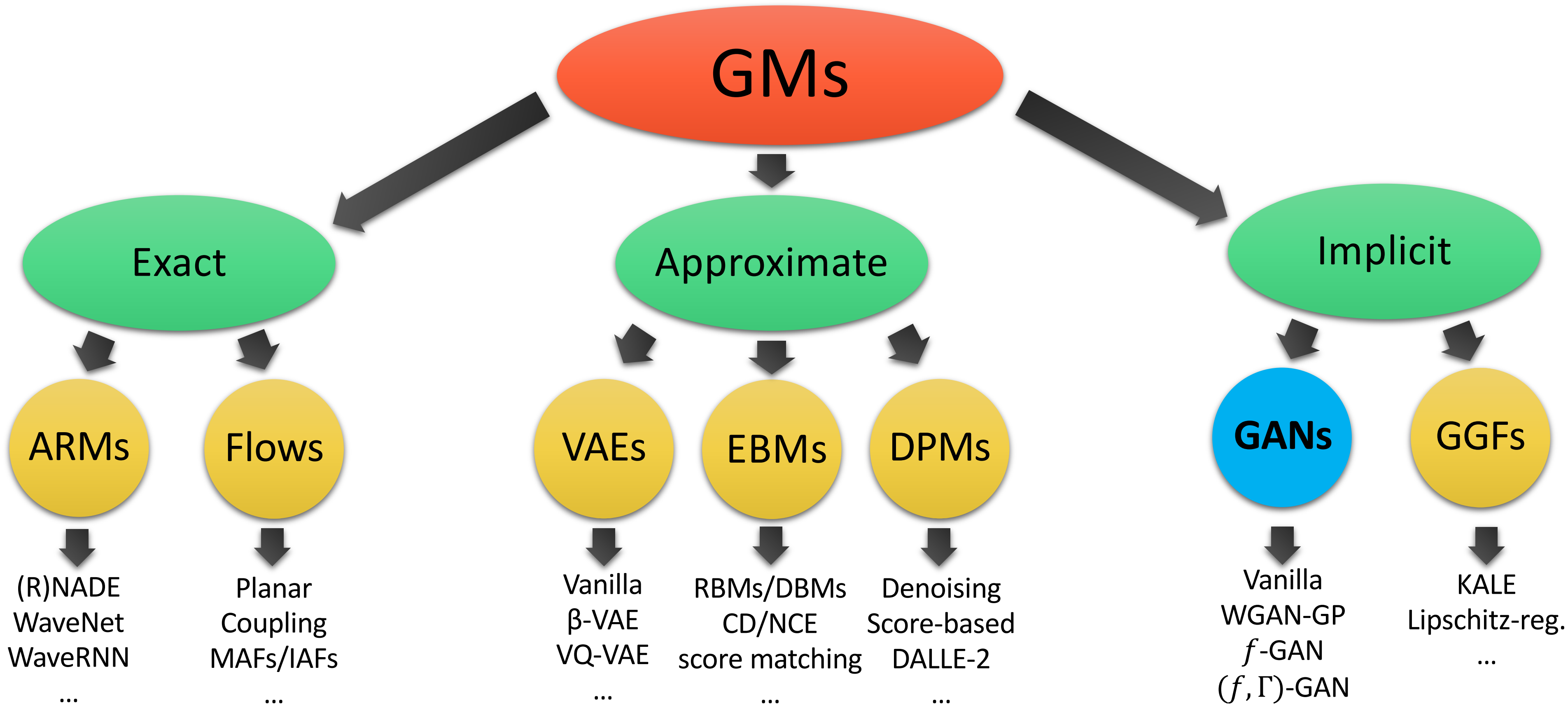
Introduction to Deep Generative Modeling

Lecture #17

HY-673 – Computer Science Dep., University of Crete
Professors: Yannis Pantazis, Yannis Stylianou
Teaching Assistant: Michail Raptakis

Taxonomy of GMs

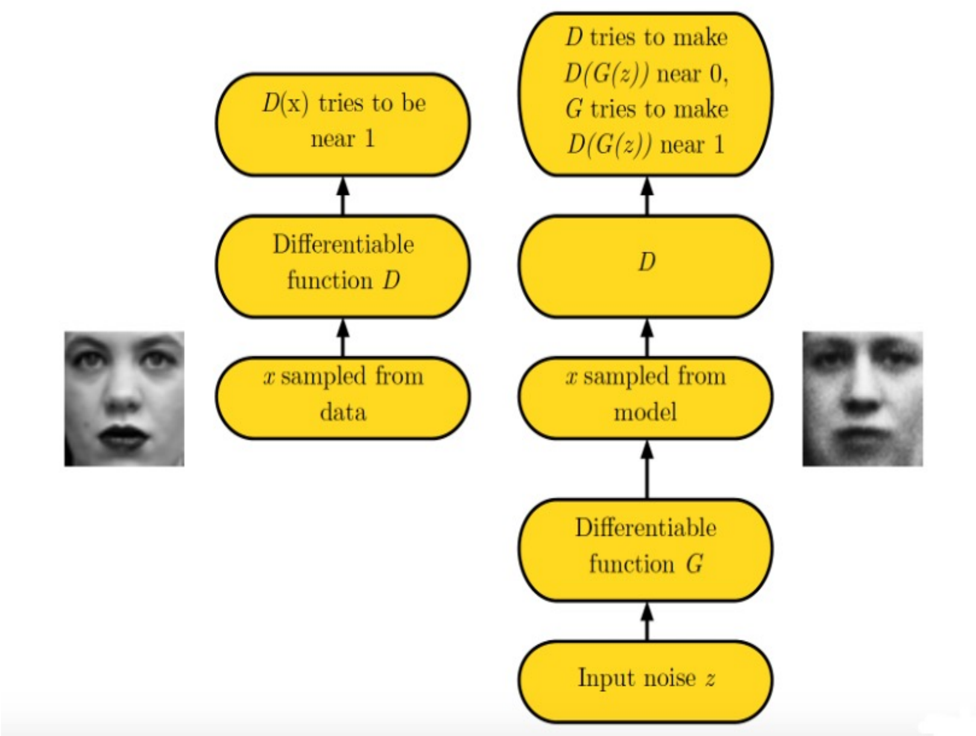


Recap - GANs

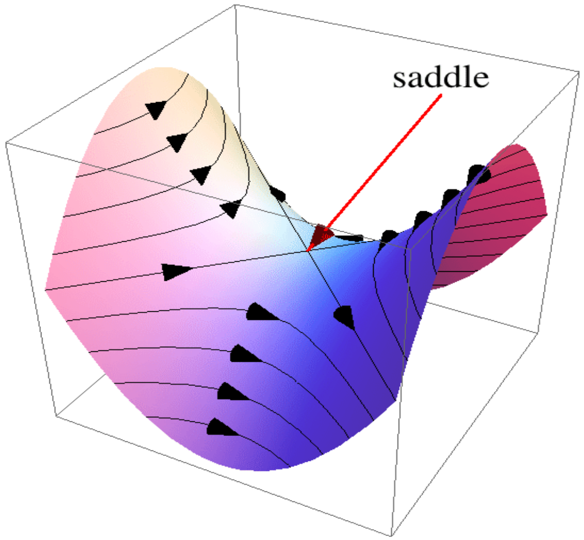
Lecture #17

- Training objective for both generator and discriminator:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p_Z}[\log(1 - D(G(x)))].$$



The joint optimum (G^*, D^*) is a saddle point.



Recap - GANs

- With the optimal discriminator D_G^*, we can see that a GAN minimizes a scaled and shifted Jensen-Shannon divergence:

$$\min_G 2D_{\mathrm{JSD}}\left[p_{\mathrm{data}}, p_G\right] - \log 4.$$

- Parametrize D by ϕ and G by θ. Prior distribution $p(z)$:

$$\min_\theta \max_\phi \mathbb{E}_{x\sim p_{\mathrm{data}}}\left[\log D_\phi(x)\right] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D_\phi(G_\theta(z))\right)\right].$$

- Likelihood-free training.

Recap - GAN Training Algorithm

- Sample minibatch of m training points $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ from $\mathcal{D}$
- Sample minibatch of m noise vectors $z^{(1)}, z^{(2)}, \ldots, z^{(m)}$ from p_Z
- Update the discriminator parameters ϕ by stochastic gradient **ascent**

$$\nabla_\phi V(G_\theta, D_\phi) = \frac{1}{m} \nabla_\phi \sum_{i=1}^{m} \left[\log D_\phi(x^{(i)}) + \log(1 - D_\phi(G_\theta(z^{(i)}))) \right].$$

- Update the generator parameters θ by stochastic gradient **descent**

$$\nabla_\theta V(G_\theta, D_\phi) = \frac{1}{m} \nabla_\theta \sum_{i=1}^{m} \log(1 - D_\phi(G_\theta(z^{(i)}))).$$

- Repeat for fixed number of iterations

Recap – Optimization Challenges

Lecture #17

- *Optimization instabilities:* the generator and discriminator loss keeps oscillating during GAN training; no stopping criterion in practice

- *Mode collapse:* the generator of a GAN collapses to one of few samples (dubbed as "modes")

- *Evaluation criteria:* no analog to log-lihelihood; has to define "new" metrics such as Inception Score (IS) and Frenchel Inception Distance (FID) for image generation

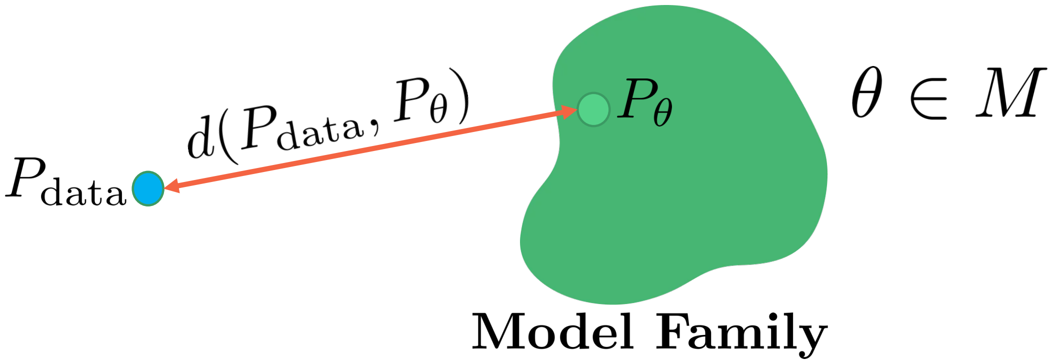
Today's Plan

- Rich class of likelihood-free objectives via f-GANS.
- Wasserstein GAN.
- Inferring latent representations via BiGAN.
- Application: Unpaired image-to-image translation via CycleGANs.

The GAN Zoo (list of all named GANs):
https://github.com/hindupuravinash/the-gan-zoo

Beyond KL and Jensen-Shannon

$x_i \sim P_{\text{data}}$
$i = 1, 2, \ldots, n$



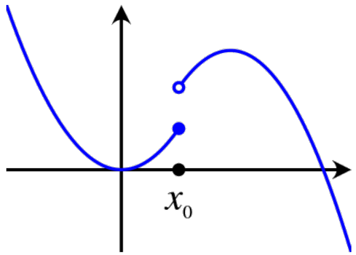
- What choices do we have for $d(\cdot, \cdot)$?
  - KL divergence: Autoregressive Models, Flow models.
  - Jensen-Shannon Divergence (scaled and shifted): Original GAN objective.

f - Divergences

- Given two densities p and q, the f-divergence is given by:

$$D_f(p||q) := \mathbb{E}_{x\sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] = \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx,$$

  where f is any convex, lower-semicontinuous function with $f(1) = 0$.



- Convex: Line joining any two points is above the function.
- Lower-semicontinuous: Function value at any point x_0 is close to $f(x_0)$ or greater than $f(x_0)$.
- Jensen's Inequality: $\mathbb{E}_{x\sim q}\left[f(p(x)/q(x))\right] \geq f\left(\mathbb{E}_{x\sim q}[p(x)/q(x)]\right) = f(1) = 0.$
- Example: KL divergence with $f(u) = u\log u$.

f - Divergences

- Many more f-divergences!

Name	$D_f(P\|Q)$	Generator $f(u)$
Total variation	$\frac{1}{2}\int \|p(x)-q(x)\|\,\mathrm{d}x$	$\frac{1}{2}\|u-1\|$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$
Reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$
Neyman χ^2	$\int \frac{(p(x)-q(x))^2}{q(x)}\,\mathrm{d}x$	$\frac{(1-u)^2}{u}$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\,\mathrm{d}x$	$(\sqrt{u}-1)^2$
Jeffrey	$\int (p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)\,\mathrm{d}x$	$(u-1)\log u$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2}+u\log u$
Jensen-Shannon-weighted	$\int p(x)\pi\log\frac{p(x)}{\pi p(x)+(1-\pi)q(x)}+(1-\pi)q(x)\log\frac{q(x)}{\pi p(x)+(1-\pi)q(x)}\,\mathrm{d}x$	$\pi u\log u-(1-\pi+\pi u)\log(1-\pi+\pi u)$
GAN	$\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x-\log(4)$	$u\log u-(u+1)\log(u+1)$
α-divergence ($\alpha\notin\{0,1\}$)	$\frac{1}{\alpha(\alpha-1)}\int\left(p(x)\left[\left(\frac{q(x)}{p(x)}\right)^{\alpha}-1\right]-\alpha(q(x)-p(x))\right)\,\mathrm{d}x$	$\frac{1}{\alpha(\alpha-1)}\left(u^{\alpha}-1-\alpha(u-1)\right)$

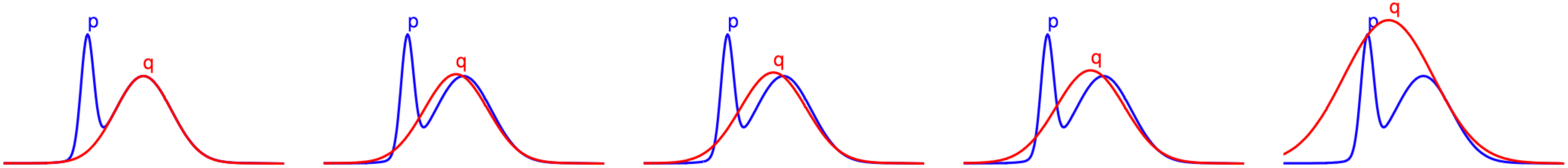
α - divergence: Mode covering vs mode seeking

Lecture #17

- α-divergence:

$$D_\alpha(p||q) := \frac{1}{\alpha(1-\alpha)} \int \alpha p(x) + (1-\alpha) q(x) + p(x)^\alpha q(x)^{1-\alpha} dx$$

- $D_\alpha(p||q) = D_{1-\alpha}(q||p)$



$\alpha = -\infty$ $\alpha = 0$ $\alpha = 0.5$ $\alpha = 1$ $\alpha = \infty$

f - GAN: Variational Divergence Minimization

- To use f-divergences as a two-sample test objective for likelihood-free learning, we need to be able to estimate it only via samples.
- Fenchel conjugate: For any function $f(\cdot)$, its convex conjugate is defined as:

$$f^*(t) = \sup_{u \in \text{dom}_f} (ut - f(u)).$$

- $f^{**} \leq f$.
- f^* is always convex and lower semi-continuous.
- Duality: $f^{**} = f$ when $f(\cdot)$ is convex, lower semi-continuous. Equivalently:

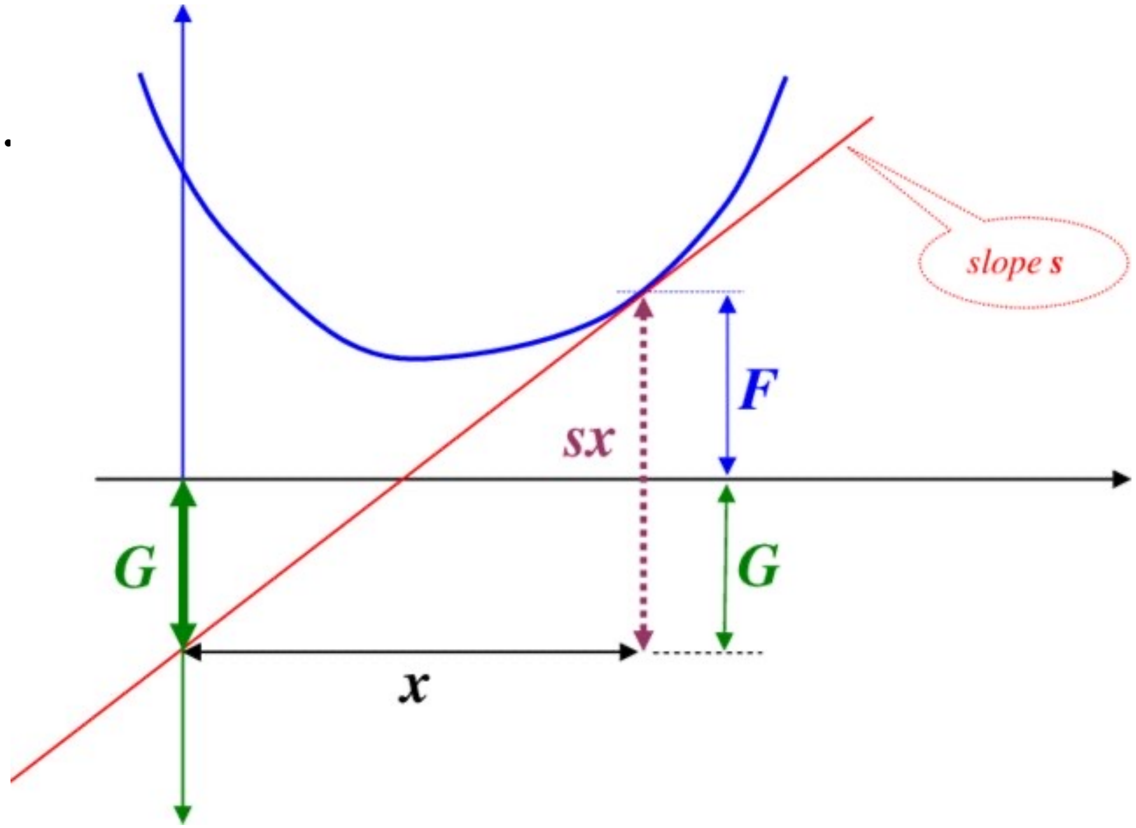
$$f(u) = f^{**}(u) = \sup_{t \in \text{dom}_{f^*}} (tu - f^*(t)).$$

f - GAN: Variational Divergence Minimization

Lecture #17

- Fenchel conjugate (a.k.a. Legendre transform):

$$f^*(t) = \sup_{u \in \mathrm{dom}_f}(ut - f(u)).$$



f - GAN: Variational Divergence Minimization

- We can obtain a lower bound to any f-divergence via its Fenchel conjugate:

$$
\begin{aligned}
D_f(p||q) &= \mathbb{E}_{x\sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right] \\
&= \mathbb{E}_{x\sim q}\left[\sup_{t\in \mathrm{dom}_{f^*}}\left(t\frac{p(x)}{q(x)} - f^*(t)\right)\right] \\
&:= \mathbb{E}_{x\sim q}\left[T^*(x)\frac{p(x)}{q(x)} - f^*\left(T^*(x)\right)\right] \\
&= \int_{\mathcal{X}} \left[T^*(x)p(x) - f^*\left(T(x)\right)q(x)\right]dx
\end{aligned}
$$

f - GAN: Variational Divergence Minimization

- We can obtain a lower bound to any f-divergence via its Fenchel conjugate:

$$
\begin{aligned}
D_f(p||q) &= \sup_{T} \int_{\mathcal{X}} \left[T(x)p(x) - f^*\left(T(x)\right)q(x)\right] dx \\
&\geq \sup_{T \in \mathcal{T}} \int_{\mathcal{X}} \left(T(x)p(x) - f^*\left(T(x)\right)q(x)\right) dx \\
&= \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim p}[T(x)] - \mathbb{E}_{x \sim q}[f^*(T(x))]\right),
\end{aligned}
$$

where $\mathcal{T} = \{T : \mathcal{X} \to \mathbb{R}\}$ is an arbitrary class of functions.

- **Note**: Lower bound is likelihood free w.r.t. p and q.

f - GAN: Variational Divergence Minimization

Lecture #17

- Variational lower bound:

$$D_f(p||q) \geq \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{x \sim p}[T(x)] - \mathbb{E}_{x \sim q}\left[f^*(T(x))\right] \right).$$

- Choose any f-divergence.
- Let $p = p_{\text{data}}$ and $q = p_G$.
- Parametrize T by ϕ and G by θ.

f - GAN: Variational Divergence Minimization

- Consider the following f-GAN objective:

$$\min_{\theta} \max_{\phi} F(\theta, \phi) = \mathbb{E}_{x \sim p_{\text{data}}}[T_\phi(x)] - \mathbb{E}_{x \sim p_{G_\theta}}\left[f^*(T_\phi(x))\right].$$

- Generator G_θ tries to minimize the divergence estimate.
- Discriminator T_ϕ tries to tighten the lower bound.
- Substitute any f-divergence and optimize the f-GAN objective.

Wasserstein GAN: Beyond f-Divergences

- The f-divergence is defined as:

$$D_f(p||q) = \mathbb{E}_{x\sim q}\left[f\left(\frac{p(x)}{q(x)}\right)\right].$$

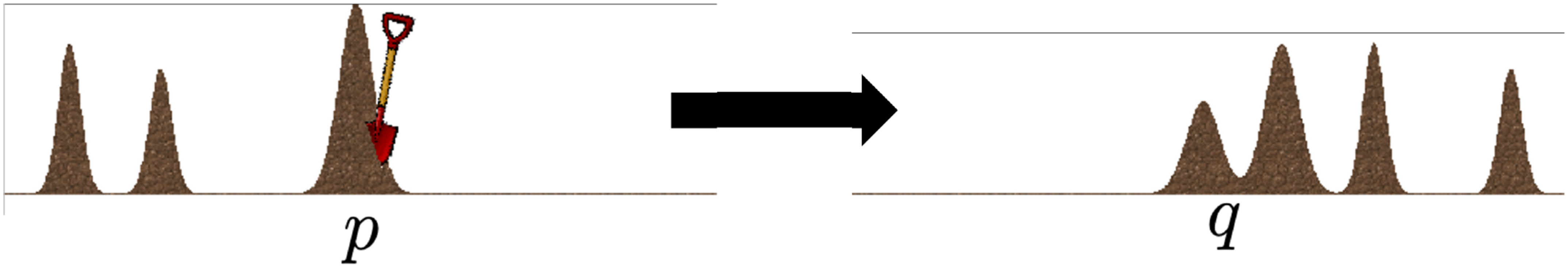
- The support of q has to cover the support of p, otherwise infinity arises in f-divergences.

Let $p(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0) \end{cases}$, and $q_\theta(x) = \begin{cases} 1, & x = \theta \\ 0, & x \neq \theta \end{cases}$, then :

$$D_{\text{KL}}(p, p_\theta) = \begin{cases} 0, & \theta = 0 \\ \infty, & \theta \neq 0 \end{cases} , \quad D_{\text{JS}}(p, q_\theta) = \begin{cases} 0, & \theta = 0 \\ \log 2, & \theta \neq 0 \end{cases} .$$

- We need a "smoother" distance $D(p, q)$ that is defined when p and q have disjoint supports

Wasserstein (Earth-Mover) Distance



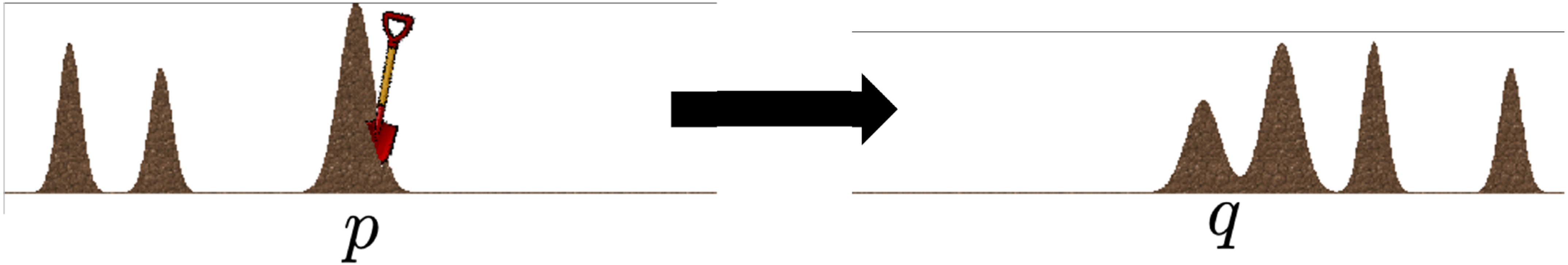
- Wasserstein distance (of order 1):

$$D_W(p,q) = \inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{(x,y)\sim\gamma}\left[||x-y||_1\right],$$

  where $\Pi(p,q)$ contains all joint distributions of (x,y) where the marginal of x is $p(x)$, and the marginal of y is $q(y)$.

- $\gamma(y|x)$: a probabilistric earth moving plan that warps $p(x)$ to $q(y)$.

Wasserstein (Earth-Mover) Distance



- Wasserstein distance:

$$D_W(p, q) = \inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{(x,y)\sim\gamma} \left[||x - y||_1 \right],$$

$$\text{Let } p(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0) \end{cases}, \text{ and } q_\theta(x) = \begin{cases} 1, & x = \theta \\ 0, & x \neq \theta \end{cases}, \text{ then :}$$

- $D_W(p, q_\theta) = |\theta|$.

Wasserstein GAN (WGAN)

- Kantorovich-Rubinstein duality:

$$D_W(p,q) = \sup_{||g||_L \leq 1} \mathbb{E}_{x\sim p}[g(x)] - \mathbb{E}_{x\sim q}[g(x)],$$

where $||g||_L \leq 1$ means the Lipschitz constant of $g(x)$ is 1. Technically:

$$\forall x, y: \ |g(x) - g(y)| \leq ||x - y||.$$

- WGAN with discriminator $D_\phi(x)$ and generator $G_\theta(z)$:

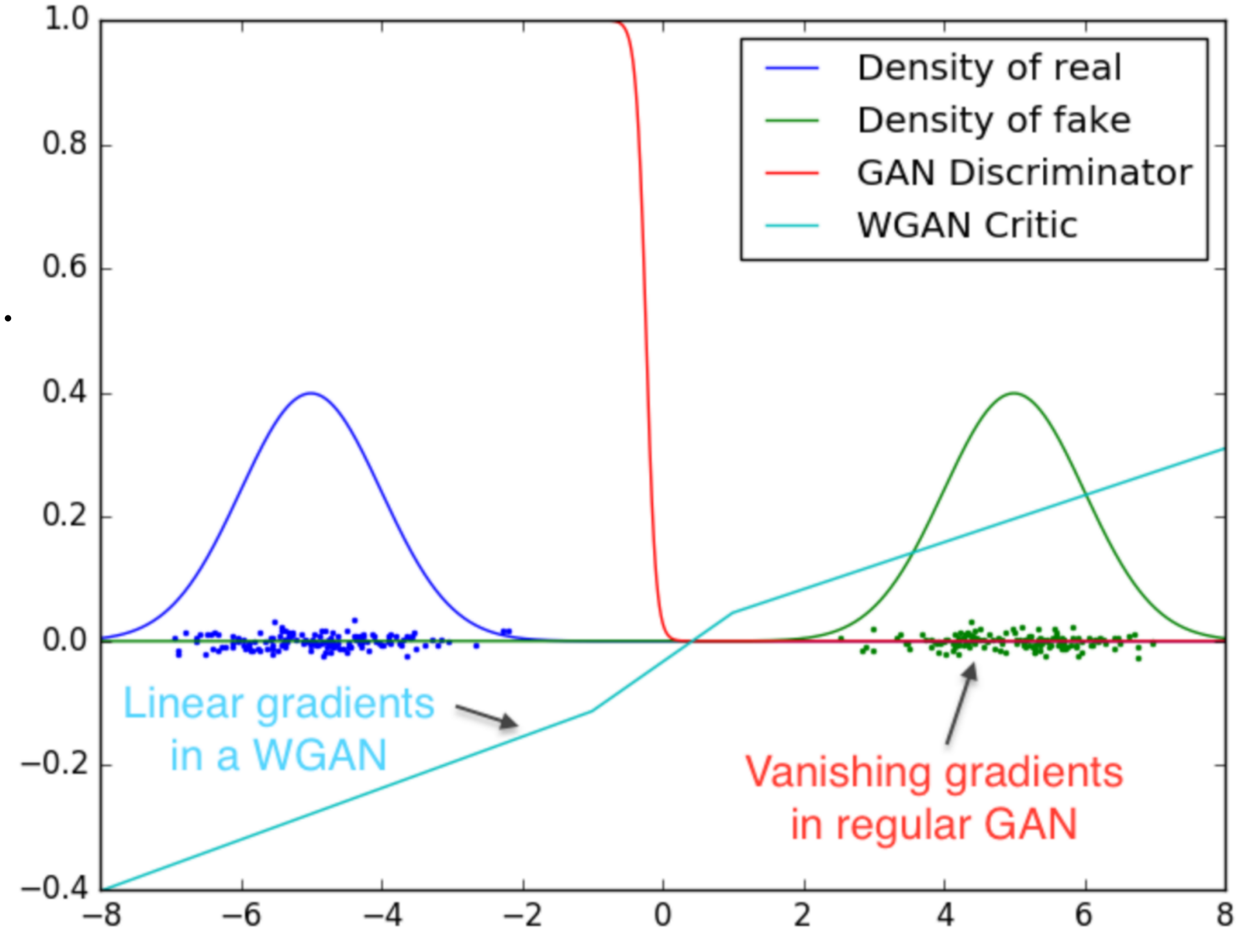
$$\min_\theta \max_\phi \mathbb{E}_{x\sim p_{\text{data}}}[D_\phi(x)] - \mathbb{E}_{z\sim p(z)}[D_\phi(G_\theta(z))].$$

- Lipschitzness of $D_\phi(x)$ is enforced through weight clipping or gradient penalty.

Wasserstein GAN (WGAN)

Lecture #17

- More training stability.
- Less mode collapse.
- Via discriminator constraining.



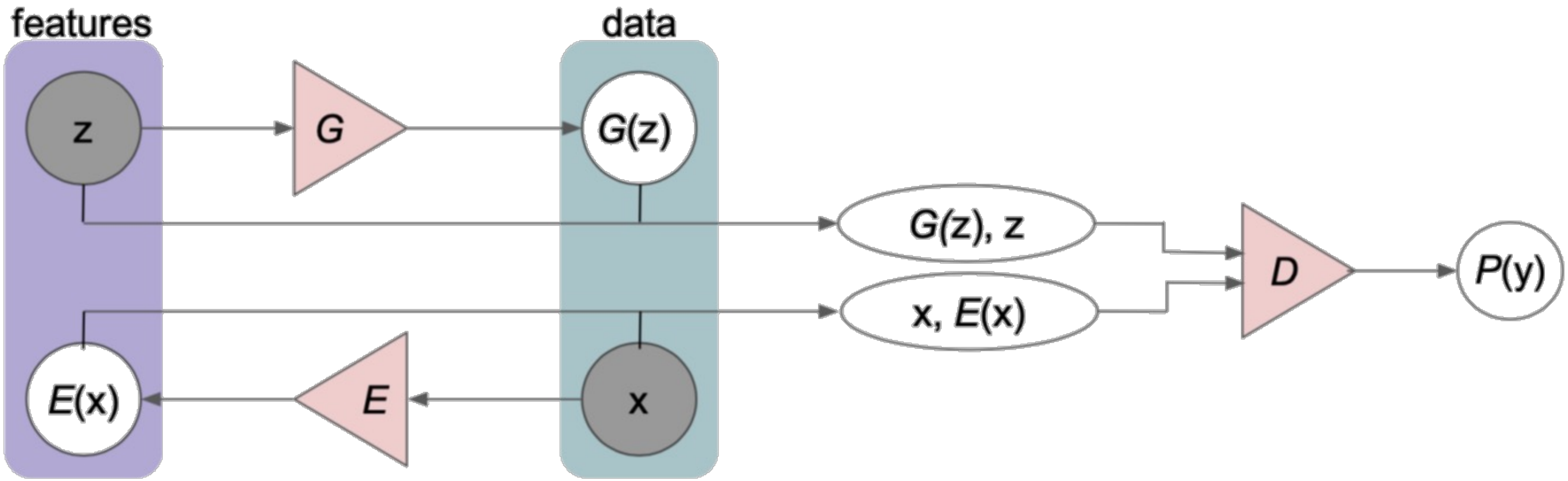
Inferring Latent Representations in GANs

- The generator of a GAN is typically a directed, latent variable model with latent variables z and observed variables x. How can we infer the latent feature representations in a GAN?
- Unlike a normalizing flow model, the mapping $G : z \to x$ is not necessarily invertible.
- Unlike a variational autoencoder, there is no inference network $q(\cdot|x)$ which can learn a variational posterior over latent variables.
- **Solution 1**: For any point x, use the activations of the prefinal layer of a discriminator as a feature representation.
- Intuition: Similar to supervised deep neural networks, the discriminator would have learned useful representations for x while distinguishing real and fake x's.

Inferring Latent Representations in GANs

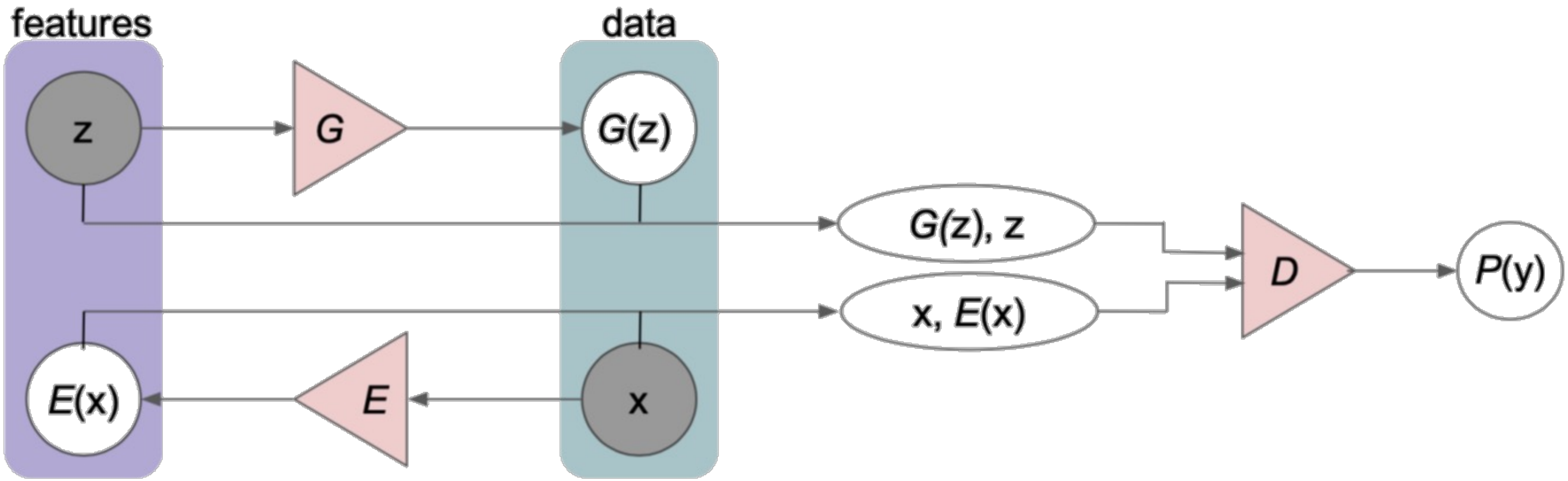
- If we want to directly infer the latent variables z of the generator, we need a different learning algorithm.
- A regular GAN optimizes a two-sample test objective that compares samples of x from the generator and the data distribution.
- **Solution 2**: To infer latent representations, we will compare samples of (x, z) from the joint distributions of observed and latent variables as per the model and the data distribution.
- For any x generated via the model, we have access to z (sampled from a simple prior $p(z)$).
- For any x from the data distribution, the z is however unobserved (latent).

Bidirectional Generative Adversarial Networks (BiGANs)



- In a BiGAN, we introduce an encoder network E in addition to the generator network G.
- E only observes $x \sim p_{\text{data}}(x)$ during training to learn a mapping $E : x \to z$.
- As before, G only observes the samples from the prior $z \sim p(z)$ during training to learn a mapping $G : z \to x$.

Bidirectional Generative Adversarial Networks (BiGANs)

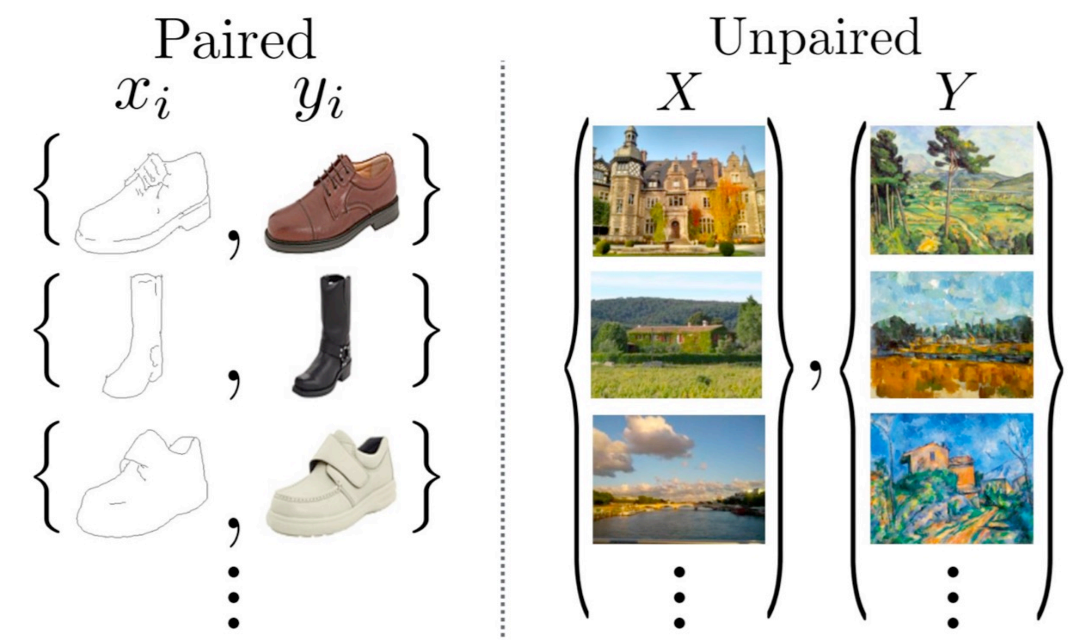


- D observes samples from G, i.e., $(z, G(z))$ pairs, and from the encoding distribution $(E(x), x)$.
- The goal of D is to maximize the two-sample test objective between $(z, G(z))$, and $(E(x), x)$.
- After training is complete, new samples are generated via G and latent representations are inferred via E.

Translating Across Domains

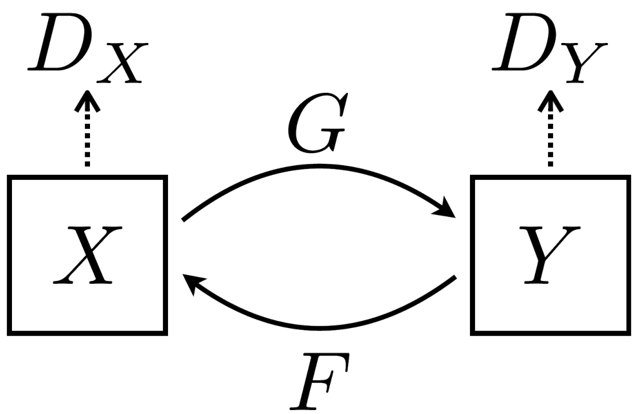
Lecture #17

- Image-to-image translation: We are given images from two domains, $\mathcal{X}$ and $\mathcal{Y}$.
- Paired vs. unpaired examples:
- Paired examples can be expensive to obtain. Can we translate from $\mathcal{X} \leftrightarrow \mathcal{Y}$ in an unsupervised manner?



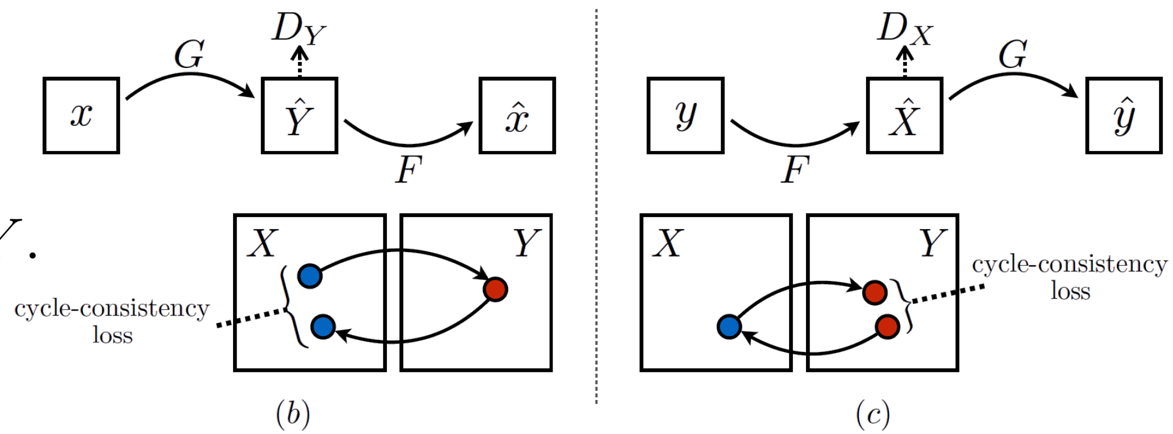
CycleGAN: Adversarial Training Across Two Domains

- To match the two distributions, we learn two parameterized conditional generative models $G : \mathcal{X} \leftrightarrow \mathcal{Y}$ and $F : \mathcal{Y} \leftrightarrow \mathcal{X}$
- G maps an element of $\mathcal{X}$ to an element of $\mathcal{Y}$. A discriminator $D_{\mathcal{Y}}$ compares the observed dataset Y and the generated samples $\hat{Y} = G(X)$.
- Similarly, F maps an element of $\mathcal{Y}$ to an element of $\mathcal{X}$. A discriminator $D_{\mathcal{X}}$ compares the observed dataset X and the generated samples $\hat{X} = F(Y)$.



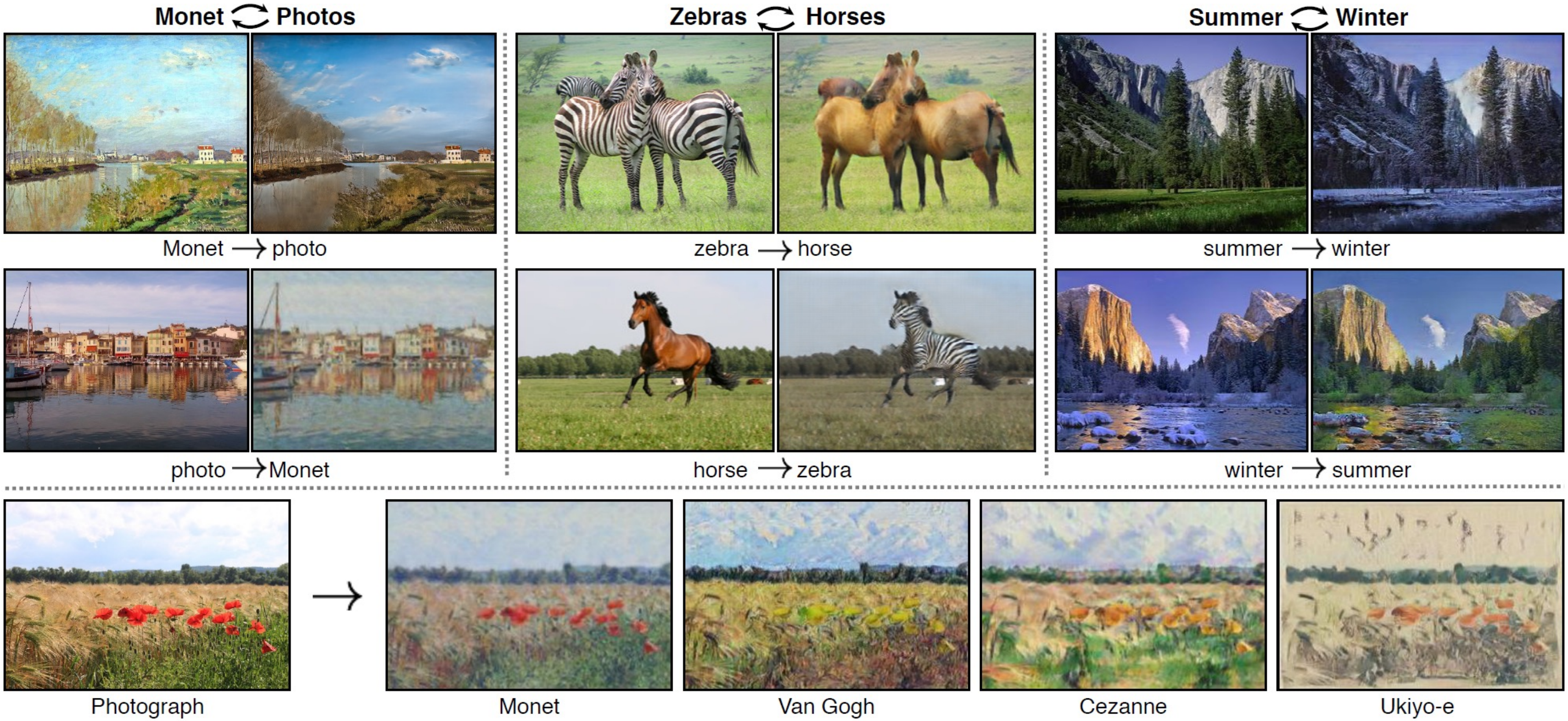
CycleGAN: Cycle Consistency Across Domains

- **Cycle Consistency:** If we can go form X to $\hat{Y}$ via G, then it should also be possible to go from $\hat{Y}$ back to X via F:
  - $F(G(X)) \approx X$.
  - Similarly, vice versa: $G(F(Y)) \approx Y$.
  - Overall loss function:



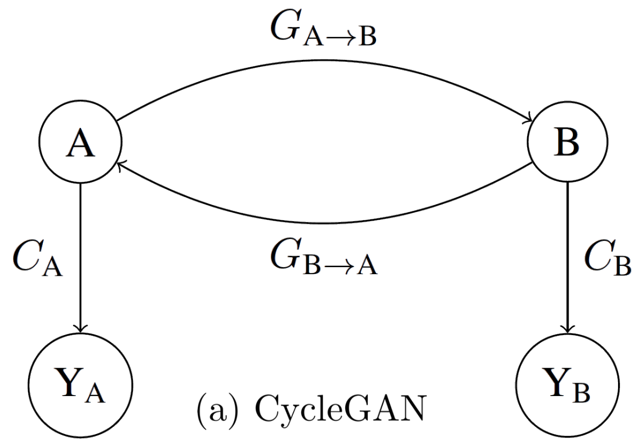
$$\min_{F,G,D_{\mathcal{X}},D_{\mathcal{Y}}} \mathcal{L}_{\text{GAN}}(G, D_{\mathcal{Y}}, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_{\mathcal{X}}, X, Y)$$
$$+ \underbrace{\lambda \left(\mathbb{E}_X \left[||F(G(X)) - X||_1 \right] + \mathbb{E}_Y \left[||G(F(Y)) - Y||_1 \right] \right)}_{\text{cycle consistency}}.$$

CycleGAN in Practice



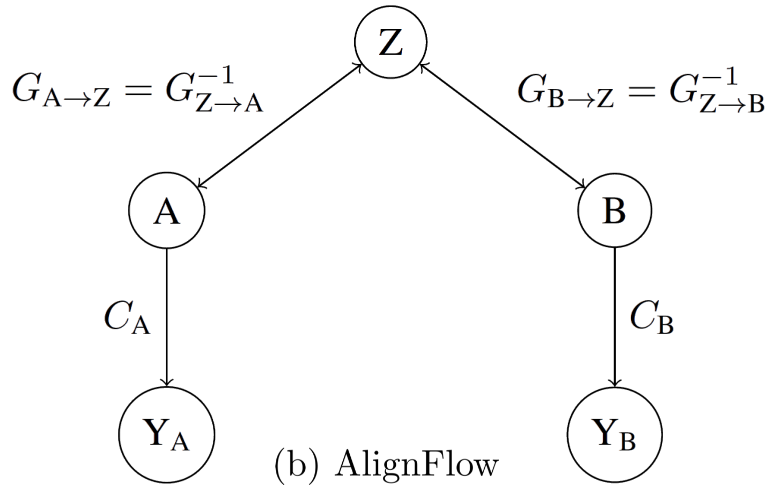
AlignFlow (Grover et al.)

- What if G is a flow model?
- No need to parametrize F separately: $F = G^{-1}$.
- Can train via MLE and/or adversarial learning.
- Exactly cycle consistent: $F(G(X)) = X$, $G(F(Y)) = Y$.



(a) CycleGAN

- Unlike CycleGAN, AlignFlow specifies a single invertible mapping $G_{A\to Z} \circ G^{-1}_{B\to Z}$ that is exactly cycle-consistent, represents a shared latent space Z between the two domains, and can be trained via both adversarial training and exact MLE. Doubleheaded arrows denote invertible mappings. Y_A and Y_B are r.v.s denoting the output of the critics used for adversarial training.

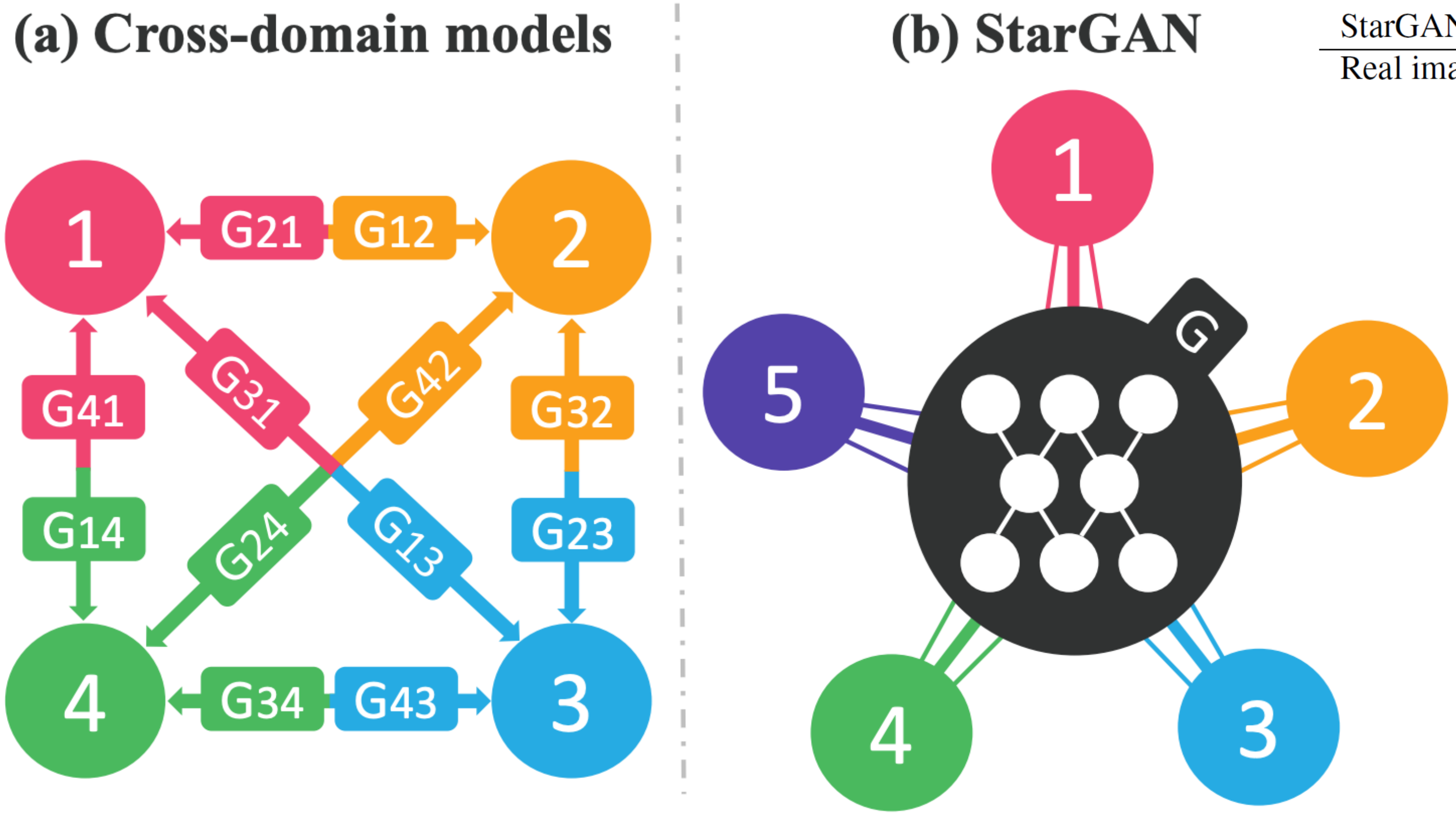


(b) AlignFlow

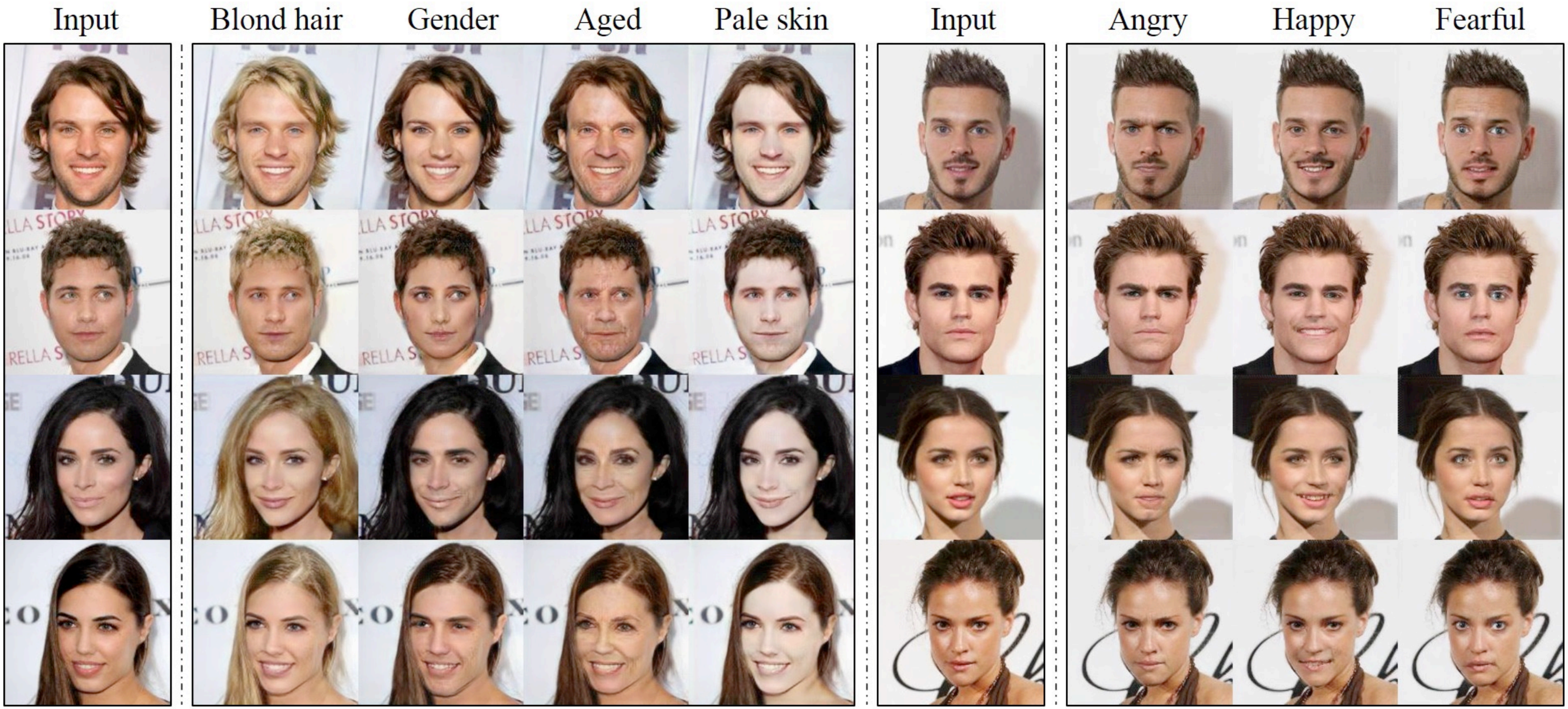
StarGAN (Choi et al.)

- What if there are multiple domains?

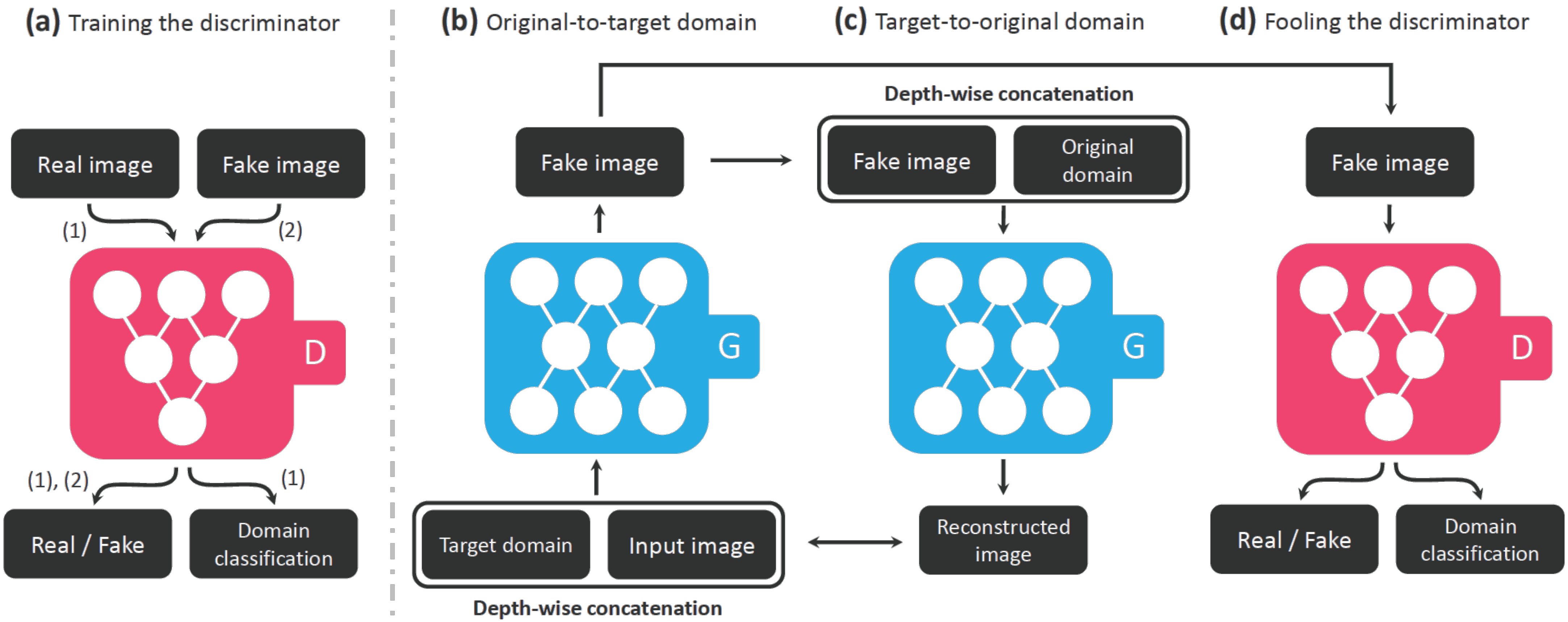
Method	Classification error	# of parameters
DIAT	4.10	52.6M × 7
CycleGAN	5.99	52.6M × 14
IcGAN	8.07	67.8M × 1
StarGAN	**2.12**	**53.2M** × **1**
Real images	0.45	-



StarGAN (Choi et al.)



StarGAN (Choi et al.)



Summary of GANs

- Key observation: Samples and likelihoods are not correlated in practice.
- Two-sample test objectives allow for learning generative models only via samples (likelihood-free).
- Wide range of two-sample test objectives covering f-divergences and Wasserstein distances (and more).
- Latent representations can be inferred via BiGAN.
- Cycle-consistent domain translations via CycleGAN and StarGAN.

References

1. https://deepgenerativemodels.github.io
2. Nowozin et al.,"f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization", 2016, NeurIPS.
3. Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", 2019, AAAI Conference on Artificial Intelligence 34(04):4028-4035.
4. Choi et al., "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation", 2017, IEEE Conference on Computer Vision and Pattern Recognition.

Introduction to Deep Generative Modeling

Lecture #17

HY-673 – Computer Science Dep., University of Crete

Professors: Yannis Pantazis, Yannis Stylianou

Teaching Assistant: Michail Raptakis