上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Lecture 27: Generative Adversarial Networks

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

Which of The Faces are Fake?

ALL faces are generated by GAN.

StyleGAN3 faces from thispersondoesnotexist.com

By the end of this lecture, you will know how a network learned to do this — without ever being told what a face looks like.

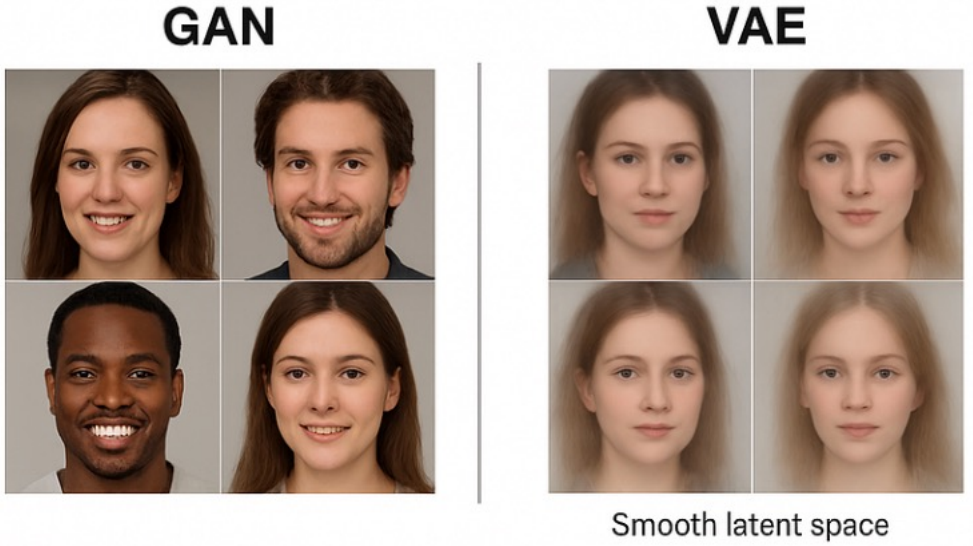
Where we left off

L23 — Autoregressive

- exact likelihood, sequential sampling
- no z, no semantic latent

L24 — VAE

- latent z — but blurry samples
- Gaussian decoder + recognition gap
  both push toward the mean
- *Maximum likelihood rewards covering all modes.*
- *Photorealism rewards committing to one.*



Can we remove likelihood objective?

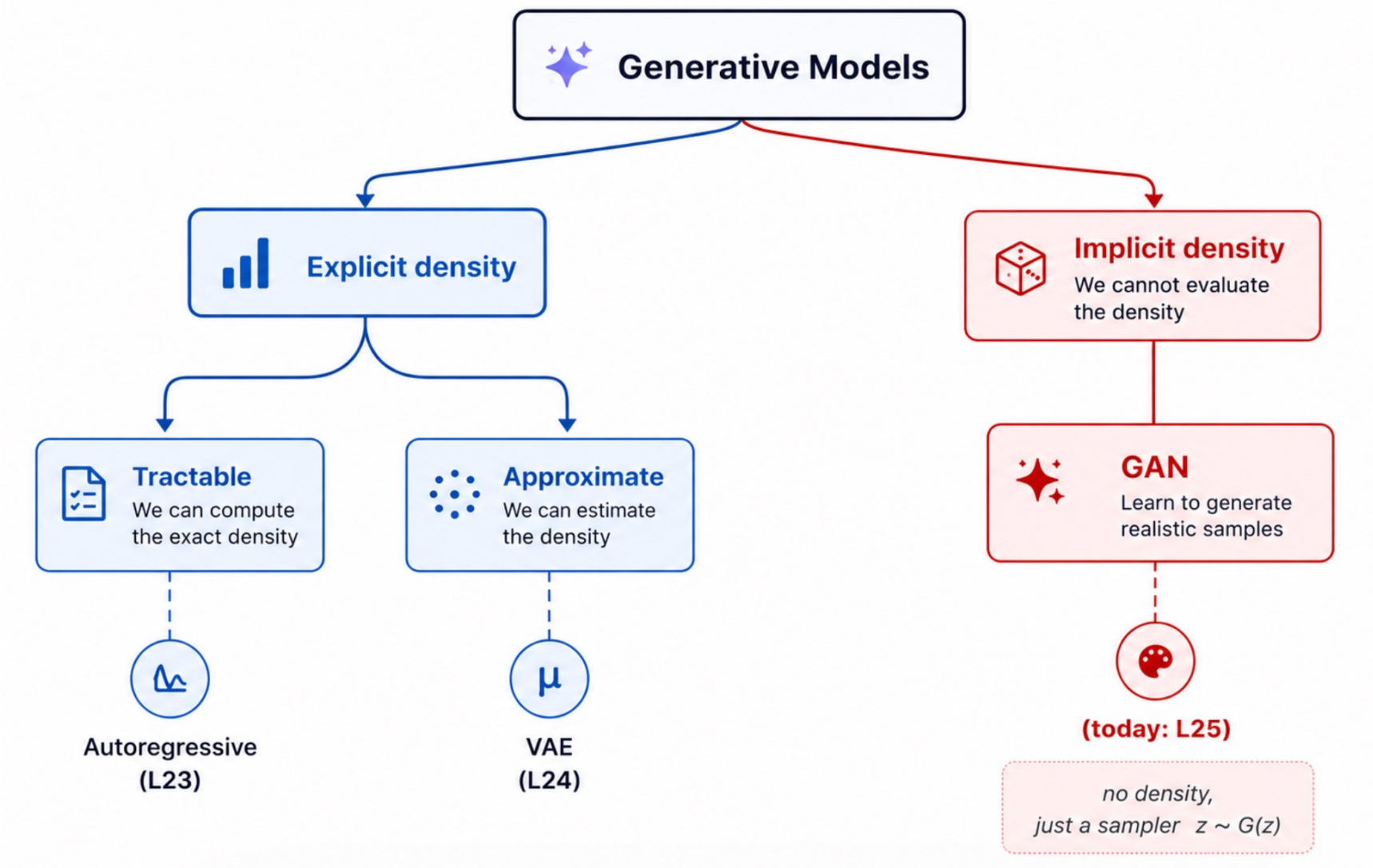
Objectives

After this lecture, you will be able to:

- **Explain** the minimax game between generator and discriminator, and **derive** the optimal discriminator in closed form.
- **Show** that the global GAN equilibrium minimizes the **Jensen–Shannon (JS) divergence** between p_{data} and p_g.
- **Diagnose** the three classical GAN failure modes — vanishing gradients, mode collapse, training oscillation — and **evaluate** which architectural fix targets which failure.

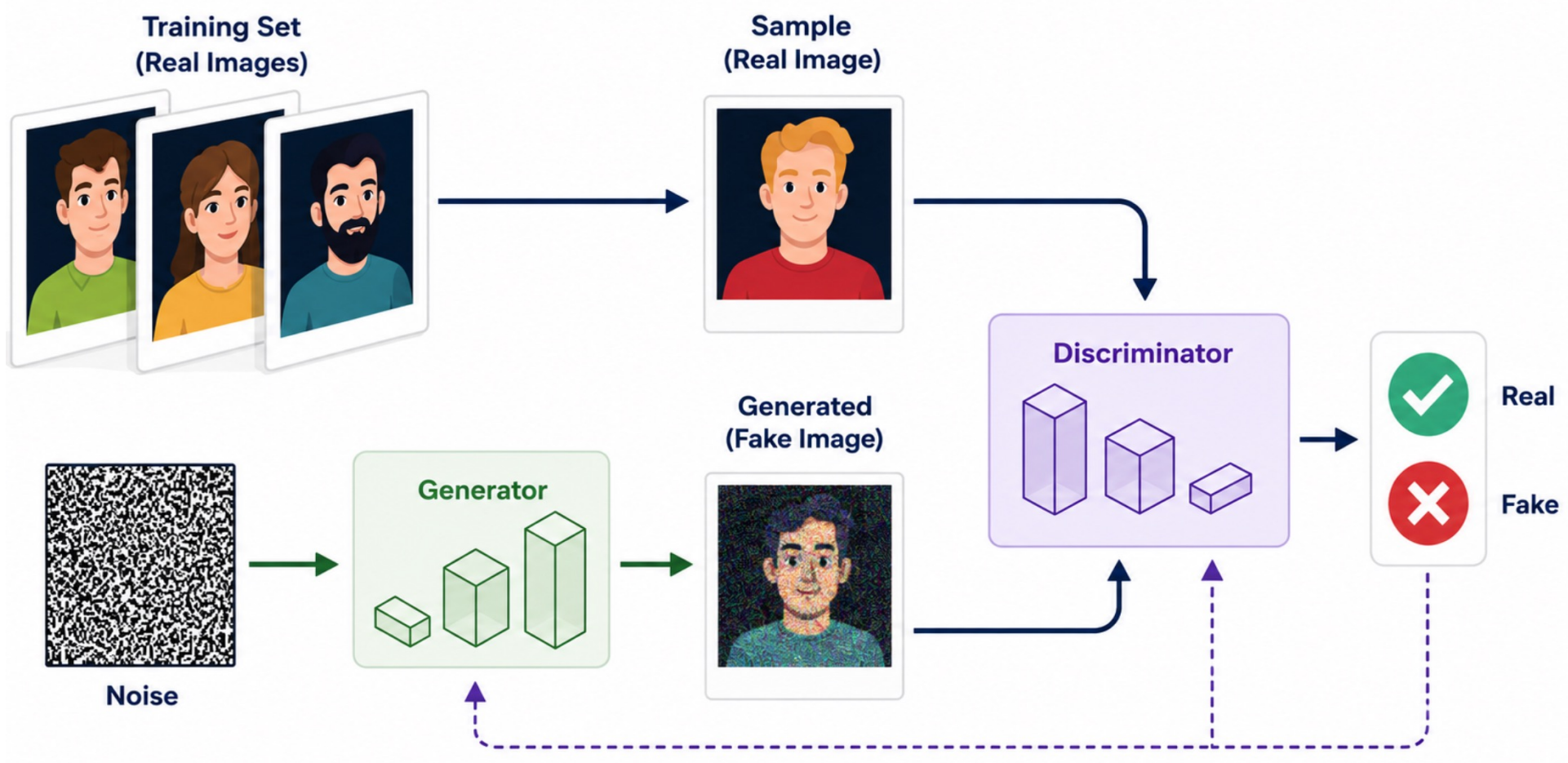
1. The Adversarial Idea

Implicit vs Explicit Generative Models



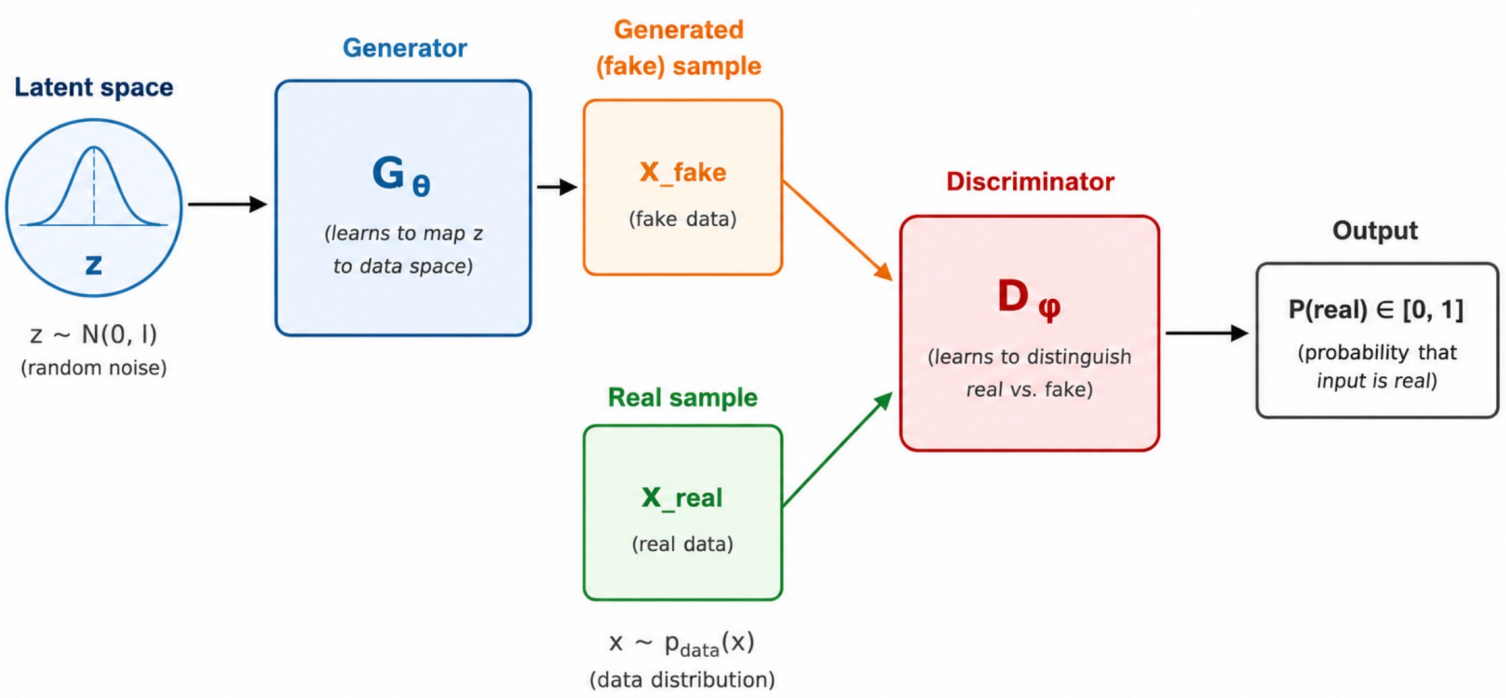
The Generator and Discriminator

- Generative Adversarial Networks (Goodfellow et al., 2014)



The Setup

- Real data: $x \sim p_{data}(x)$ (e.g., CelebA images)
- Latent noise: $z \sim p_z(z)$ (typically N(0, I) or U[−1, 1]^d)
- Generator: $G_\theta : z \to x$ — neural network. Defines implicit p_g.
- Discriminator: $D_\varphi : x \to [0,1]$ — outputs probability that x is real.



The Minimax Objective

- D maximizes — push real toward 1, fake toward 0.
- G minimizes — make D's second term fail.

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

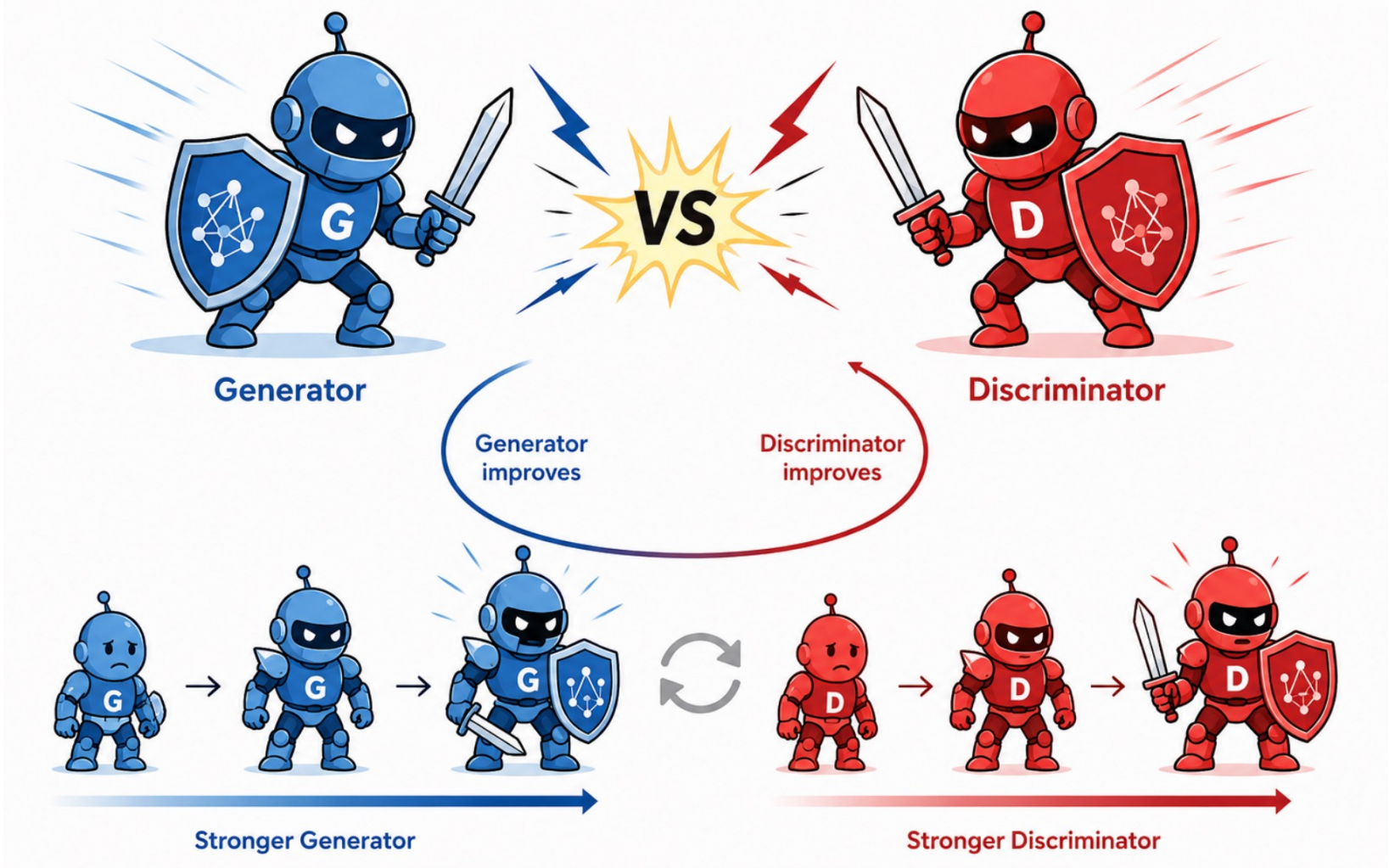
D rewards real images
Pushes $D(x) \to 1$

D penalizes fake images
Pushes $D(G(z)) \to 0$

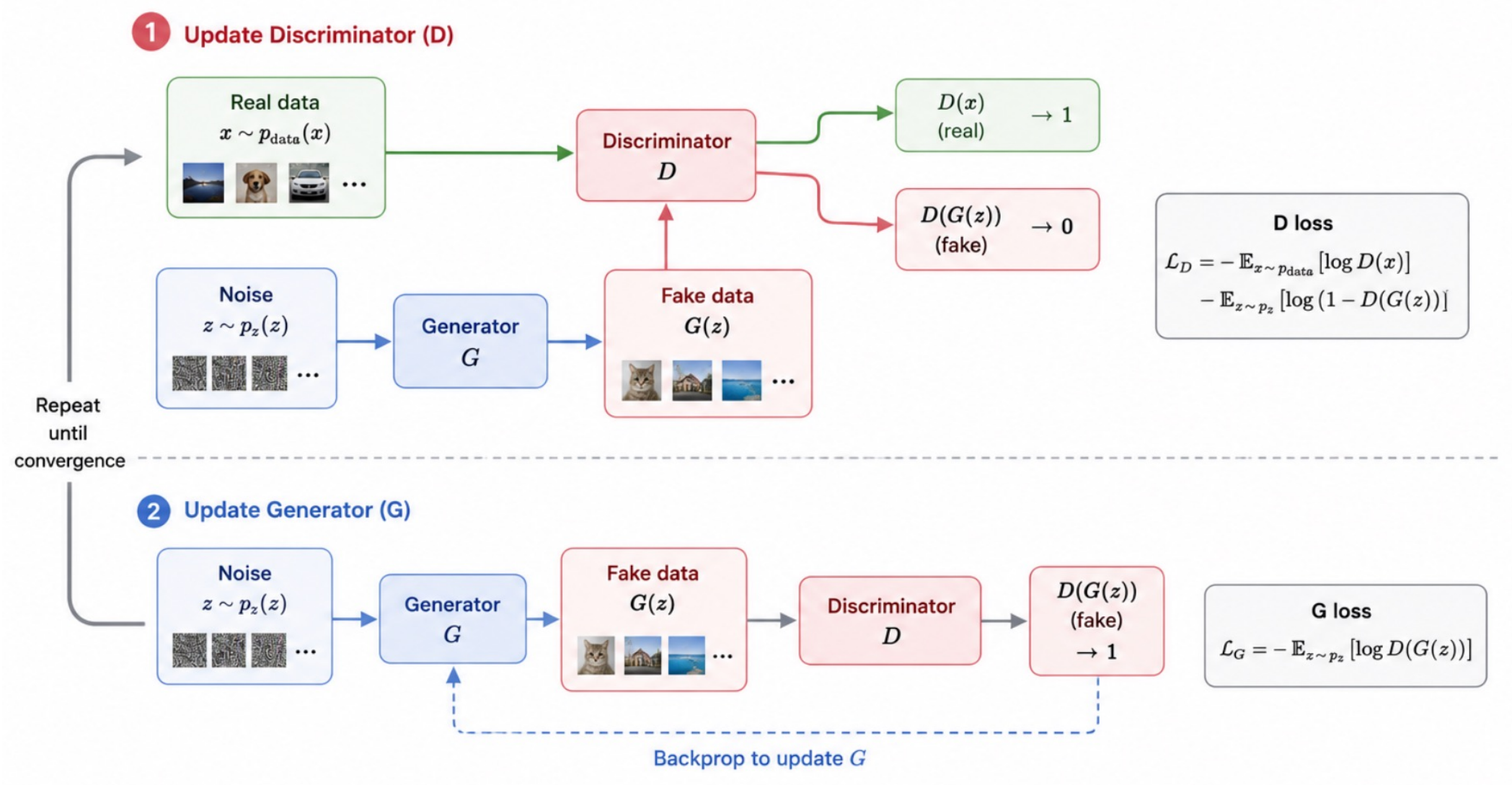
G tries to fool D. D tries to catch G.
Through this adversarial game, both get better.

The Minimax Objective

- Both generator and discriminator get better by learning from each other.



Algorithm: Vanilla GAN Training



Discussion

- *Why must D be retrained as G improves? What goes wrong if we freeze D after pretraining it on real data?*

- Answer: A frozen D defines a fixed loss landscape. G can find a single x* that maximally fools it.
- ***D's adaptiveness is what forces G to model the distribution.***

2. The Math

What does the minimax game actually optimize?

Step 1: Fix G, Solve for Optimal D

1. For fixed G, V(D, G) is a functional of D alone:

$$V(D,G) = \int_x p_{\text{data}}(x) \log D(x) + p_g(x) \log(1 - D(x))\ dx$$

2. Pointwise: maximize $a \cdot \log b + (1-a) \cdot \log(1-b)$ over $b \in [0,1]$ for each x.
3. Take derivative, set to zero:

$$D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}$$

The optimal discriminator is a density ratio.

Step 2: Substitute D* back

- Plugging D* into V gives the virtual training criterion for G:

$$C(G) = \mathbb{E}_{p_{\text{data}}}\left[\log\left(\frac{p_{\text{data}}}{p_{\text{data}} + p_g}\right)\right] + \mathbb{E}_{p_g}\left[\log\left(\frac{p_g}{p_{\text{data}} + p_g}\right)\right]$$

- Two expectations, both look like log-ratios.

Step 3: Recognize the JS Divergence

- Add and subtract $\log 2$ in each term:

$$C(G) = -\log 4 + \mathrm{KL}\left(p_{\text{data}} \,\middle\|\, \frac{p_{\text{data}} + p_g}{2}\right) + \mathrm{KL}\left(p_g \,\middle\|\, \frac{p_{\text{data}} + p_g}{2}\right)$$

The Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the Kullback–Leibler divergence $D(P \parallel Q)$. It is defined by

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M),$$

where $M = \frac{1}{2}(P + Q)$ is a mixture distribution of P and Q.

- The two KL terms are exactly $2 \cdot \mathrm{JSD}(p\_data \parallel p\_g)$.

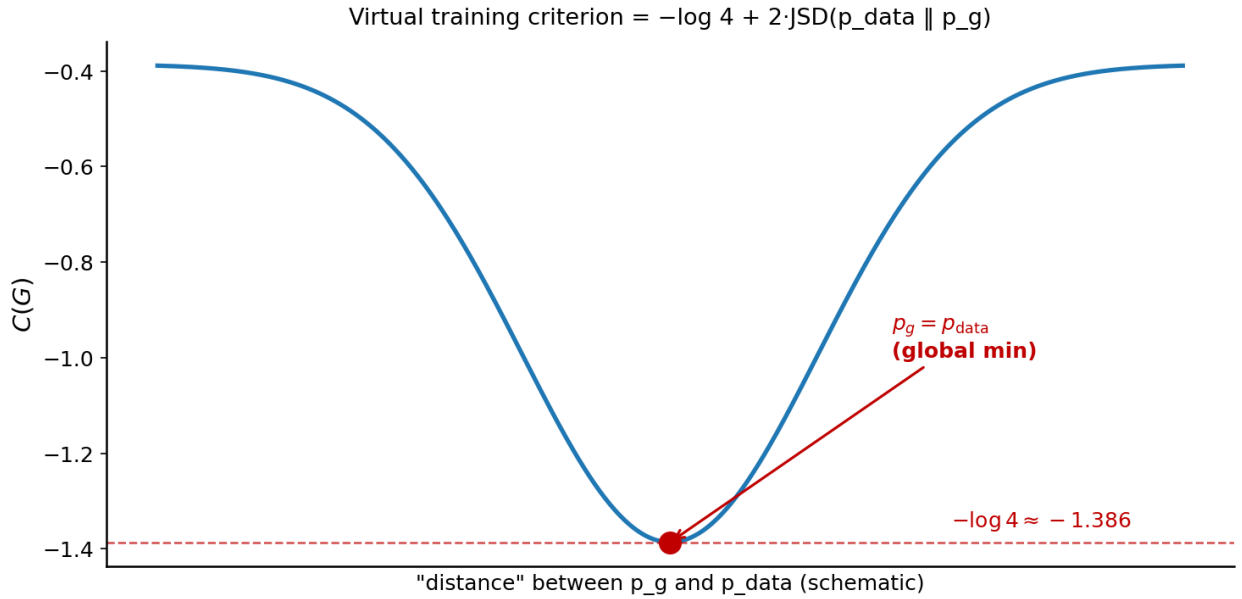
$$C(G) = -\log 4 + 2 \cdot \mathrm{JSD}(p_{\text{data}} \,\|\, p_g)$$

Global Optimum

- JSD ≥ 0, with equality iff (if and only if) p_g = p_data.

- Therefore:
  C(G) = −log 4
  achieved iff p_g = p_data.

- At that point:
  D*(x) = 1/2 everywhere
  — *the discriminator is reduced to coin-flipping*.



But: A Warning

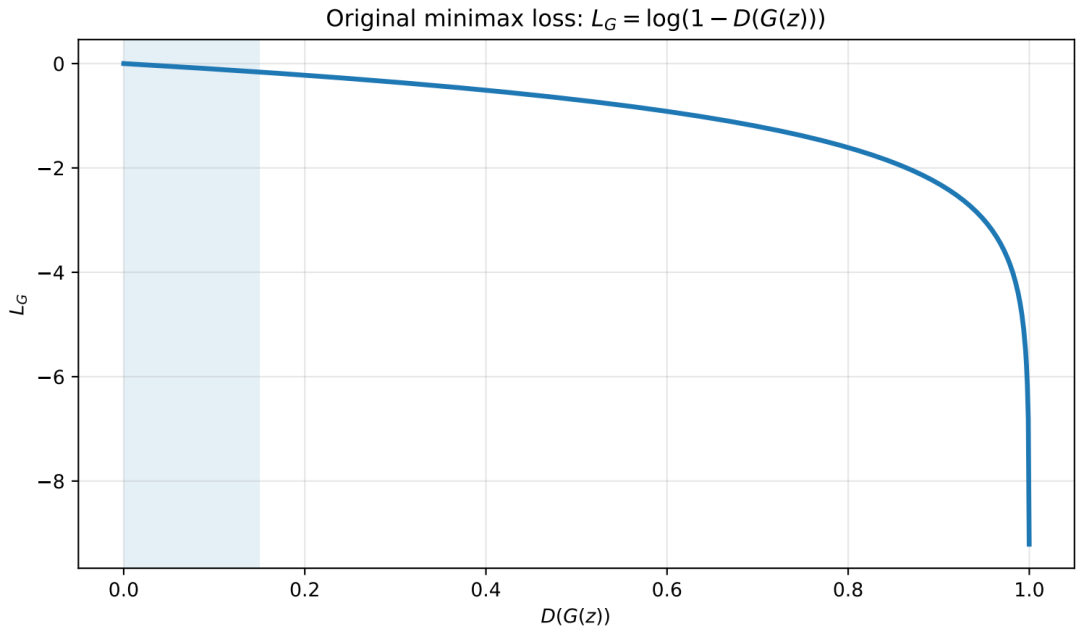
The proof we just walked through assumes:

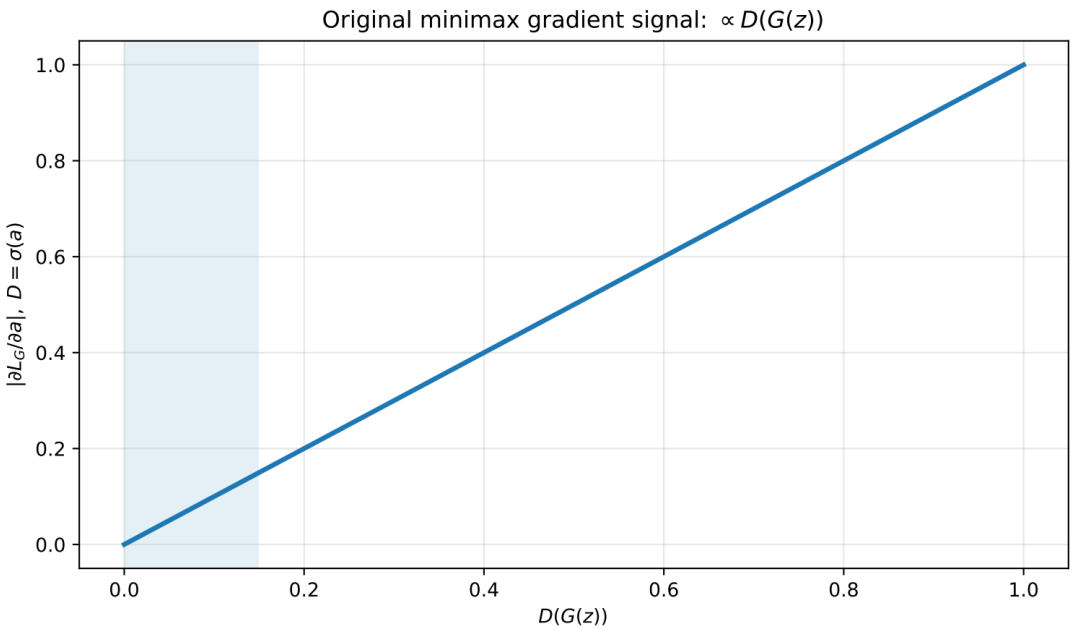
1. Infinite capacity in G and D — any function is representable.
2. D is trained to optimum at every step of G.
3. We optimize in distribution space p_g, not parameter space θ.

- *In practice: finite networks, alternating SGD, parameter space.*
- ***None of the assumptions hold.***

The Vanishing Gradient Problem

- Original G loss: $L_G = \mathbb{E}_z \left[\log\left(1 - D(G(z))\right)\right]$
- Early training: D easily flags fakes → $D(G(z)) \approx 0 \rightarrow \log(1 - D(G(z))) \approx 0$
  → $\nabla_\theta L_G \approx 0$. **G stops learning exactly when it most needs to.**

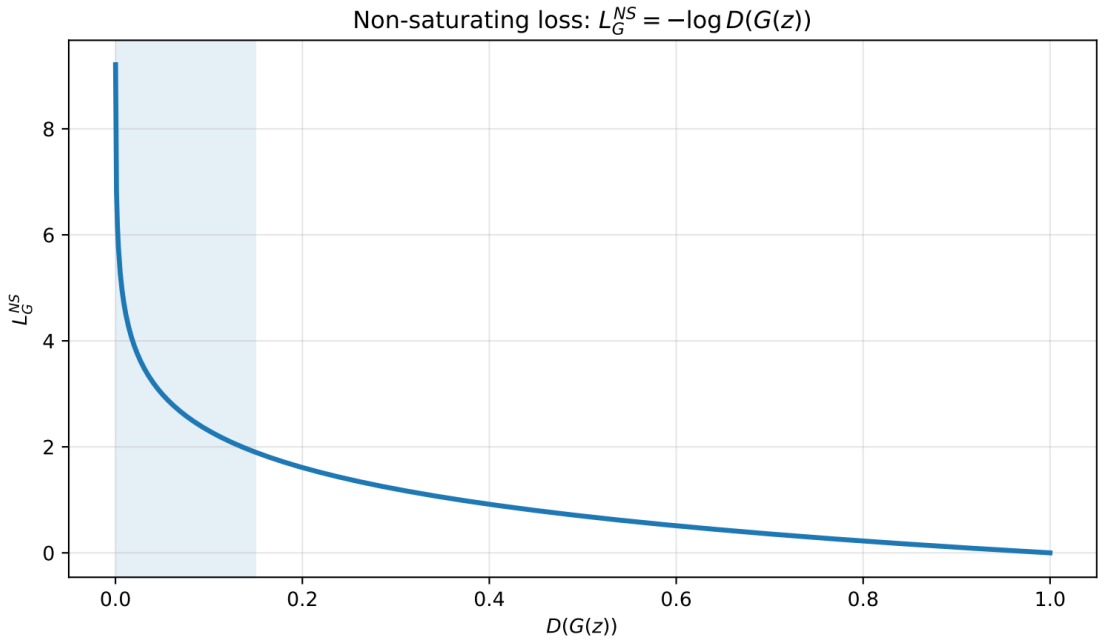


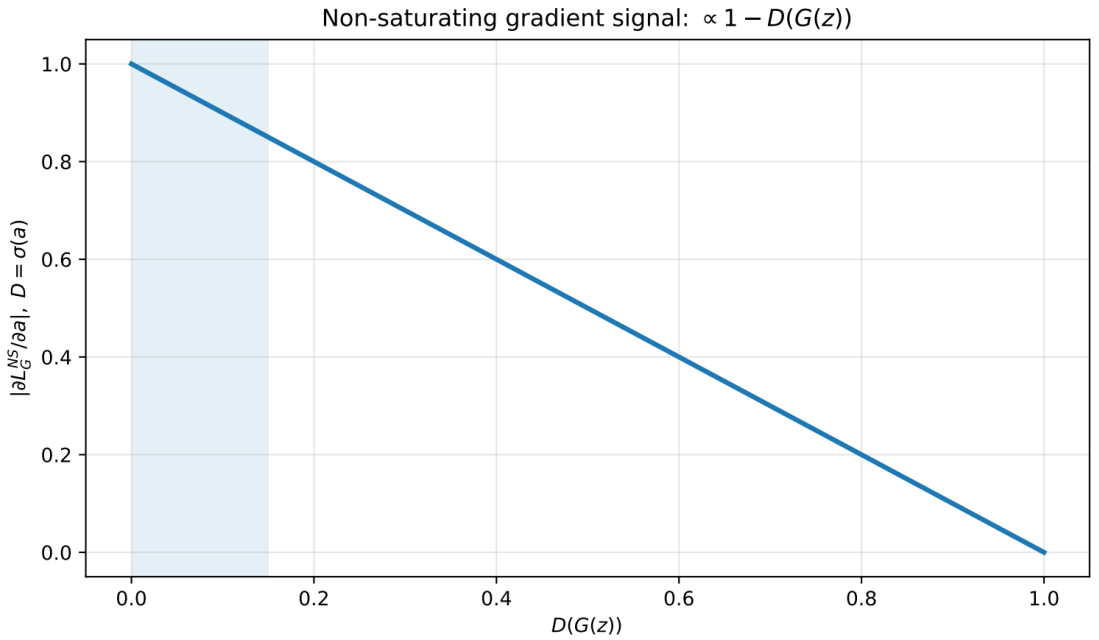


Fix: The Non-Saturating Loss

- Goodfellow's fix (in the same paper):

$$L_G^{\text{NS}} = -\mathbb{E}_z \left[\log D(G(z))\right]$$





Math Recap

1. Optimal D

- density ratio p_data / (p_data + p_g)

2. Optimal G

- minimizes JSD(p_data ‖ p_g) → unique global min at p_g = p_data

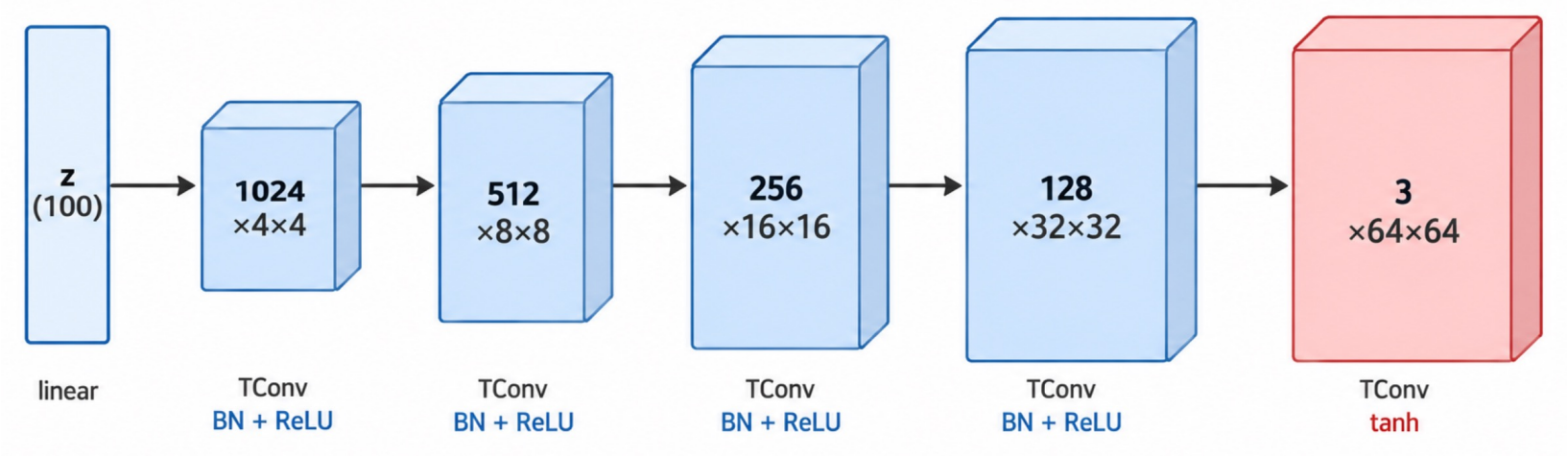
3. In practice

- non-saturating loss for gradients + alternating SGD in parameter space
- *no convergence guarantee.*

3. Making It Work

DCGAN (Radford et al., 2016)

- *The first architectural recipe that made GANs train reliably on images.*



Key changes from MLP-GAN:

- Strided conv (D) and transposed conv (G), no pooling
- BatchNorm in both networks (except G output and D input)
- ReLU in G, LeakyReLU in D. No fully-connected hidden layers.

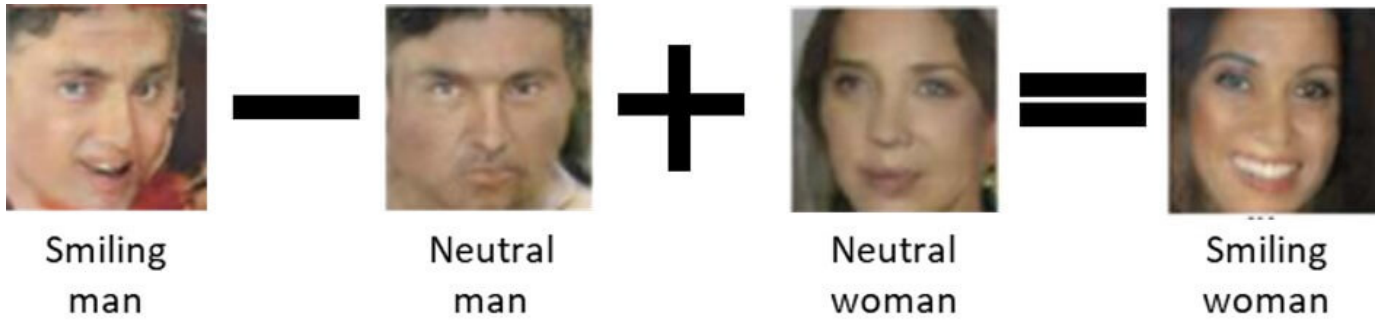
DCGAN: What It Showed

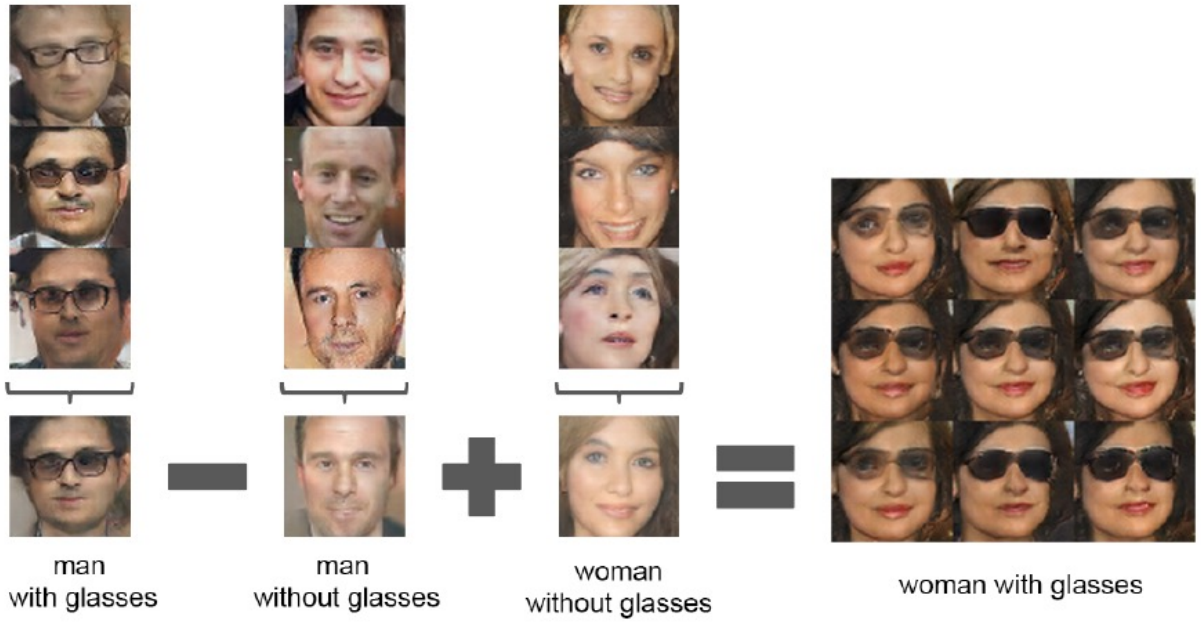
Latent space arithmetic:

- vec(man with glasses)
  - − vec(man)
  - + vec(woman)
  - ≈ **vec(woman with glasses)**
- Smooth latent interpolations between bedrooms, faces.

- *Proof that the implicit p_g learned semantic structure*

— the same property VAEs offered, now without blur.





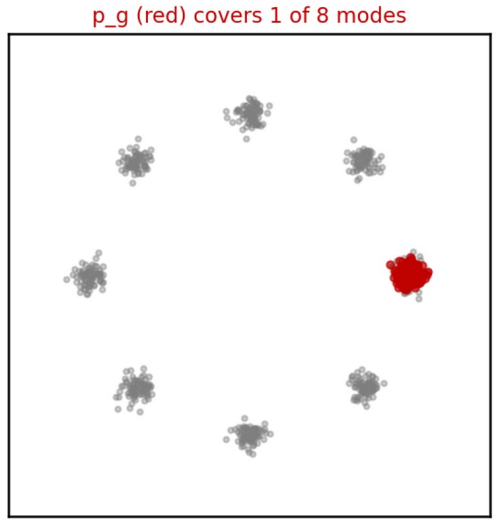
Mode Collapse: The Failure Mode

Definition.

- G learns to produce a small subset of p_data's modes, ignoring the rest.
- ***Extreme case: G outputs the same image regardless of z.***

Why it happens.

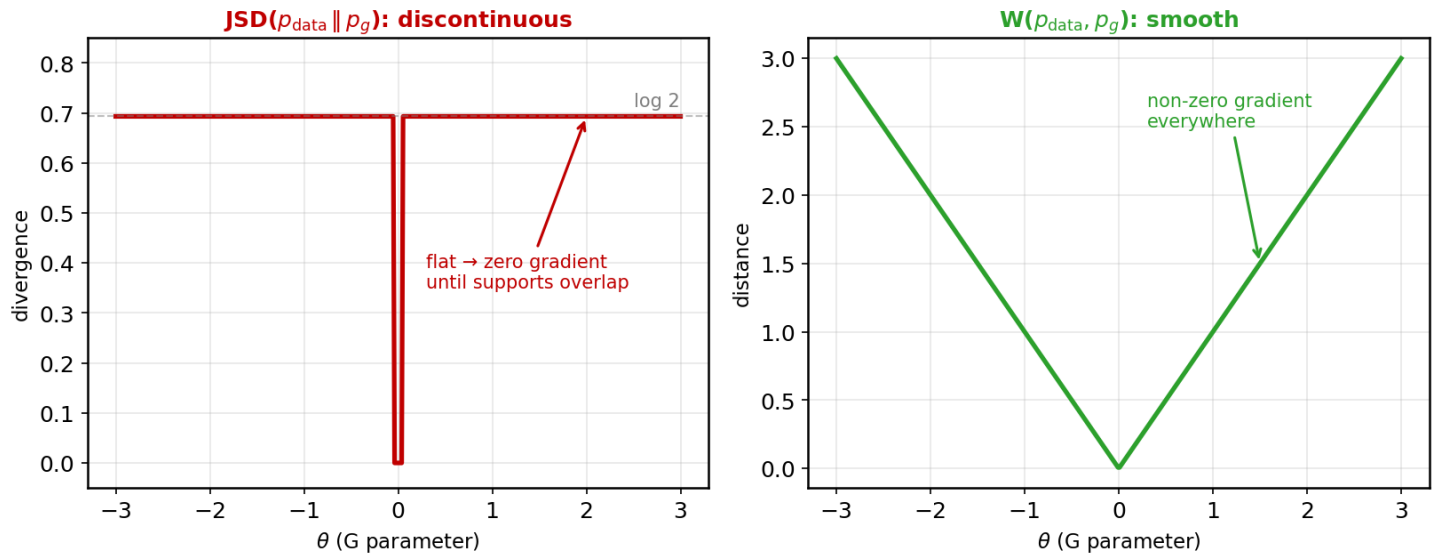
- Non-saturating G loss is mode-seeking.
- Alternating optimization can find a narrow x* that fools the current D,
- then D adapts only locally,
- *then G shifts to another narrow x*.*



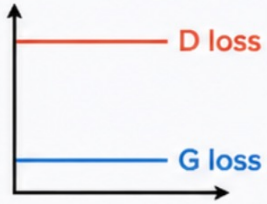
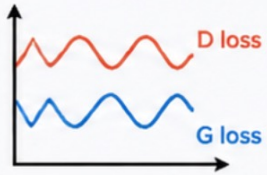
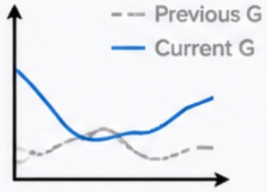
Wasserstein GAN (Arjovsky et al., 2017)

- **Diagnosis:** JSD is discontinuous when p_data and p_g have disjoint supports — gradient is zero or undefined.
  => Explanation: If the real distribution and generated distribution do not overlap, the JSD provides almost no useful gradient.
- **Fix. Replace JSD with Wasserstein-1 (Earth Mover's) distance:**

$$W(p_{\text{data}}, p_g) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{p_{\text{data}}}[f(x)] - \mathbb{E}_{p_g}[f(x)]$$

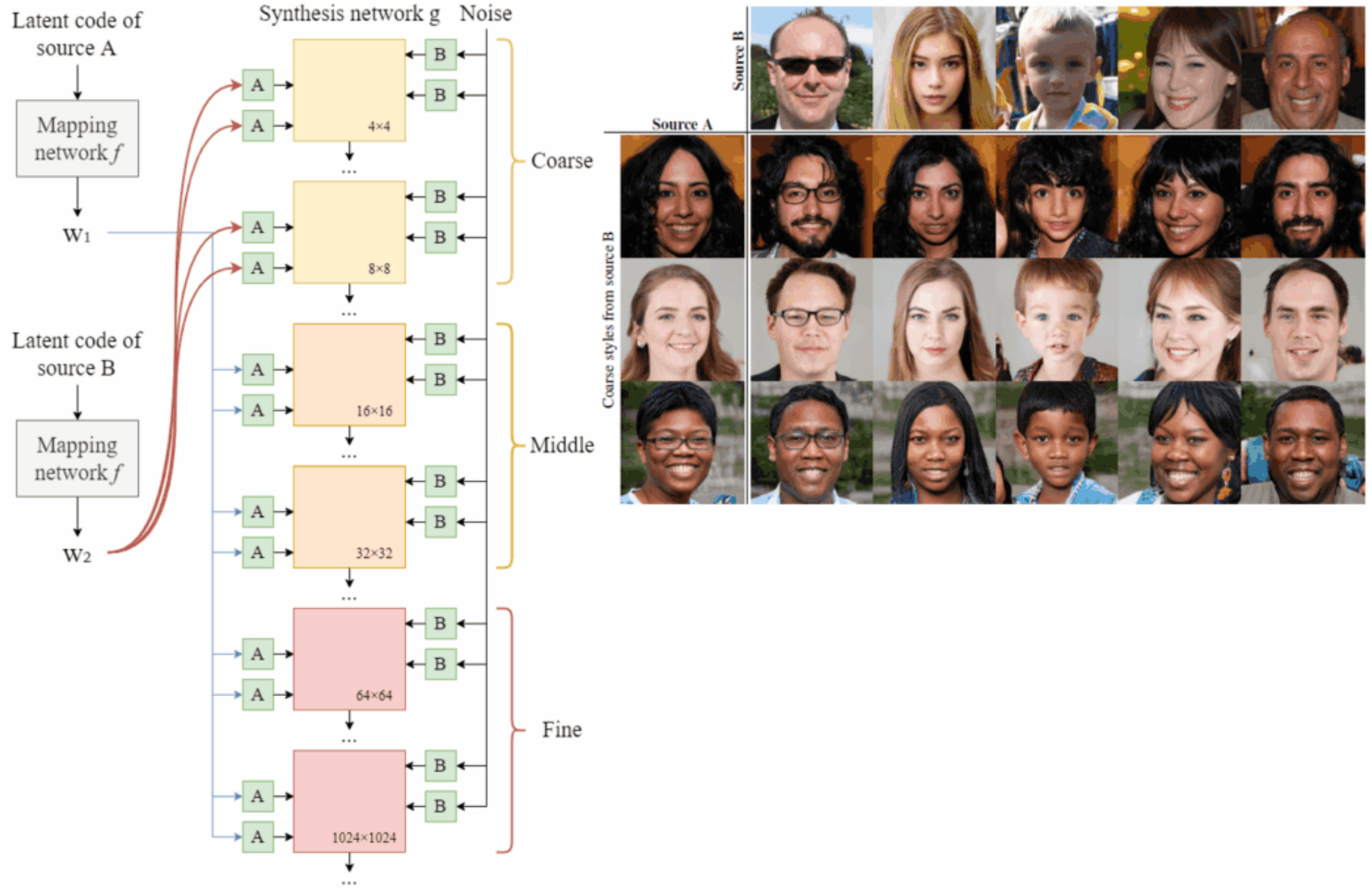


Diagnosing failures

Symptom	Likely cause	Fix
 **G loss → 0, D loss → 0, samples are noise**	• D too strong • Vanishing gradient	• Use non-saturating loss • Lower D learning rate
 **Samples sharp but identical**	• Mode collapse	• Minibatch discrimination • Unrolled GAN; WGAN
 **Both losses oscillate forever**	• No convergence in parameter space	• TTUR (two time-scale update rule) • Spectral normalization • Gradient penalty
 **Improving then catastrophic worsening**	• D over-fits a checkpoint of G	• Historical averaging • Experience replay

Progressive growing → StyleGAN

- ProgGAN (Karras et al. 2017): start at 4×4, grow resolution layer-by-layer to 1024×1024.
- StyleGAN (Karras et al. 2019): map z → style w, inject at every layer via AdaIN.
  Disentangles coarse (pose) from fine (texture).
- StyleGAN 2 / 3: fix artifact issues, equivariance under translation/rotation.



Style-mixing in StyleGAN

Summary

- The idea: implicit generator + adaptive critic. Replaces likelihood with a learned distance.
- The theory: at idealized equilibrium p_g = p_data, JSD minimized — but it is a saddle point.
- The practice: NS loss + DCGAN arch + Wasserstein + spectral norm + progressive growing.