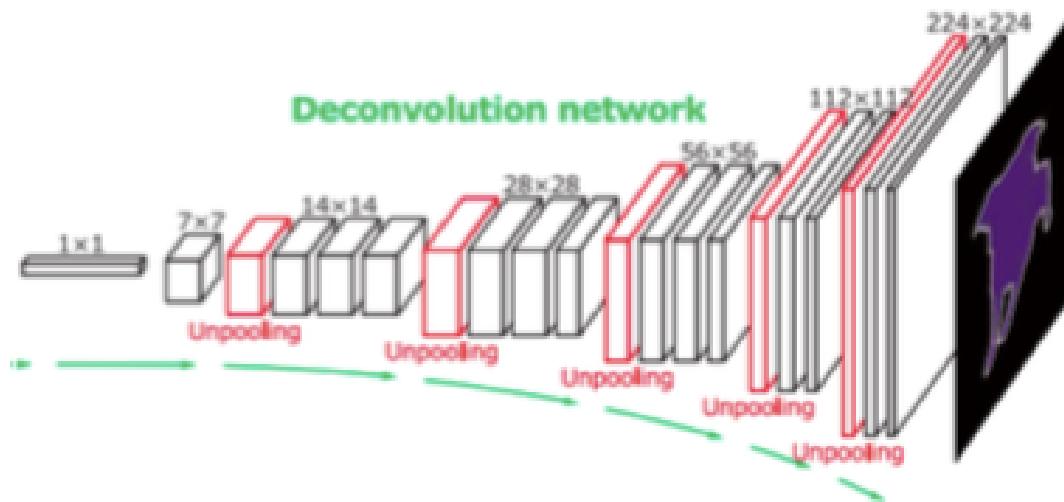
David McAllester, Winter 2020

## **Generative Adversarial Networks (GANs)**

# Representing a Distribution with a Generator

$$z \sim \mathcal{N}(0, I)$$

$$y \sim p_\Phi$$



## Generative Adversarial Networks (GANs)

Let  $y$  range over images. We have a generator  $p_\Phi$ . For  $i \in \{-1, 1\}$  we define a probability distribution over pairs  $\langle y, i \rangle$  by

$$\begin{aligned} \Psi : \text{dis} & \quad \tilde{p}_\Phi(i=1) = 1/2 \\ \varphi : \text{gen} & \quad \tilde{p}_\Phi(y|i=1) = \text{pop}(y) \\ & \quad \tilde{p}_\Phi(y|i=-1) = p_\Phi(y) \end{aligned} \quad \xrightarrow{\text{Sample from generator}}$$

We also have a discriminator  $P_\Psi(i|y)$ .

$$\Phi^* = \underset{\Phi}{\text{argmax}} \min_{\Psi} E_{\langle y, i \rangle \sim \tilde{p}_\Phi} - \ln P_\Psi(i|y)$$

*"adversarial"*

Assuming universality of both  $p_\Phi$  and  $P_\Psi$  we have  $p_{\Phi^*} = \text{pop}$ .

when  $i, y$  have same <sup>3</sup> information  $\Rightarrow$

## The Discriminator Tends to Win

The log loss for the binary discrimination classifier is quickly driven to very near zero.

This causes the learning gradient to also become essentially zero and the learning stops.

## Review of Binary Classification

In the case of binary classification cross-entropy loss becomes the log loss of the margin

$$\Psi^* = \operatorname{argmin}_{\Psi} E_{(i,y) \sim \tilde{p}_\Phi} - \ln P_\Psi(i|y)$$

only 2 classes

make one

$$= \operatorname{argmin}_{\Psi} E_{(i,y) \sim \tilde{p}_\Phi} \ln(1 + e^{-m})$$

$\nearrow$  = minus of  
the other

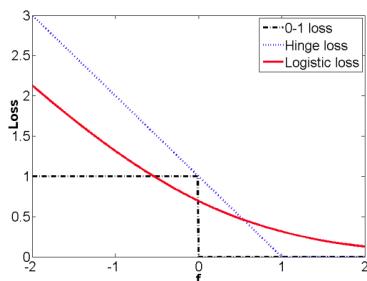
$$m = 2is_\Psi(i|y) \quad \text{for} \quad s_\Psi(-1|y) = -s_\Psi(1|y)$$

$$P_\Psi(i|y) = \operatorname{softmax}_i s_\Phi(i|y)$$

↙

diff of

2 slope



## Vanishing Gradients

For  $i = 1$  and  $y \sim \text{pop}$ :

$$\Psi += \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Psi} m \approx 0 \text{ for } m \gg 1$$

For  $i = -1$  and  $y \sim p_{\Phi}$ :

$$\Psi += \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Psi} m \approx 0 \text{ for } m \gg 1$$

$$\Phi -= \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Phi} m \approx 0 \text{ for } m \gg 1$$

The gradients vanish when the discriminator achieves large margins.

## A Heuristic Patch

Replace

$$\Phi -= \eta \frac{e^{-m}}{1 + e^{-m}} \nabla_{\Phi} m \approx 0 \text{ for } m \gg 1$$

with

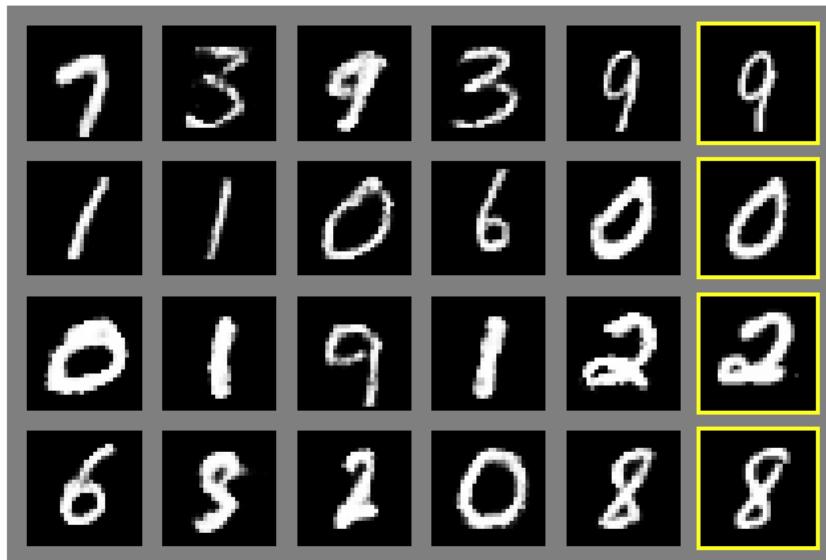
$$\Phi -= \eta \nabla_{\Phi} m$$

This allows the generator to recover.

Change  $\eta$  by + weight  
not correct but works

# Generative Adversarial Nets

## Goodfellow et al., June 2014



The rightmost column (yellow boarders) gives the nearest neighbor in the training data to the adjacent column.

Not sampled  
from training  
set

assume  $\Psi$  perfect  $\Rightarrow \Psi^*$

Assuming Universality of  $\Psi$  Only

optimal  $\Psi$  learn

$\rightarrow$  true dist

$$\Psi^*(\Phi) = \arg \min_{\Psi} E_{(i,y) \sim \tilde{p}_{\Phi}} - \ln \tilde{p}_{\Phi}(i|y)$$

$$\tilde{p}_{\Phi}(i|y) = \frac{\tilde{p}_{\Phi}(i \wedge y)}{\tilde{p}_{\Phi}(y)}$$

$\hookrightarrow$  sample from generator,

= true pop when

$$\Psi = \Psi^*$$

$$\tilde{p}(1|y) = \frac{\frac{1}{2}\text{pop}(y)}{\frac{1}{2}\text{pop}(y) + \frac{1}{2}p_{\Phi}(y)}$$

$$\tilde{p}(-1|y) = \frac{\frac{1}{2}p_{\Phi}(y)}{\frac{1}{2}\text{pop}(y) + \frac{1}{2}p_{\Phi}(y)}$$

9

$\rightarrow$  sample

$$\left\{ \begin{array}{ll} \frac{1}{2} & \text{pop} \\ \frac{1}{2} & p_{\Phi} \end{array} \right.$$

## Assuming Universality of $\Psi$ Only

$$\begin{aligned} & E_{(i,y) \sim p_\Phi} \ln \tilde{p}_\Phi(i|y) \\ &= \frac{1}{2} E_{y \sim \text{pop}} \ln \frac{\frac{1}{2}\text{pop}(y)}{\frac{1}{2}\text{pop}(y) + \frac{1}{2}p_\Phi(y)} + \frac{1}{2} E_{y \sim p_\Phi} \ln \frac{\frac{1}{2}p_\Phi(y)}{\frac{1}{2}\text{pop}(y) + \frac{1}{2}p_\Phi(y)} \\ &= \frac{1}{2} \left( KL \left( \text{pop}, \frac{\text{pop} + p_\Phi}{2} \right), KL \left( p_\Phi, \frac{\text{pop} + p_\Phi}{2} \right) \right) - \ln 2 \\ &= \text{JSD}(\text{pop}, p_\Phi) - \ln 2 \end{aligned}$$

## Assuming Universality of $\Psi$ Only

$$\begin{aligned} & E_{(i,y) \sim \tilde{p}_\Phi} - \ln \tilde{p}_\Phi(i|y) \\ &= \ln 2 - JSD(\text{pop}, p_\Phi) \end{aligned}$$

## Contrastive GANs

A GAN can be built with a “contrastive” discriminator. Rather than estimate the probability that  $y$  is from the population, the discriminator must select which of  $y_1, \dots, y_N$  is from the population.

$$\psi \xrightarrow{\text{choose}} \text{chance} = \frac{1}{N}$$

More formally, for  $N \geq 2$  let  $\tilde{p}_\Phi^{(N)}$  be the distribution defined by drawing one “positive” from  $p_\Phi$  and  $N - 1$  IID negatives from  $p_\Phi$ ; then inserting the positive at a random position among the negatives; and returning  $(i, y_1, \dots, y_N)$  where  $i$  is the index of the positive.

## Contrastive GANs

$$\Psi^*(\Phi) = \operatorname{argmin}_{\Psi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_\Phi^{(N)}} - \ln P_\Psi(i|y_1, \dots, y_N)$$

$$\Phi^* = \operatorname{argmax}_{\Phi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_\Phi^{(N)}} - \ln P_{\Psi^*(\Phi)}(i|y_1, \dots, y_N)$$

$\tilde{p}_\Phi^{(2)}(i|y_1, y_2)$  requires a choice between two  $y$ 's while  $\tilde{p}_\Phi(i|y)$  classifies a single  $y$  — these are different.

The discrimination gets more difficult as  $N$  gets larger.

## Contrastive GANs

$$\Psi^*(\Phi) = \operatorname{argmin}_{\Psi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_\Phi^{(N)}} - \ln P_\Psi(i | y_1, \dots, y_N)$$

$$\Phi^* = \operatorname{argmax}_{\Phi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_\Phi^{(N)}} - \ln P_{\Psi^*(\Phi)}(i | y_1, \dots, y_N)$$

Assuming universality

$$\mathcal{L}_{\text{Discr}}(\Psi^*(\Phi)) = H_\Phi(i | y_1, \dots, y_N)$$

$$p_{\Phi^*} = \text{pop} \quad H_{\Phi^*}(i | y_1, \dots, y_N) = \ln N$$

# Noise Contrastive Estimation

## Gutmann and Hyvärinen, 2010

$$\Psi^* = \operatorname{argmin}_{\Psi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_\Phi^{(N)}} - \ln P_\Psi(i|y_1, \dots, y_N)$$

$p_\Phi$  is fixed “noise”

Assume  $p_\Phi$  is both samplable and computable — we can sample from  $p_\Phi$  and for any given  $y$  we can compute  $p_\Phi(y)$ .

Assume  $P_\Psi(i|y_1, \dots, y_N) = \operatorname{softmax}_i s_\Psi(y_i)$

Assume  $\Psi$  universal

# Noise Contrastive Estimation

$$\Psi^* = \operatorname{argmin}_{\Psi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_{\Phi}^{(N)}} - \ln P_{\Psi}(i | y_1, \dots, y_N)$$

$p_{\Phi}$  is fixed “noise” noise probability  
of  $y$

Theorem:  $\text{pop}(y) = \text{softmax}_y s_{\Psi^*}(y) + \ln p_{\Phi}(y)$

$\downarrow$   $\downarrow$   
*Computable*

We then have a computable score function (energy function) for the population. We do not have the partition function  $Z$ .

$Z$ : Partition func  
but not get  $Z$ , but can calculate ratio

## Noise Contrastive Estimation

$$\Psi^* = \operatorname{argmin}_{\Psi} E_{(i, y_1, \dots, y_N) \sim \tilde{p}_{\Phi}^{(N)}} - \ln P_{\Psi}(i | y_1, \dots, y_N)$$

$p_{\Phi}$  is fixed “noise”

Lemma:  $P_{\Psi^*}(i | y_1, \dots, y_N) = \operatorname{softmax}_i \ln \frac{\operatorname{pop}(y_i)}{p_{\Phi}(y_i)}$

## Lemma Proof

$$\tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \dots, y_N) = \frac{1}{N} \text{pop}(y_i) \prod_{j \neq i} p_\Phi(y_j)$$

$$= \alpha \frac{\text{pop}(y_i)}{p_\Phi(y_i)}, \quad \alpha = \frac{1}{N} \prod_i p_\Phi(y_i)$$

$$\tilde{p}_\Phi^{(N)}(i \mid y_1, \dots, y_N) = \frac{\tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \dots, y_N)}{\sum_i \tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \dots, y_N)} = \frac{1}{Z} \frac{\text{pop}(y_i)}{p_\Phi(y_i)}$$

$$= \underset{i}{\text{softmax}} \left( \ln \frac{\text{pop}(y_i)}{p_\Phi(y_i)} \right)$$

## Theorem Proof

$$\text{softmax } s_{\Psi^*}(y_i) = \text{softmax}_i \ln \frac{\text{pop}(y_i)}{p_{\Phi}(y_i)}$$

is solved by

$$s_{\Psi^*}(y) = \ln \frac{\text{pop}(y)}{p_{\Phi}(y)} - \ln Z$$

giving

$$\text{pop}(y) = \frac{1}{Z} \exp(s_{\Psi}(y) + \ln p_{\Phi}(y))$$

$P \Rightarrow P = e^{s + Z + p_{\Phi}}$

## Another Theorem

$$E_{(i,y_1,\dots,y_N) \sim \tilde{p}_\Phi^{(N)}} - \ln p_{\Psi^*}^{(N)}(i|y_1, \dots, y_N)$$

$$\geq \ln N - \frac{N-1}{N} (KL(\text{pop}, p_\Phi) + KL(p_\Phi, \text{pop}))$$

Note that the bound holds with equality for  $p_\Phi = \text{pop}$ .

This is analogous to the JSD expression for the optimal discriminator loss.

## Proof Part A.

$$\begin{aligned}
& E_{(i,y_1,\dots,y_N) \sim \tilde{p}_{\Phi}^{(N)}} \ln p_{\Psi^*}(i|y_1, \dots, y_N) \\
&= E_{(i,y_1,\dots,y_N) \sim \tilde{p}_{\Phi}^{(N)}} \ln \left( \underset{i}{\text{softmax}} \ln \frac{\text{pop}(y_i)}{p_{\Phi}(y_i)} \right) [i] \\
&= E_{(i,y_1,\dots,y_N) \sim \tilde{p}_{\Phi}^{(N)}} \ln \frac{\text{pop}(y_i)}{p_{\Phi}(y_i)} - \ln \left( \sum_j \frac{\text{pop}(y_j)}{p_{\Phi}(y_j)} \right) \\
&= E_{(i,y_1,\dots,y_N) \sim \tilde{p}_{\Phi}^{(N)}} \ln \frac{\text{pop}(y_i)}{p_{\Phi}(y_i)} - \ln \left( \frac{1}{N} \sum_j \frac{\text{pop}(y_j)}{p_{\Phi}(y_j)} \right) - \ln N
\end{aligned}$$

## Proof Part B.

$$\begin{aligned}
& E_{(i,y_1,\dots,y_N) \sim \tilde{p}_\Phi^{(N)}} \ln \frac{\text{pop}(y_i)}{p_\Phi(y_i)} - \ln \left( \frac{1}{N} \sum_j \frac{\text{pop}(y_j)}{p_\Phi(y_j)} \right) - \ln N \\
& \leq E_{(i,y_1,\dots,y_N) \sim \tilde{p}_\Phi^{(N)}} \ln \frac{\text{pop}(y_i)}{p_\Phi(y_i)} - \frac{1}{N} \sum_j \ln \frac{\text{pop}(y_j)}{p_\Phi(y_j)} - \ln N \\
& = E_{y \sim \text{pop}} \ln \frac{\text{pop}(y)}{p_\Phi(y)} - E_{(i,y_1,\dots,y_N) \sim \tilde{p}_\Phi^{(N)}} \frac{1}{N} \sum_j \ln \frac{\text{pop}(y_j)}{p_\Phi(y_j)} - \ln N \\
& = \frac{N-1}{N} (KL(\text{pop}, p_\Phi) + KL(p_\Phi, \text{pop})) - \ln N
\end{aligned}$$

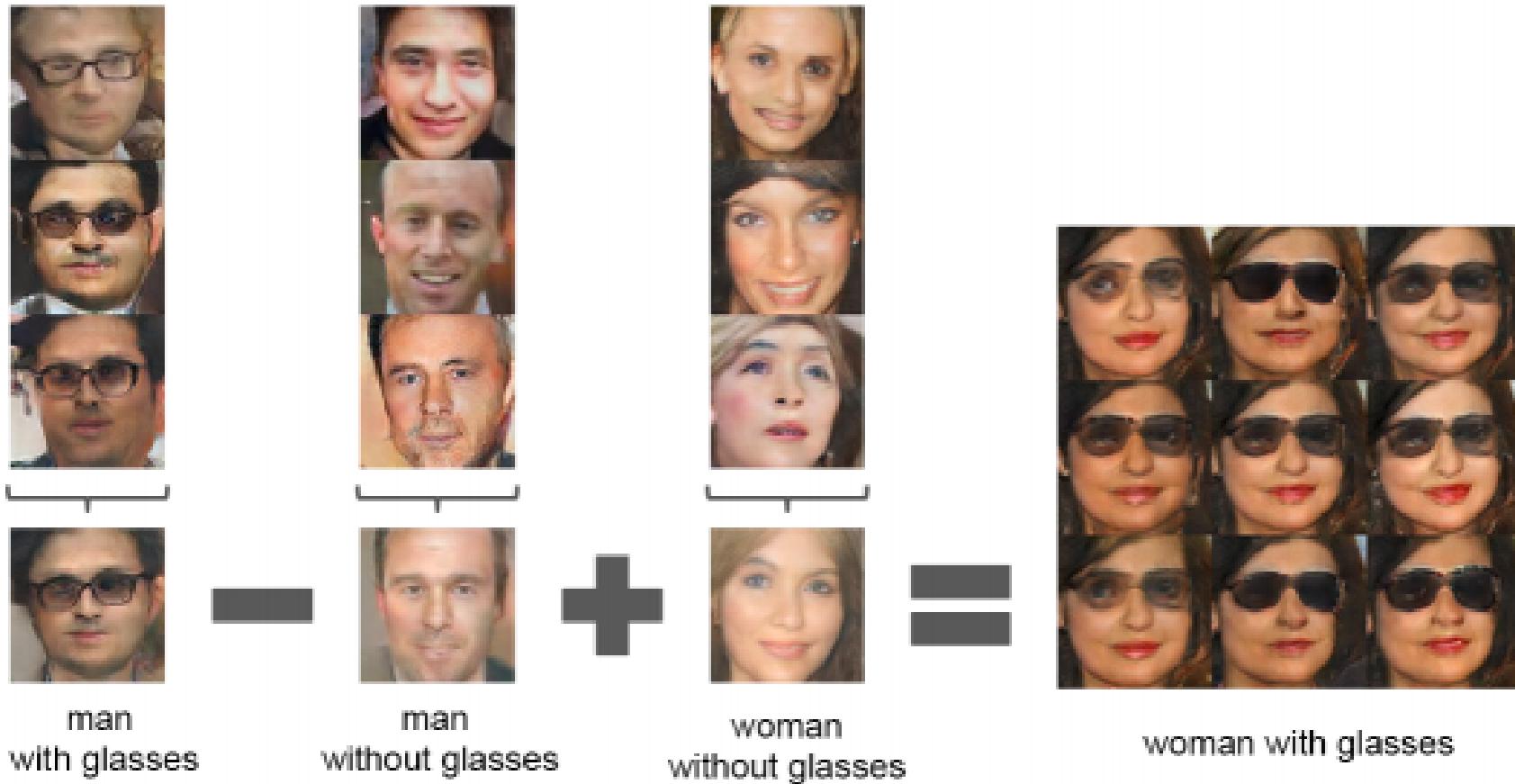
# Unsupervised Representation Learning ... (DC GANS)

## Radford et al., Nov. 2015



# Unsupervised Representation Learning ... (DC GANS)

Radford et al., Nov. 2015



# Interpolated Faces

[Ayan Chakrabarti, January 2017]



# Progressive Growing of GANs

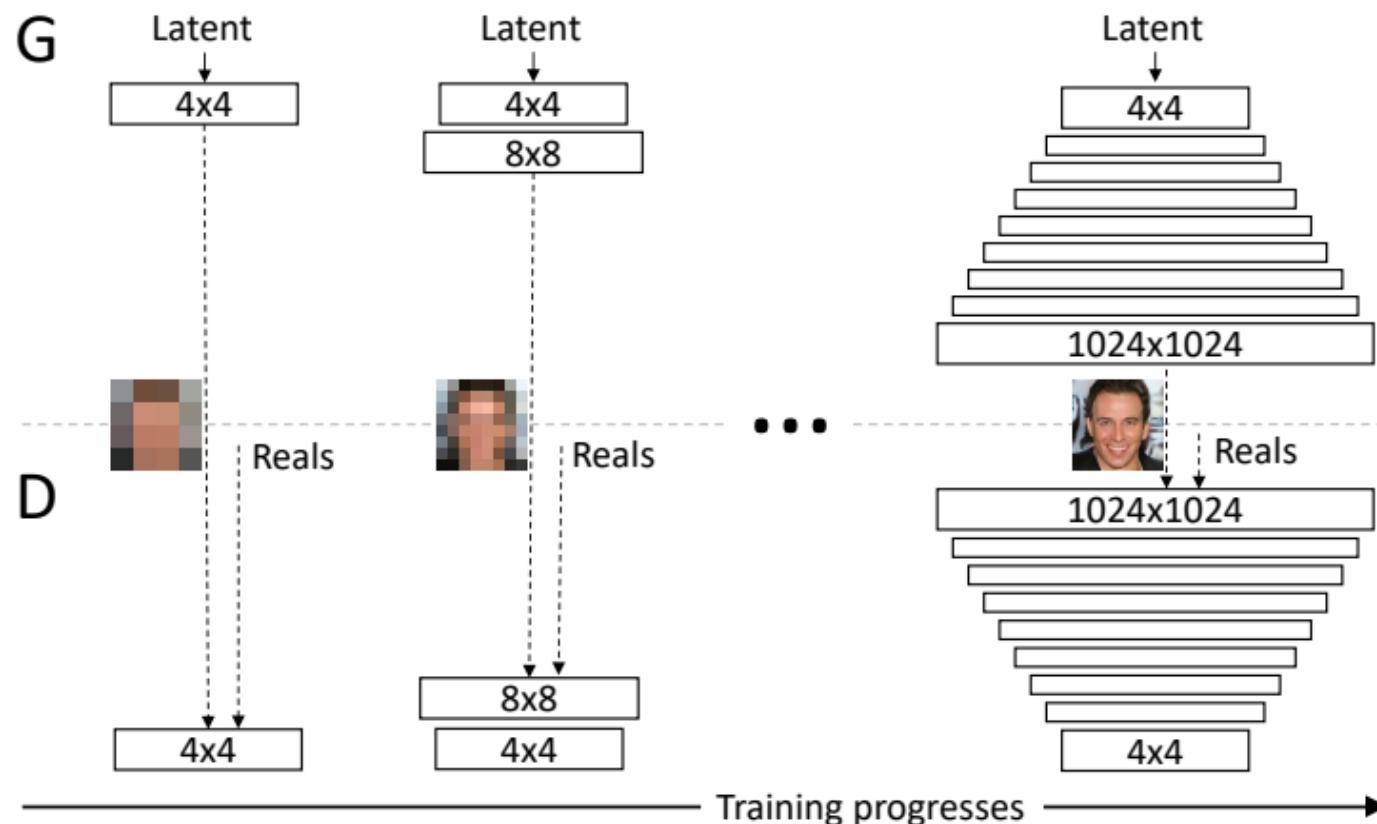
Karras et al., Oct. 2017



Figure 5:  $1024 \times 1024$  images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

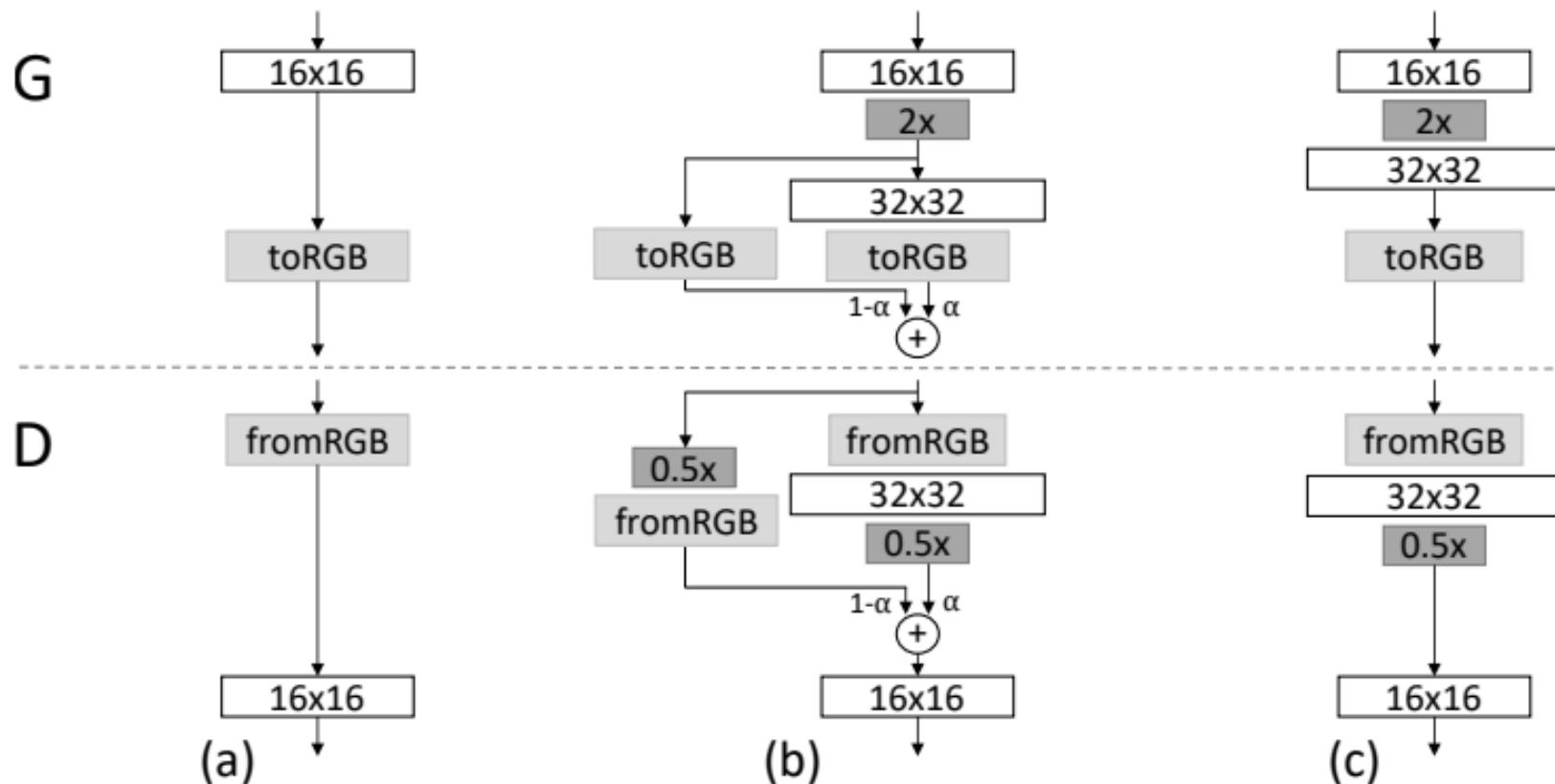
# Progressive Growing of GANs

Karras et al., Oct. 2017

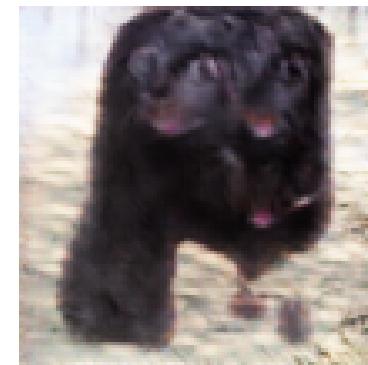


# Progressive Growing of GANs

Karras et al., Oct. 2017



# Early GANs on ImageNet



# Large Scale GAN Training

Brock et al., Sept. 2018



**Figure 1:** Class-conditional samples generated by our model.

This is a class-conditional GAN — it is conditioned on the imangenet class label.

This generates 512 X 512 images without using progressive training.

# Issues

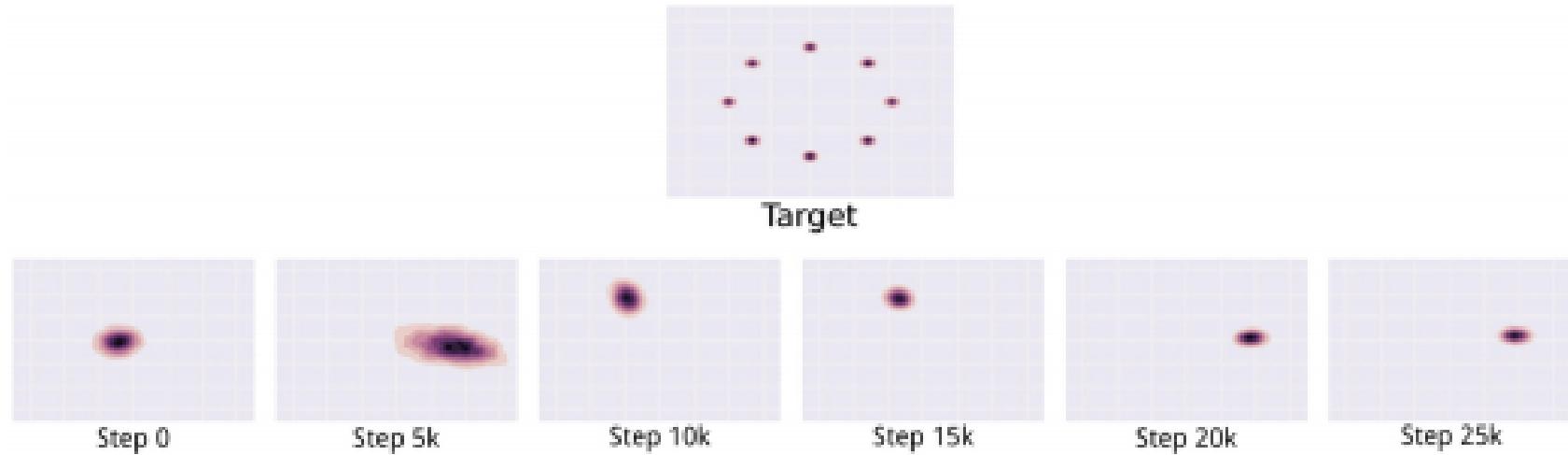
Mode Collapse

Unstable Training

Measuring Performance

# Mode Collapse a.k.a Mode Dropping

The generator distribution drops portions of the population.



## Unstable Training

Joint SGD is not the same as nested max-min.

Consider

$$\max_x \min_y xy$$

A Nash equilibrium is  $x = y = 0$ .

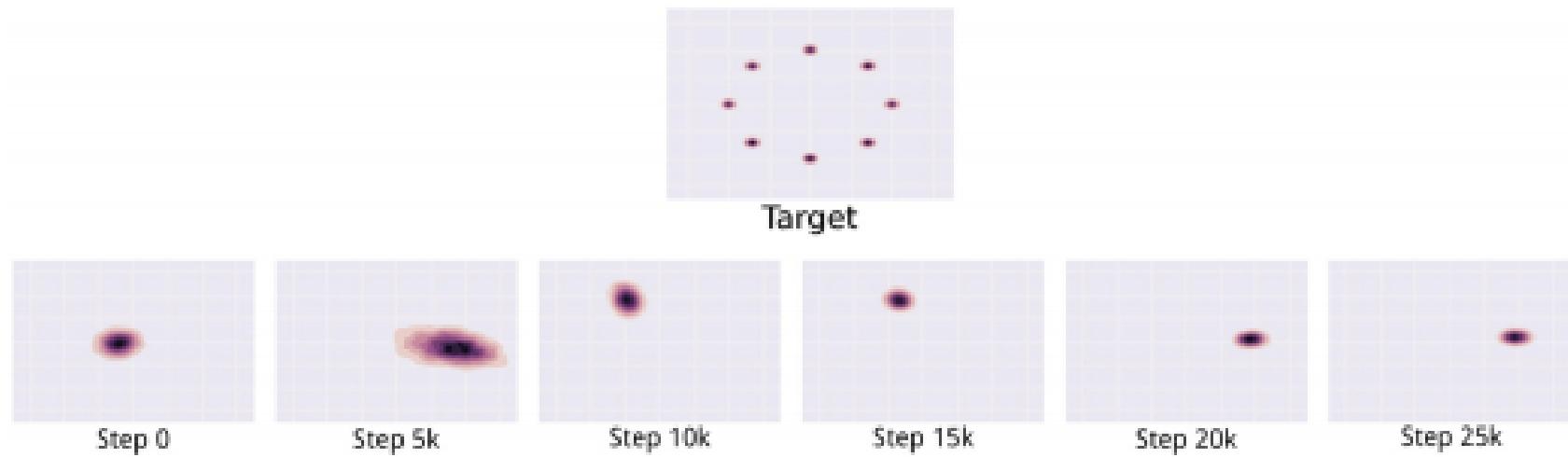
Simultaneous gradient flow yields

$$\frac{dx}{dt} = y \quad \frac{dy}{dt} = -x$$

This goes in a circle.

# Unstable Training

The generator distribution drifts as the discriminator follows.



# Pros and Cons of GAN Evaluation Measures

Borji, Oct 2018

We would like a rate-distortion metric on distribution models.

This has not yet been achieved for GANs.

Evaluation of GANs always involves, at least in part, subjective judgments of naturalness.

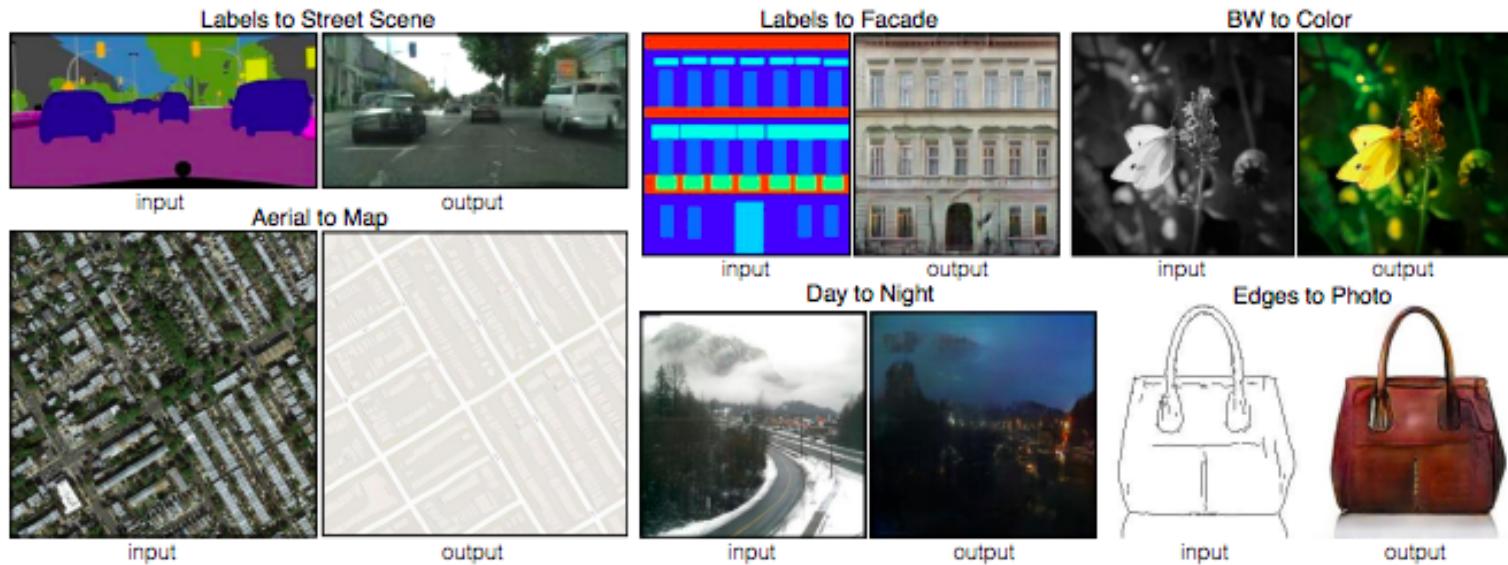
Sometimes automated metrics are also used.

The above paper discusses various proposed automated metrics of GAN performance. Current automated metrics are questionable.

# Image-to-Image Translation (Pix2Pix)

Isola et al., Nov. 2016

We assume a corpus of “image translation pairs” such as images paired with semantic segmentations.



## Conditional GANS

All unconditional distribution modeling methods apply to conditional distribution modeling.

Let  $y$  range over images. We have a generator  $p_\Phi$ . For  $i \in \{-1, 1\}$  we define a probability distribution over triple  $\langle x, y, i \rangle$  by

$$\begin{aligned}\tilde{p}_\Phi(i = 1) &= 1/2 \\ \tilde{p}_\Phi(y|i = 1) &= \text{pop}(y|x) \\ \tilde{p}_\Phi(y|i = -1) &= p_\Phi(y|x)\end{aligned}$$

We also have a discriminator  $P_\Psi(i|y)$ .

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \min_{\Psi} E_{\langle x, y, i \rangle \sim \tilde{p}_\Phi} - \ln P_\Psi(i|x, y)$$

if  $y$  real given  $x$

## Conditional GANs

$$\Phi^* = \operatorname{argmin}_{\Phi} \max_{\Psi} E_{x,y,i \sim \tilde{p}_{\Phi}(y|x)} \ln P_{\Psi}(i|x, y)$$

## Adversarial Discrimination as an Additional Loss

$$\Phi^* = \operatorname{argmin}_{\Phi} E_{(x,y) \sim \text{pop}} \|\cancel{x-y}\|^2 + \lambda \mathcal{L}_{\text{Discr}}(\Phi)$$

$\overset{\gamma - \gamma_x}{\cancel{x-y}}$

↓  
gold label

+ translation label

$$\mathcal{L}_{\text{Discr}}(\Phi) = \max_{\Psi} E_{x,y,i \sim \tilde{p}_{\Phi}} \ln P_{\Psi}(i|y, x)$$

## Discrimination as an Additional Loss

$$\text{L1 : } \Phi^* = \operatorname{argmin}_{\Phi} E_{(x,y) \sim \text{pop}} \|y - y_{\Phi}(x)\|_1$$

$$\text{cGAN : } \Phi^* = \operatorname{argmin}_{\Phi} \mathcal{L}_{\text{Discr}}(\Phi)$$

$$\text{L1 + cGAN : } \Phi^* = \operatorname{argmin}_{\Phi} E_{(x,y) \sim \text{pop}} \|y - y_{\Phi}(x)\|_1 + \lambda \mathcal{L}_{\text{Discr}}(\Phi)$$

# Image-to-Image Translation (Pix2Pix)

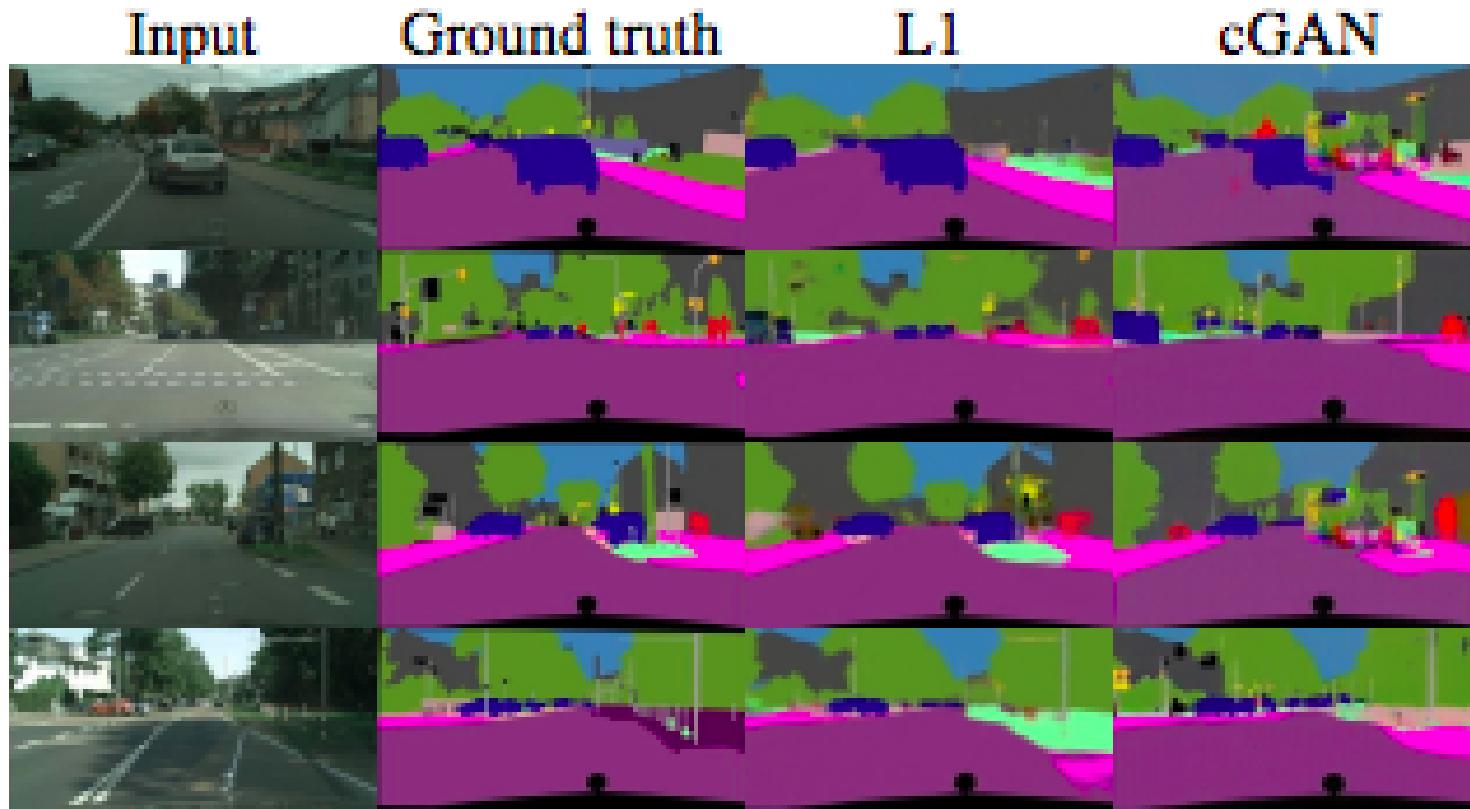
Isola et al., Nov. 2016



# Arial Photo to Map and Back

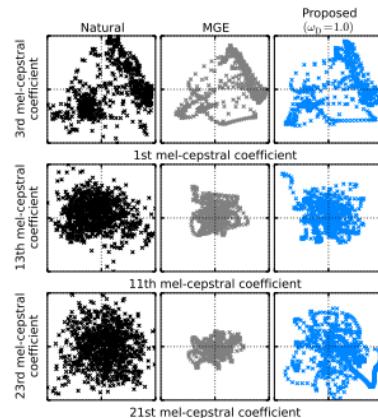


# Semantic Segmentation



# Feature Alignment by Discrimination

## Text to Speech (Saito et al. Sept. 2017)



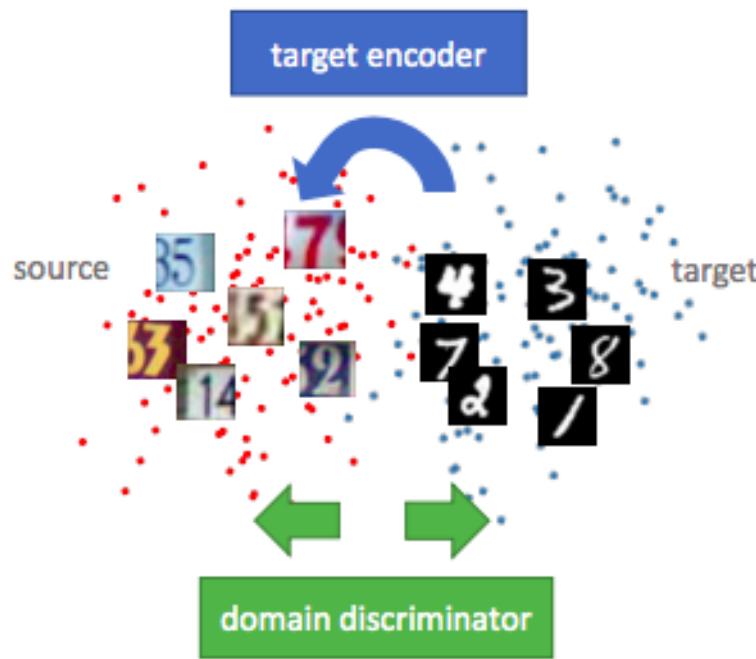
Minimum Generation Error (MGE) uses **perceptual distortion** — a distance between the feature vector of the generated sound wave and the feature vector of the original.

**Perceptual Naturalness** can be enforced by a feature discrimination loss.

$\psi$  tell which distribution come from

## Adversarial Discriminative Domain Adaptation

Tzeng et al. Feb. 2017



A feature discrimination loss can be used to align source and target features.

## Comments

I predict that in a few years adversarial discrimination will be limited to enforcing perceptual naturalness in the generation of sounds and images.

Cooperative discrimination seems more useful for predictive tasks. We will see that cooperative discrimination has been effective in pretraining.

**END**