Generation Adversarial Network

为什么要GAN？

什么是生成模型？

- 生成模型
  - 给定一组真实数据样本 $\{x_1, x_2, \ldots, x_n\}$，假设从未知的真实数据分布 $P_{data}(x)$ 中采样得到
  - 生成模型 (Generative Model)，任务是学习一个模型 $P_{model}(x)$，使其尽可能地逼近 $P_{data}(x)$
- 为什么要生成？
  - 理解数据。 一个好的生成模型抓住了数据的本质结构与变化规律
  - 创造数据。一旦学到 P_{model}，可从中采样，生成全新的、与真实数据风格一致的数据
  - 应用: 图像合成、风格迁移、数据增强、超分辨率、药物发现等
- 传统方法所面临的挑战
  - 传统的对高维数据 $P_{data}(x)$ 进行最大似然估计（MLE），通常是极其困难或不可行

为什么要GAN？

既然无法写下一个好的损失函数来衡量“真实性”，那么，我们能否让一个模型自己去学习这个损失函数？

- 传统范式(显式/近似优化)
  - 1. 定义一个固定的损失函数(如：对数似然，MSE)
  - 2. 优化生成器以最小化该损失结果：数学上可解，但感官上模糊
- GAN范式(对抗博弈)
  - 1. 学习一个动态的损失函数 (判别器D)D的目标：区分真伪
  - 2. 优化生成器G以“欺骗”这个学习中的D结果：将生成问题转化为博弈问题

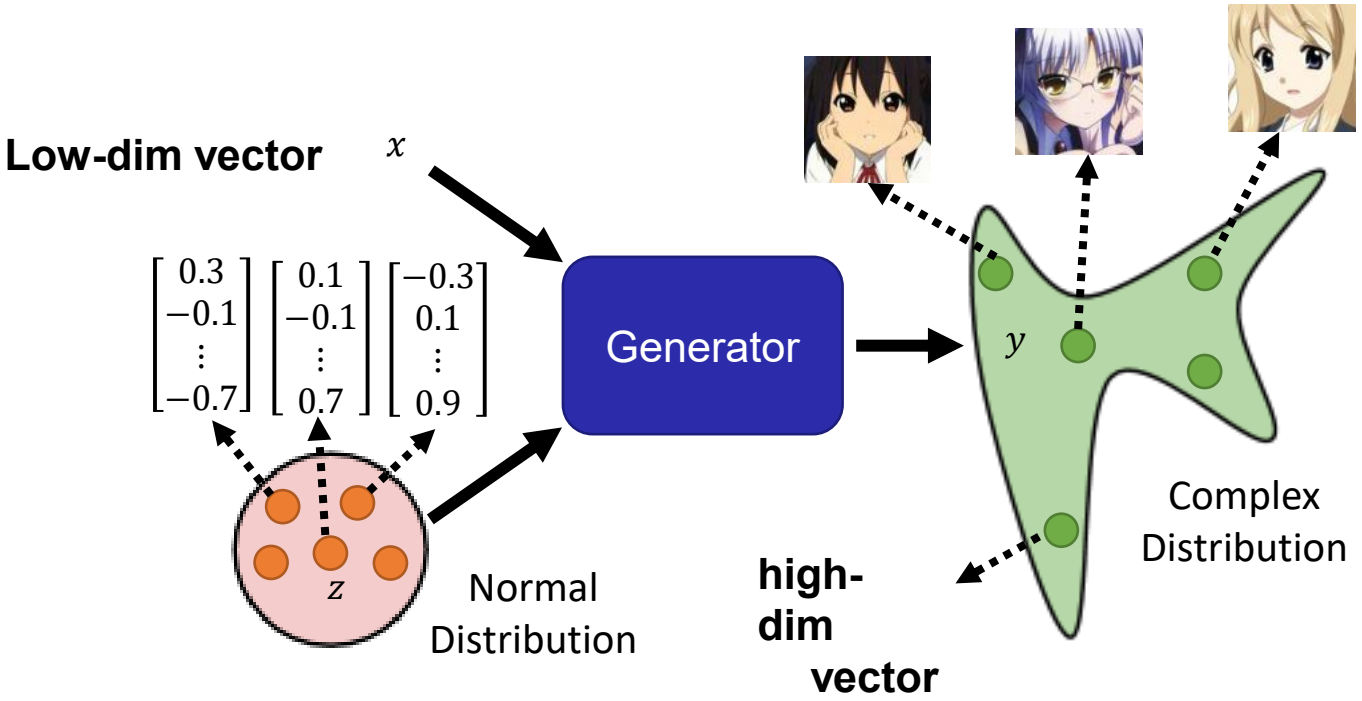
GAN概要

GAN的核心思想：直觉与类比

- GAN的根本思想：通过对抗，实现学习
- 类比：伪画制造者 (Generator) vs. 艺术品鉴赏家 (Discriminator)
  - 生成器 (Generator, G)
    - “伪造者”，目标画出能够以假乱真的“赝品”。从未见过真正的名画，只能根据鉴赏家反馈提升画技
  - 判别器 (Discriminator, D):
    - 一个经验丰富的“鉴赏家”，目标是准确区分出“真品”（来自博物馆）和“赝品”（来自生成器）
    - 它通过阅览大量的真品和赝品来提升自己的鉴别能力
- 动态博弈过程
  - 初期。G的作品很拙劣，D能轻易分辨。D的反馈（“太假了”）指导G改进
  - 中期。G的技艺提升，D必须更仔细地寻找破绽才能分辨。D的鉴别能力也在提升
  - 最终。G的画作足以乱真，D再也无法有效分辨（只能考猜）
    - 此时，我们认为G已经学到了创造“名画”的精髓
- G和D在目标上相互对抗，但在整个系统的进化上，它们共同协作，缺一不可

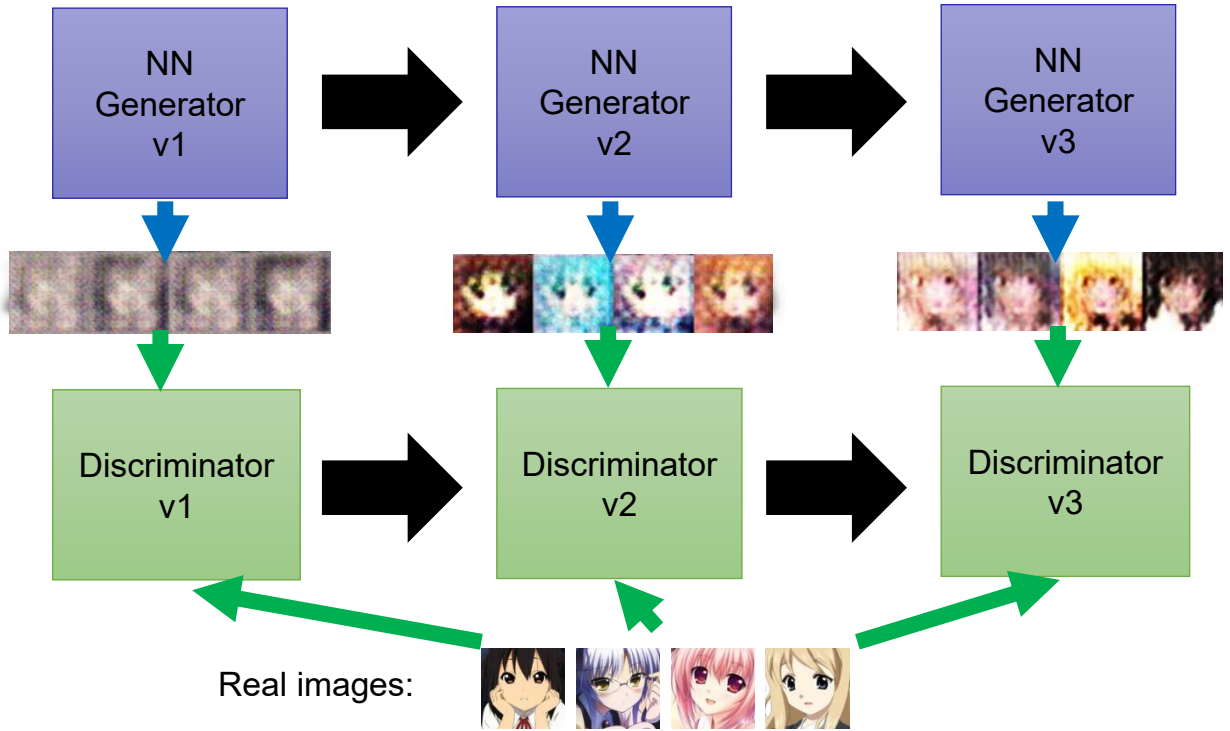
什么是生成模型：Anime Face Generation

- 呈现形式



Basic Idea of GAN

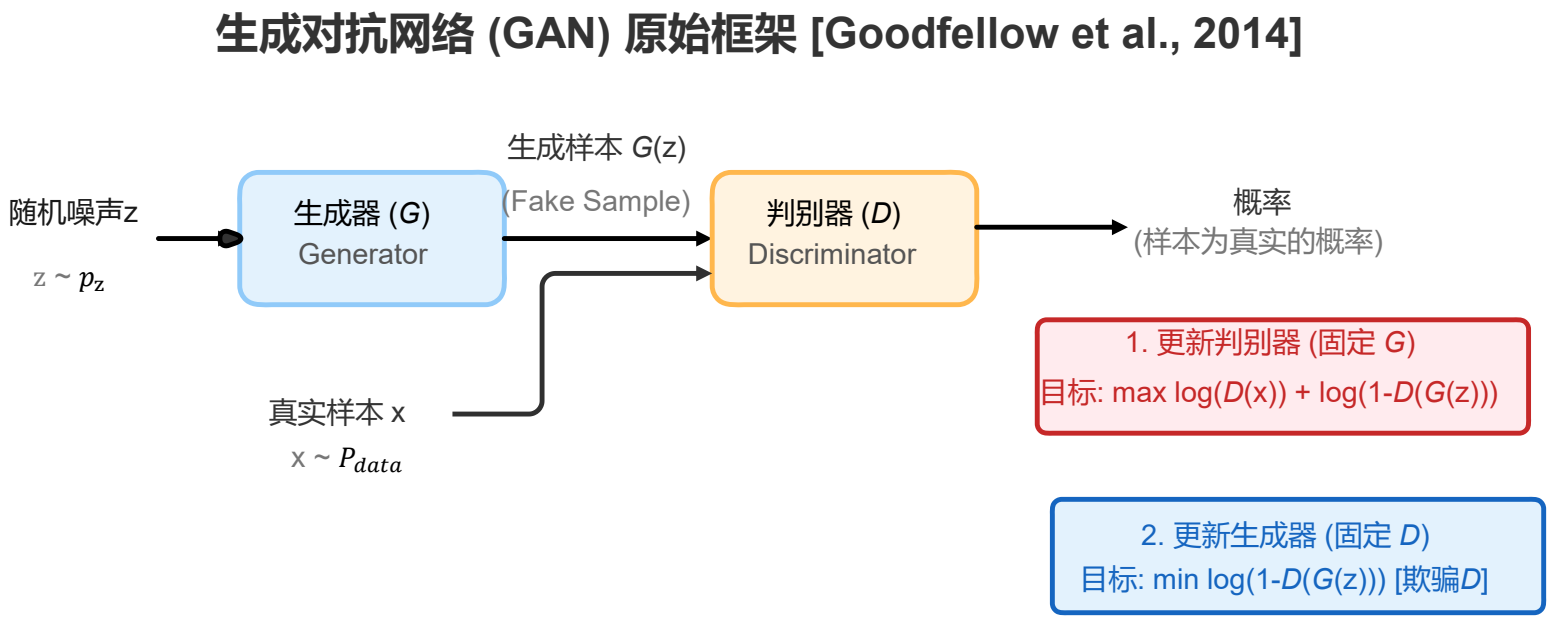
- This is where the term “adversarial” comes from



GAN模型架构

GAN的核心机制：模型架构

➢框架



GAN的核心机制：模型架构

- 生成器 (Generator)
  - 输入: 从简单先验分布（如高斯分布）中采样随机噪声向量 z，作为生成内容的“语义种子”
  - 网络: 通常是一个深度神经网络（如反卷积网络）
  - 输出: 生成的数据 $G(z)$，其维度和结构与真实数据 x 相同（例如，一张图片）
  - 目标: 生成的数据 $G(z)$ 的分布 P_G 要尽可能接近真实数据分布 P_{data}
- 判别器 (Discriminator)
  - 输入: 真实数据 x，或者生成数据 $G(z)$
  - 网络: 通常为标准的分类神经网络（如CNN）
  - 输出: 一个标量值 $D(x)$，表示输入数据为“真实”的概率（或分数）
  - 目标: 对真实数据 x 输出高分（接近1），对生成数据 $G(z)$ 输出低分c（接近0）

对抗博弈的数学表达：价值函数

- GAN的训练过程是一个博弈过程，其目标函数为

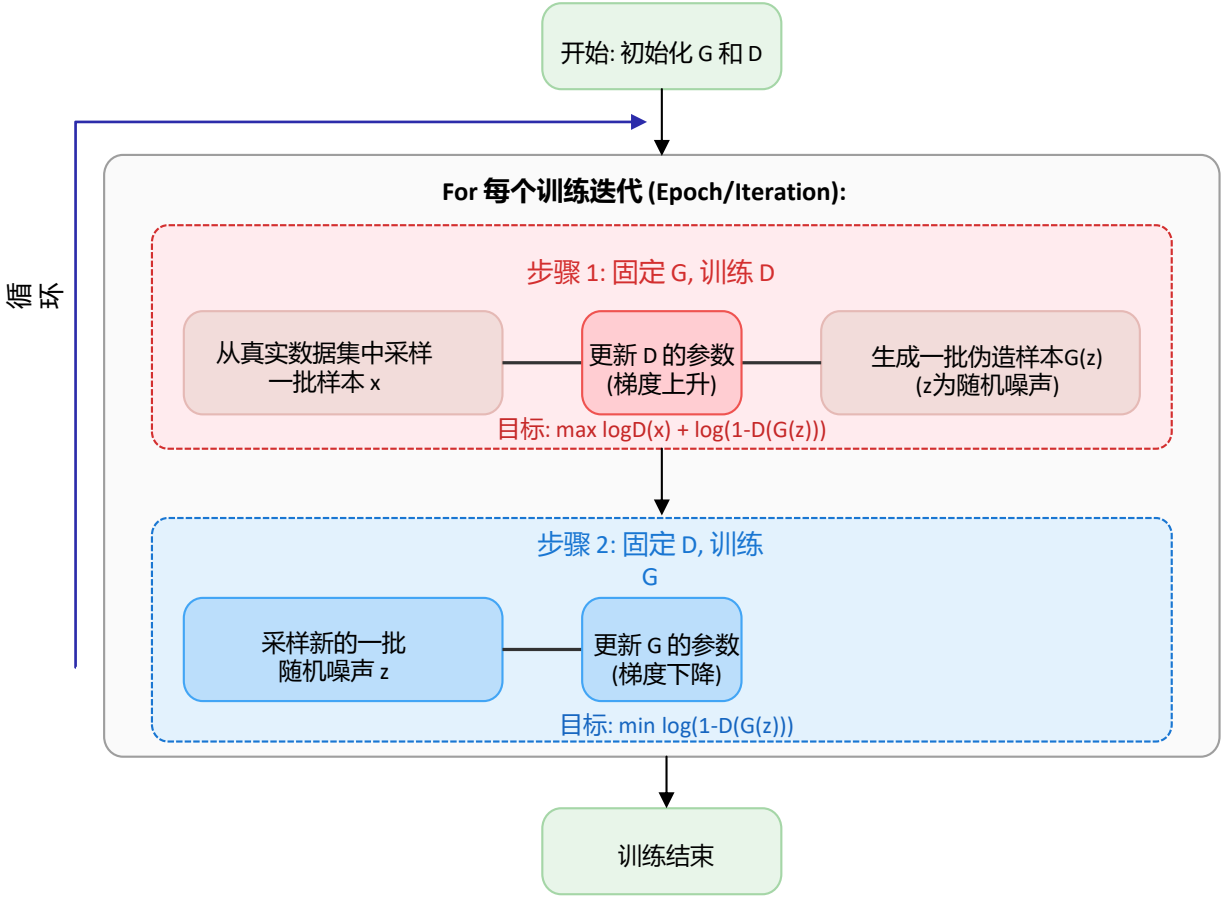
$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))]$$

  - 最大化判别器D(max D)：学习区分真实数据和生成数据
    - $\mathbb{E}_{x\sim P_{data}(x)}[\log D(x)]$: 真实数据 x，最大化 $D(x)$，即让 $\log D(x)$最大
    - $\mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))]$: 生成数据 $G(z)$ ，最小化 $D(G(z))$
      - $\log(1-D(G(z)))$ 最大
    - 等价于训练一个二元分类器D，正确区分真实样本和生成样本
      - 最小化标准的二元交叉熵损失
  - 最小化生成器 G (min G)：学习欺骗判别器
    - G 无法影响第一项 $\mathbb{E}_{x\sim P_{data}(x)}[\log D(x)]$
    - 最大化 $D(G(z))$ ，等价于最小化 $\mathbb{E}_{z\sim P_z(z)}[\log(1-D(G(z)))]$

模型训练过程

训练流程图

- 交替


开始: 初始化 G 和 D
For 每个训练迭代 (Epoch/Iteration):
循环
步骤 1: 固定 G, 训练 D
从真实数据集中采样
一批样本 x
更新 D 的参数
(梯度上升)
生成一批伪造样本G(z)
(z为随机噪声)
目标: max logD(x) + log(1-D(G(z)))
步骤 2: 固定 D, 训练
G
采样新的一批
随机噪声 z
更新 G 的参数
(梯度下降)
目标: min log(1-D(G(z)))
训练结束

训练算法：交替优化

- 直接求解minmax问题很困难，采用交替迭代优化的方式训练:
  - Step 1: 固定 G，优化 D
    - 从真实数据集中采样一批样本 $\{x_1, \ldots, x_m\}$
    - 从噪声分布 P_z 中采样一批向量 $\{z_1, \ldots, z_m\}$，通过生成器得到一批伪造样本 $\{G(z_1), \ldots, G(z_m)\}$
    - 使用这些真实样本和伪造样本，通过梯度上升来更新判别器 D 的参数，以最大化目标 $V(D, G)$

$$\nabla_{\theta_d} \frac{1}{m} \sum_{\mathrm{i}=1}^{\mathrm{m}} [\log D(x_i) + \log(1 - D(G(z_i)))]$$

  - Step 2: 固定 D，优化 G
    - 从噪声分布 P_z 中采样新的一批向量 $\{z_1, \ldots, z_m\}$
    - 通过梯度下降来更新生成器 G 的参数，以最小化目标 $V(D, G)$

$$\nabla_{\theta_g} \frac{1}{m} \sum_{\mathrm{i}=1}^{\mathrm{m}} \log(1 - D(G(z_i)))$$

    - 训练初期，$\log(1 - D(G(z)))$ 梯度较小（饱和区）。实践中常使用“非饱和”目标函数，即最大化 $\log(D(G(z)))$，这提供了更强的梯度信号

GAN的理论分析

理论的优雅：GAN的终极目标

- GAN的收敛点与JS散度
- 原始论文证明了一个关键理论
  - 在理想情况下 (即判别器D有无限建模能力)，对于固定的G，最优判别器 $D^*(x)$ 为

$$D^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_G(x)}$$

    - 代表，判定来自真实数据的概率
  - 将 $D^*(x)$代入整个博弈的价值函数，证明了GAN的理论目标等价于最小化JS散度
    - 该博弈的理论价值为：$C(G) = \max_D V(D,G) = 2 \cdot JSD(P_{data}||P_G) - 2\log 2$
  - 注意：实践中G优化的是独立的损失函数，而非直接优化$C(G)$。对抗训练是间接逼近该理论目标的过程
- 对抗训练与纳什均衡
  - 当 $P_G = P_{data}$ 时，JS散度为0，$D^*(x) = 1/2$，判别器无法区分真假，博弈达到纳什均衡

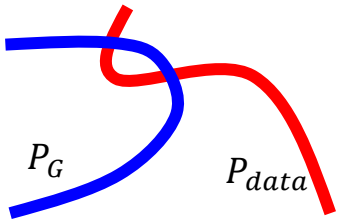
理论与实践的鸿沟：GAN为何难以训练

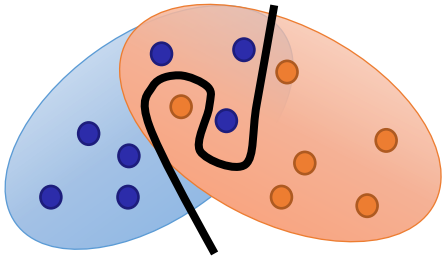
实践困境：为什么GAN难训练？

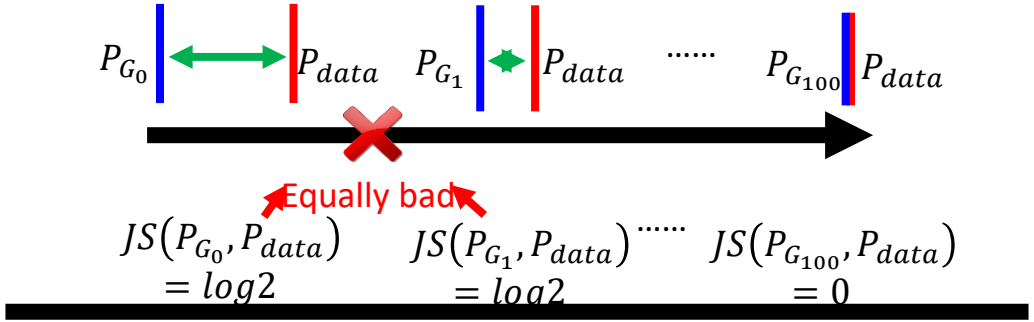
- 理论根源：流形假设 (Manifold Hypothesis)
  - 高维数据（如图像）分布在空间中的低维流形上。在训练初期，两个流形几乎没有重叠
- 导致的连锁反应
  - 判别器过于强大：由于两个分布不重叠，判别器可以轻易地找到一个超平面将它们完美分开
    - 对于真实样本，$D(x) \to 1$；对于生成样本，$D(G(z)) \to 0$
  - JS散度失效
    - 当分布不重叠时，$JSD(P_{data}||P_G) = \log 2$，无法衡量“远近”，其梯度为0
  - 生成器梯度消失
    - 生成器的损失函数 $\log(1 - D(G(z)))$ 的梯度正比于 D 的梯度
    - 当 $D(G(z))$ 接近0时，判别器损失函数的梯度也趋近于0
    - 生成器 G 接收不到有效的梯度信号，无法学习和改进。训练停滞

JS divergence is not suitable

- In most cases, P_G and P_{data} are not overlapped.
- 1. The nature of data
  - Both P_{data} and P_G are low-dim manifold in high-dim space.
- 2. Sampling
  - Even though P_{data} and P_G have overlap.







Intuition: If two distributions do not overlap, binary classifier achieves 100% accuracy.
The accuracy (or loss) means nothing during GAN training.

WGAN

Wasserstein GAN (WGAN)

- 更好的Wasserstein-1 距离 (Earth Mover's Distance, 推土机距离)
  - 将一个概率分布（一堆沙土）变换成另一个（一个沙坑）所需移动沙土的“最小平均代价”
  - 即使两个分布完全不重叠，依然能提供一个有意义的、平滑的距离度量
- WGAN的目标函数(利用Kantorovich-Rubinstein对偶性)

$$W(P_{data}, P_G) = \sup_{\|f\|_L \le 1} \left\{ \mathbb{E}_{x \sim P_{data}}[f(x)] - \mathbb{E}_{x \sim P_G}[f(x)] \right\}$$

  - f 为判别器（在WGAN中称为Critic），它不再输出概率，而是输出一个实数分数
  - $\| f \|_L \le 1$ 是关键的 1-Lipschitz约束，要求Critic函数必须足够“平滑”
- WGAN的minimax博弈

$$\min_G \max_{D \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_{data}}[D(x)] - \mathbb{E}_{z \sim P_z}\left[D\left(G(z)\right)\right] \right\}$$

  - 其中 $\mathcal{D}$ 是所有1-Lipschitz函数的集合

WGAN为何能解决梯度消失？

- 原始GAN判别器
  - 输出的是概率，经过Sigmoid激活
  - 像“悬崖”，底部梯度饱和
- WGAN Critic
  - 输出的是分数，受Lipschitz约束
  - 像“平滑的斜坡”，处处有梯度

- Lipschitz约束保证了无论G多差，总能获得有效的梯度来学习

WGAN的实践：如何实现Lipschitz约束？

- 强制Critic满足1-Lipschitz约束是WGAN成功的关键
- Weight Clipping (原始WGAN)
  - 每次更新Critic参数后，将其裁剪到一个小范围 [-c, c] 内
  - 简单粗暴。会导致Critic倾向于学习非常简单的函数，或者在裁剪边界处梯度消失/爆炸，降低了模型容量和训练稳定性
- Gradient Penalty (WGAN-GP)
  - 在Critic的损失函数中加入一个惩罚项，惩罚其梯度范数偏离1的行为
  - $\lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2]$，其中 $\hat{x}$ 是在真实样本和生成样本之间的随机插值点
  - 效果显著优于Weight Clipping，成为后续很多GAN模型的基础
- Spectral Normalization (SN-GAN)
  - 对Critic网络中的每一层的权重矩阵进行谱范数归一化，使其Lipschitz常数被约束为1
  - 计算开销小，实现简单，且能提供比Gradient Penalty更稳定的训练过程。目前成为主流方法

Conditional Generation

GAN生态：架构的进化

- DCGAN (2015)
  - 奠定了GAN用于图像生成的架构指南。使用卷积层替代全连接层，用BatchNorm稳定训练，消除了池化层。
- Progressive GAN (2017)
  - 渐进式增长。从低分辨率（4x4）开始训练，逐步增加网络层数以生成更高分辨率（1024x1024）的图像。极大地提升了生成质量和训练稳定性。
- StyleGAN (2018-2020)
  - 风格控制与解耦。引入映射网络（Mapping Network）将输入噪声 z 映射到中间隐空间 w，并通过仿射变换（AdaIN）在不同尺度上控制生成图像的“风格”。实现了前所未有的生成质量和可控性。

GAN生态：可控生成 (cGAN)

- 如何控制GAN生成的内容？（例如，生成指定数字的MNIST图像）
- 解决方案: 将条件信息 y (如类别标签、文本描述) 同时提供给生成器和判别器。
  - 生成器 G: 输入变为 (z, y)，目标是生成符合条件 y 的图像 $G(z, y)$。
  - 判别器 D: 输入变为 (x, y)，目标是判断图像 x 是否真实 并且 是否与条件 y 相匹配。
- cGAN目标函数:

$$\min_G \max_D V(D,G) = \mathbb{E}_{(x,y)\sim P_{data}}[\log D(x,y)] + \mathbb{E}_{z\sim P_z, y\sim P_y}[\log(1 - D(G(z,y)))]$$

- 应用: Text-to-Image, Image-to-Image Translation (pix2pix) 等。

Text-to-image

red eyes
black hair

yellow hair
dark circles

x

red eyes

Generator

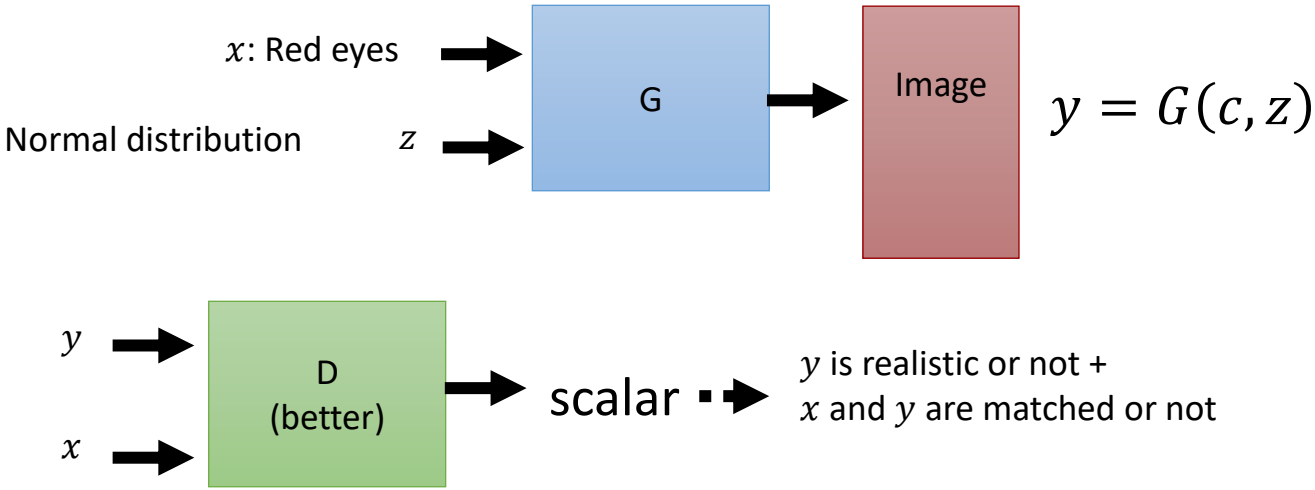
y

z

red hair,
green eyes

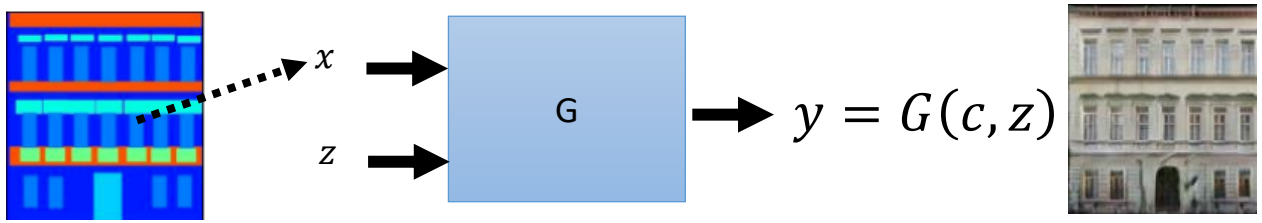
blue hair,
red eyes

Conditional GAN



Conditional GAN

https://arxiv.org/abs/1611.07004



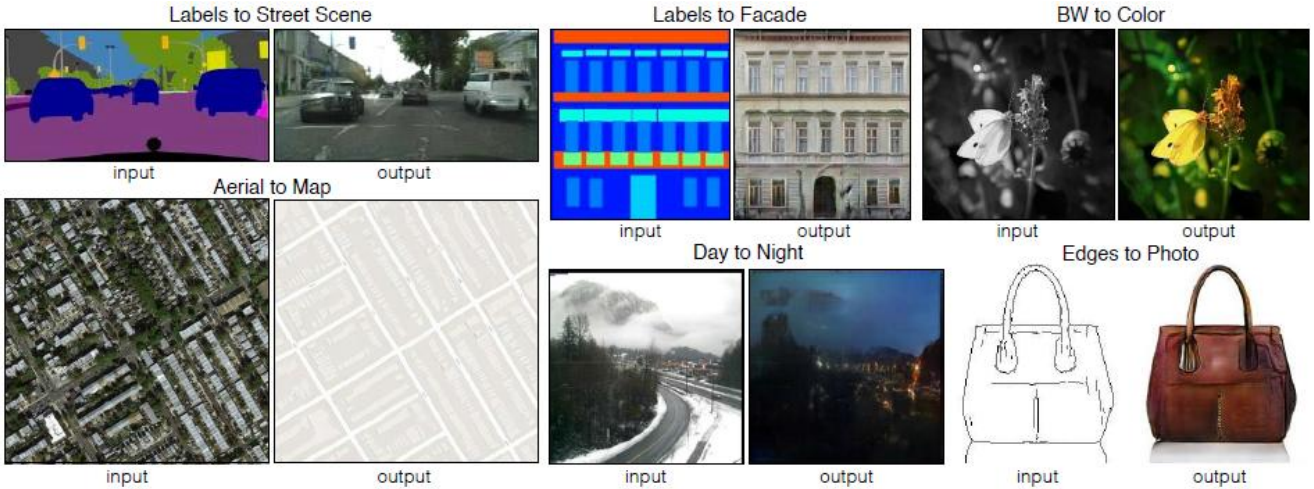
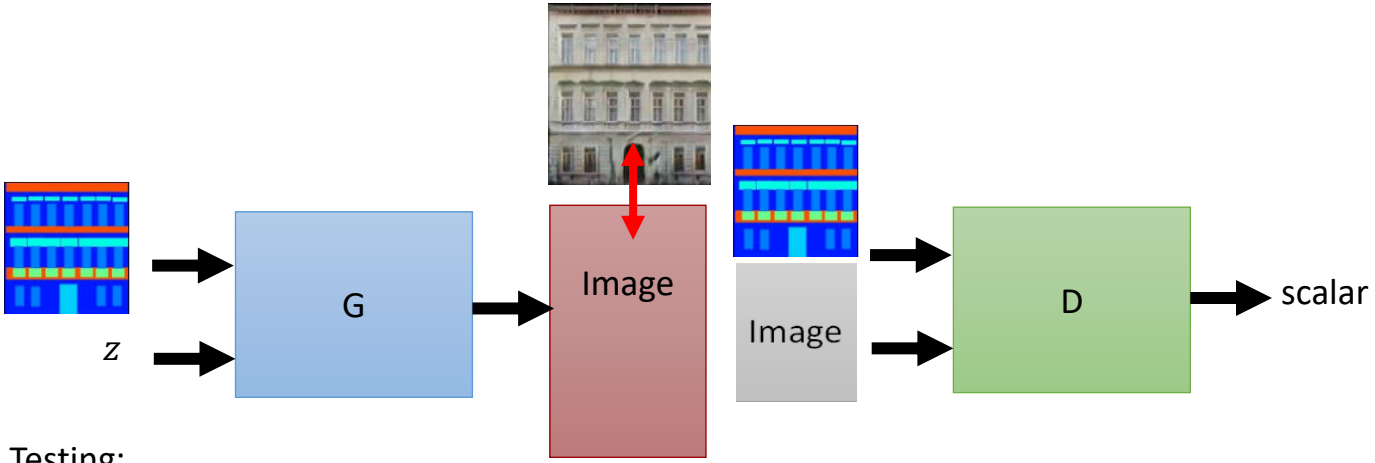
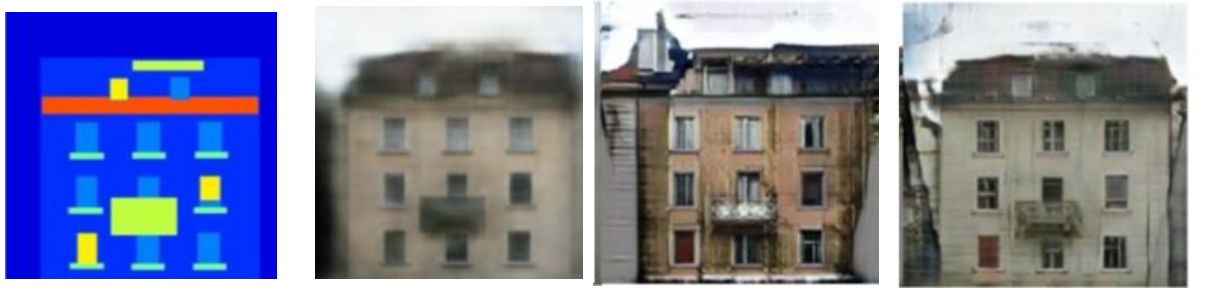


Image translation, or pix2pix

Conditional GAN



Testing:



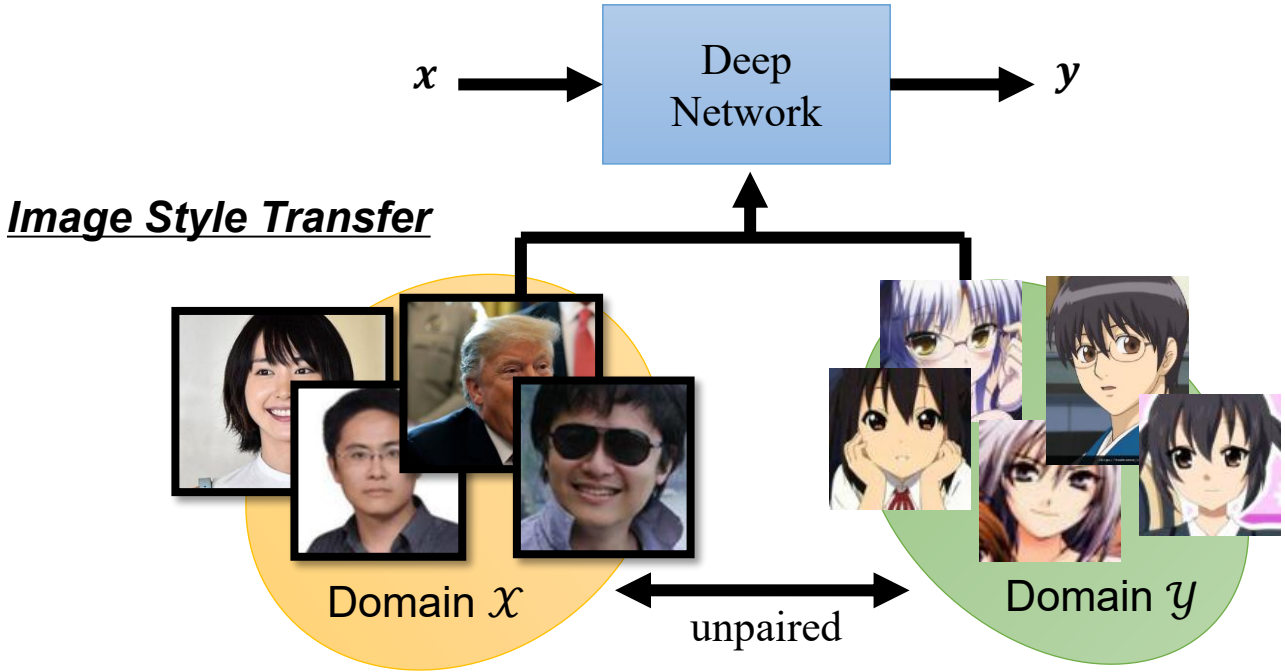
input supervised GAN GAN + supervised

GAN生态：无监督转换 (CycleGAN)

- 如何在没有成对的训练数据的两个领域之间进行图像转换？
  - 如，马 -> 斑马，但没有“同一匹马”和“其对应的斑马”照片
- 核心思想：循环一致性 (Cycle Consistency)
  - 双向生成: 训练两个生成器，$G_{X\to Y}$ 和 $G_{Y\to X}$，以及两个对应的判别器 D_Y 和 D_X
  - 对抗损失: $G_{X\to Y}$ 试图生成 D_Y 无法分辨的“假Y”；$G_{Y\to X}$ 试图生成 D_X 无法分辨的“假X”
  - 循环一致性损失: 一个从X域转换到Y域的图像，应该能被“转换回来”并恢复原状
    - Forward cycle: $G_{Y\to X}(G_{X\to Y}(x)) \approx x$
    - Backward cycle: $G_{X\to Y}(G_{Y\to X}(y)) \approx y$
    - 这个损失（通常是L1或L2范数）强制生成器保留原始图像的内容结构，只改变其风格
  - 总损失 = 对抗损失 + $\lambda \times$ 循环一致性损失

Learning from Unpaired Data

Learning from Unpaired Data



Can we learn the mapping without any paired data?

Unsupervised Conditional Generation

GAN生态：无监督转换 (CycleGAN)

- Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

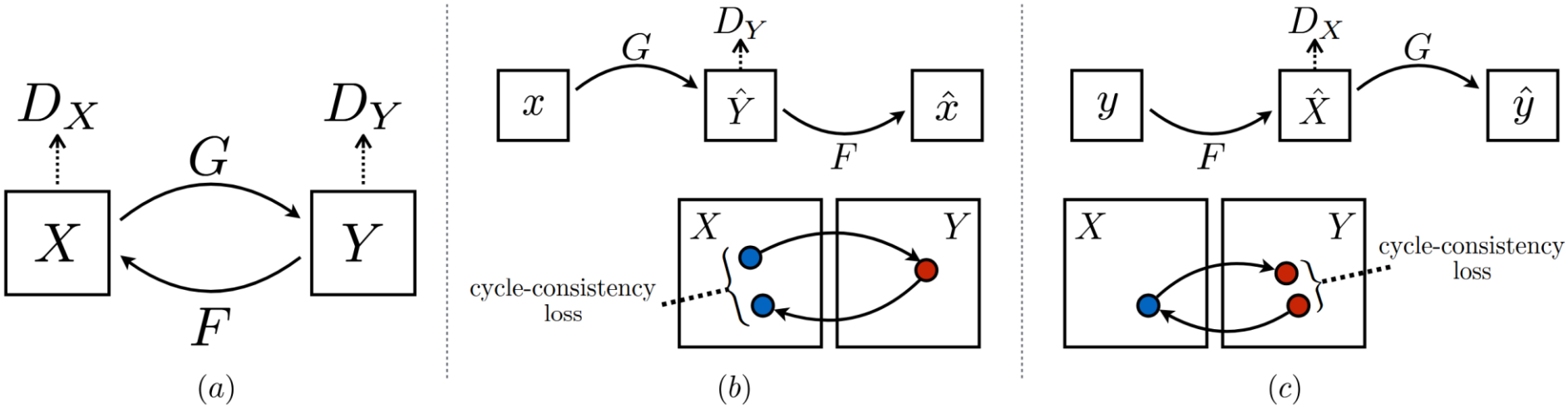
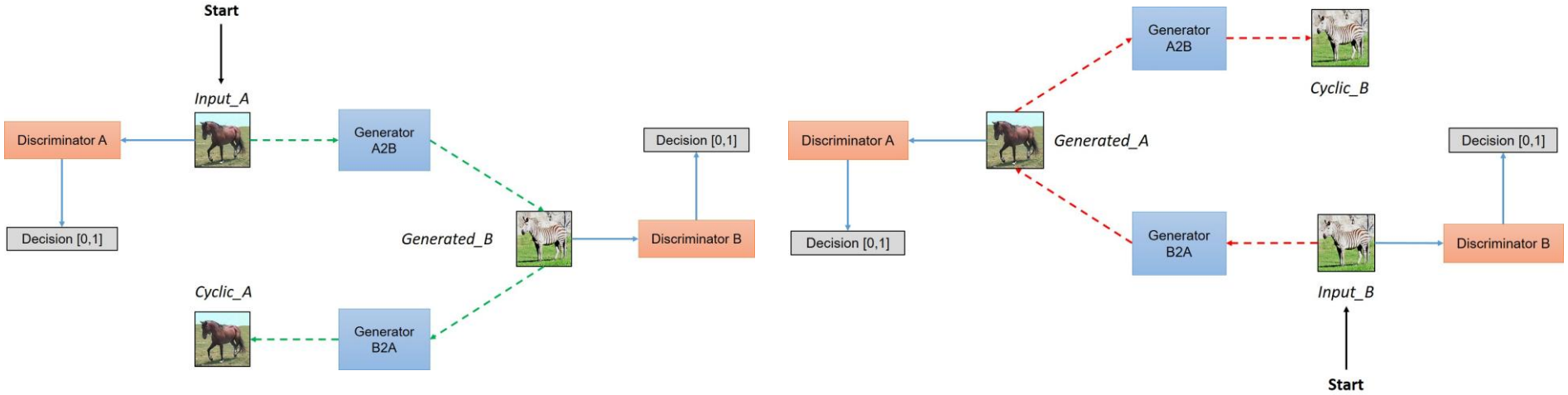
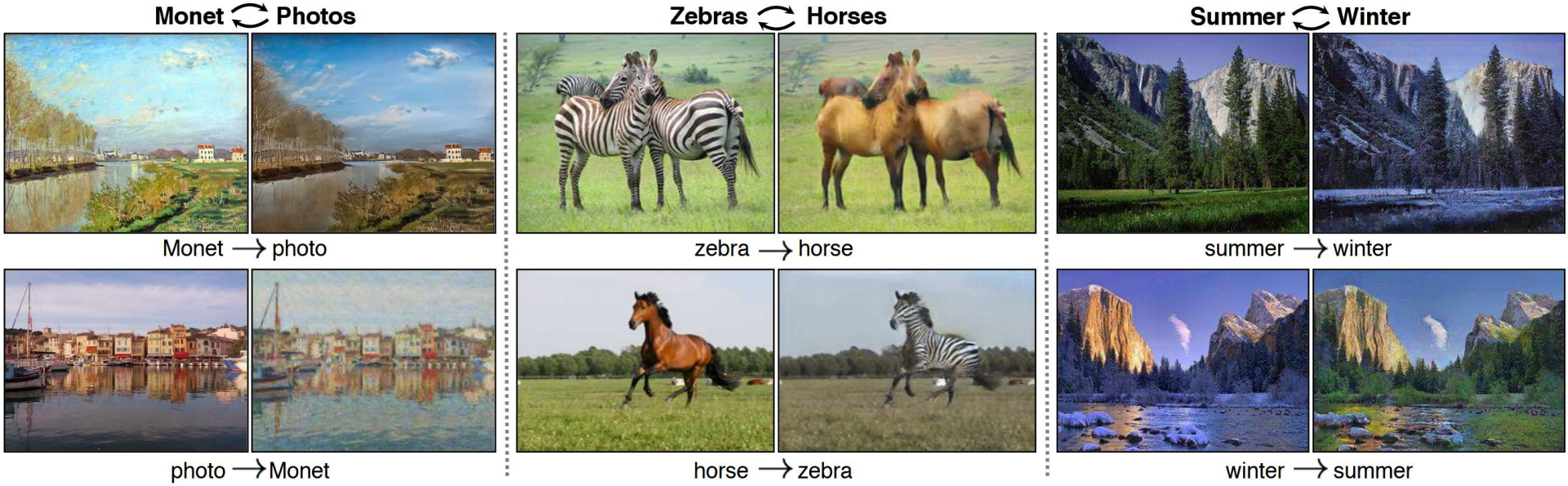


Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X, F, and X. To further regularize the mappings, we introduce two "cycle consistency losses" that capture the intuition that if we translate from one domain to the other and back again we should arrive where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

网络框架

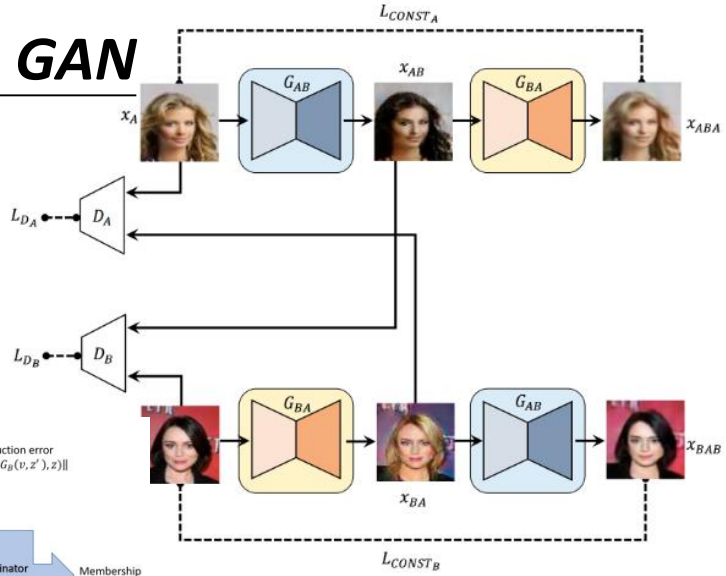

Start
Input_A
Discriminator A
Generator A2B
Decision [0,1]
Decision [0,1]
Generated_B
Discriminator B
Cyclic_A
Generator B2A
Generator A2B
Cyclic_B
Discriminator A
Generated_A
Decision [0,1]
Decision [0,1]
Generator B2A
Discriminator B
Input_B
Start

结果


Monet ⇄ Photos
Zebras ⇄ Horses
Summer ⇄ Winter
Monet → photo
zebra → horse
summer → winter
photo → Monet
horse → zebra
winter → summer

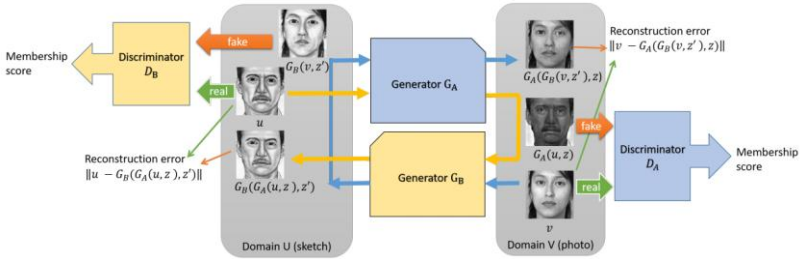
Disco GAN

https://arxiv.org/abs/1703.05192



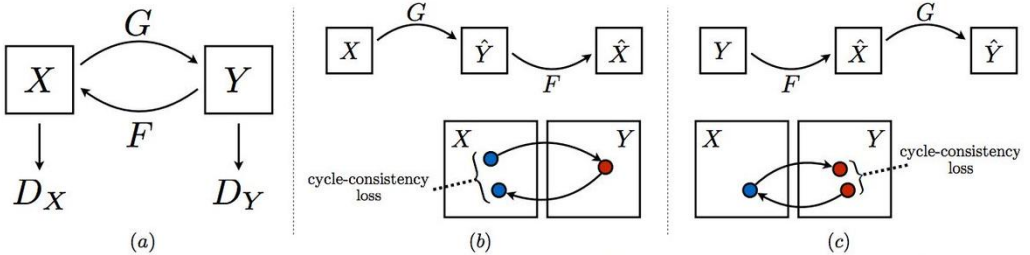
Dual GAN

https://arxiv.org/abs/1704.02510



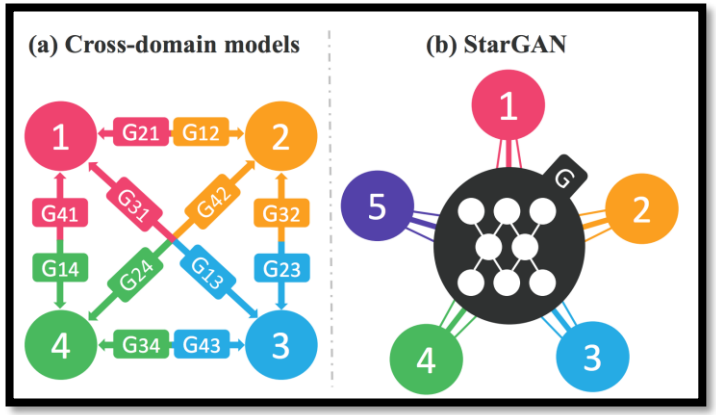
Cycle GAN

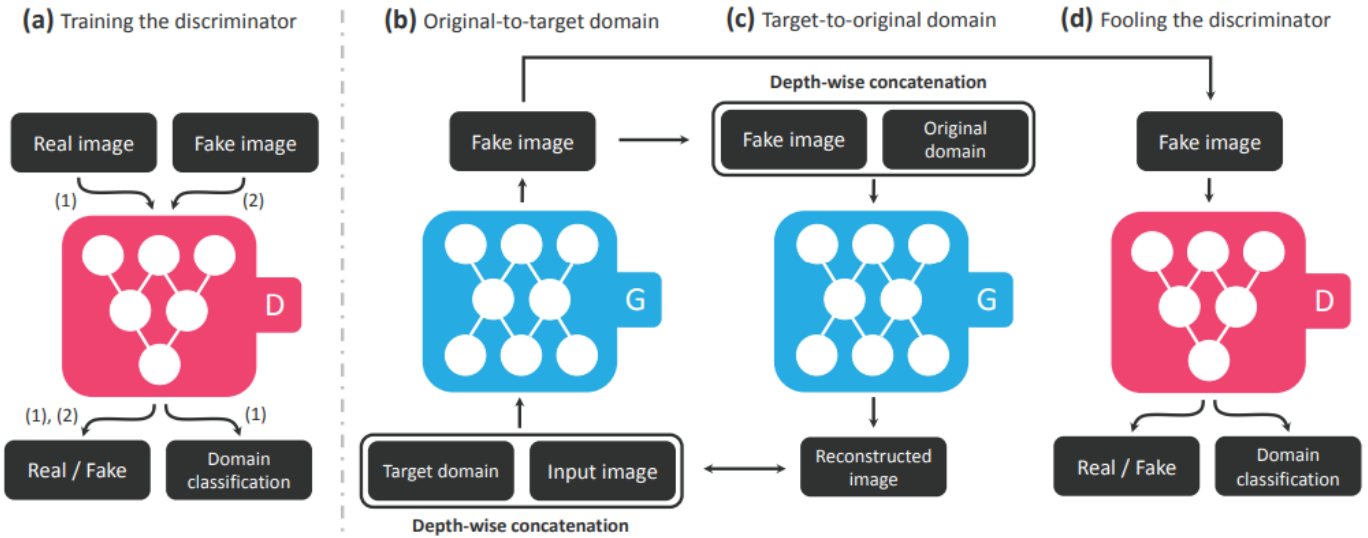
https://arxiv.org/abs/1703.10593



StarGAN

https://arxiv.org/abs/1711.09020





如何评价GAN的好坏？

- 评估生成模型的质量和多样性是一个开放性难题
- Inception Score (IS) - 越高越好
  - 使用一个在ImageNet上预训练的Inception网络
    - 质量: 对于一个清晰的生成图像 x，其类别预测分布 $p(y|x)$ 应该具有低熵（即分类器很确定它是什么）
    - 多样性: 对于所有生成图像，其类别预测的边缘分布 $p(y) = \int p(y|x = G(z))dz$ 应该具有高熵（即生成的图像涵盖了多种类别）
  - $IS = \exp(\mathbb{E}_{x\sim P_G}[KL(p(y|x)||p(y))])$
  - 缺点: 对噪声和模式坍塌敏感，不与人类感知强相关
- Fréchet Inception Distance (FID) - 分数越低越好
  - 比较真实图像和生成图像在Inception网络特征空间中的分布差异
    - 将真实图像和生成图像输入Inception网络，提取某个中间层的激活特征
    - 将两组特征向量分别建模为多元高斯分布，计算它们的均值 (μ_r, μ_g) 和协方差矩阵 (Σ_r, Σ_g)
    - 计算这两个高斯分布之间的Fréchet距离
  - $FID = \| \mu_r - \mu_g \|_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$
  - 优点: 对模式坍塌更敏感，与人类感知的一致性更好。目前是评估图像生成质量的黄金标准
- 其它量化指标
  - Perceptual Path Length (PPL)：衡量生成器潜在空间(latent space)的平滑性
  - Kernel Inception Distance (KID)：使用核方法（kernel method）来计算距离

Evaluation of Generation

永恒的挑战：如何评价GAN

- 评估生成模型的质量和多样性是一个开放性难题。
- Inception Score (IS) - 越高越好
  - 使用预训练的Inception网络，同时衡量质量和多样性。
  - 质量: 清晰图像的类别预测 $p(y|x)$应有低熵 (分类器很确定)。
  - 多样性: 所有生成图像的平均类别预测 $(p(y|x) \parallel p(y))$应有高熵 (涵盖多种类别)
    - $\text{IS} = \exp\left(\mathbb{E}_{x\sim P_G}[D_{KL}(p(y|x) \parallel p(y))]\right)$
- Fréchet Inception Distance (FID) - 越低越好 (黄金标准)
  - 比较真实图像和生成图像在Inception网络特征空间中的分布差异
  - 将两组特征向量分别建模为多元高斯分布，计算它们的均值(μ_r, μ_g)和协方差矩阵(Σ_r, Σ_g)
  - 计算这两个高斯分布之间的Fréchet距离
    - $\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)$
  - 对模式坍塌更敏感，与人类感知的一致性更好

评估指标的局限性

- IS的缺陷：无法有效检测类内模式坍塌
  - 例如：一个模型只生成ImageNet中“狗”这一类的1000张不同图片。它的IS分数会很高（质量高，类内多样性好），但全局多样性极差（没生成猫、车等）
- 模式坍塌 (Mode Collapse)
  - 生成器只学会了生成少数几个看起来真实的样本
- 记忆/过拟合 (Memorization)
  - 生成器只是记住了训练集样本，而不是学习其分布
  - 好的评估指标(如FID)和检查方法(如寻找最近邻)可以检测这种行为

总结与展望

- GAN的核心贡献:
  - 提出一种全新的、通过对抗博弈进行学习的生成模型框架，避免了直接计算棘手的似然函数
  - 在图像生成等领域取得了革命性的成果，生成图像的逼真度达到了前所未有的水平
- 核心挑战与未来方向
  - 训练稳定性
  - 评估: 自动、可靠且与人类感知高度一致的评估指标仍是研究热点
  - 可控性与可解释性: 如何更精细、更解耦地控制生成过程
  - 新领域应用: 将GAN的思想推广到文本、音频、3D模型、科学模拟等更广泛的领域
- GAN开启了一个充满无限可能的“生成时代”