

第六讲： 生成模型

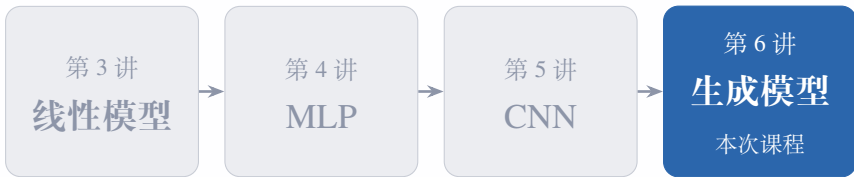
学习图像分布与逐步去噪

廖振宇

电子信息与通信学院
华中科技大学
2026 年秋季学期



本讲在课程中的位置



前面三讲：从图像预测标签（有监督）；本讲学习图像分布并生成新样本（无监督）

本讲内容

01

生成模型究竟学习什么

02

模型怎样从噪声生成图像

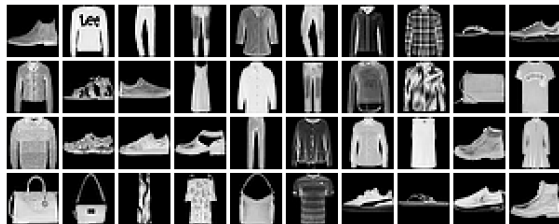
03

怎样评价生成结果

1

生成模型究竟学习什么

1 从图像识别到生成新图像？



Fashion-MNIST 测试集前 40 张图像

- 第 3 讲：线性分类基线
- 第 4 讲：非线性表示
- 第 5 讲：图像空间结构
- 第 6 讲：图像分布与新样本

思考

无监督学习：去掉类别标签后，如何“学习”生成图像？

分类任务

$$p_{\theta}(y | \mathbf{x})$$

图像 $\mathbf{x} \rightarrow$ 类别 y



输入图像 $\mathbf{x}$

已训练的分类器
 $p_{\theta}(y | \mathbf{x})$

输出类别概率
如判断为“短靴”

生成任务

$$p_{\theta}(\mathbf{x})$$

学习后采样 $\mathbf{x}'$



多张训练图像

学习

图像分布 $p_{\theta}(\mathbf{x})$



采样

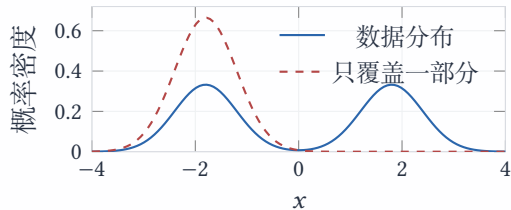


采样生成的新图像

注

包含：无条件分布 $p_{\theta}(\mathbf{x})$ 和类别条件分布 $p_{\theta}(\mathbf{x} | y)$

1 评估：生成质量与分布覆盖



- **质量**：单张图像是否合理？
- **覆盖**：不同类别是否都有？类内是否有多多样性？
- **记忆**：是否只复现训练样本？

思考

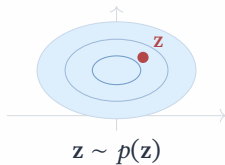
每张鞋都很清晰，但只生成一种鞋，是否算学到了服饰分布？

1 生成模型的主要思路与联系

方法	核心过程	区别与联系
AE	$\mathbf{x} \rightarrow \text{编码} \rightarrow \mathbf{z} \rightarrow \text{解码} \rightarrow \hat{\mathbf{x}}$	输入图像提供重建目标 不指定采样分布
VAE	$\mathbf{z} \sim p(\mathbf{z}) \rightarrow \text{解码器} \rightarrow \mathbf{x}$ <small>↓ 加先验约束</small>	AE + 先验约束 生成时从给定先验采样
GAN	$\mathbf{z} \sim p(\mathbf{z}) \rightarrow \text{生成器} \rightarrow \mathbf{x}$	判别器提供真假反馈 采样时只使用生成器
扩散模型	噪声 $\mathbf{x}_T \rightarrow \text{多步去噪} \rightarrow \dots \rightarrow \mathbf{x}_0$	人工加噪提供训练目标 学会逐步生成（本讲重点）

其他路线：自回归逐个预测像素或 token，归一化流（normalizing flow）用可逆变换“连接”分布
方法可以组合：潜空间扩散 = AE 学习压缩表示 + 扩散生成潜变量 + 解码成图像

简单的潜变量分布 (e.g. 高斯)

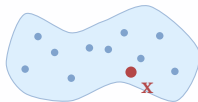


采样得到 z

学习得到的解码器
 $p_{\theta}(\mathbf{x} | z)$

再生成 $\mathbf{x}$

复杂的图像分布

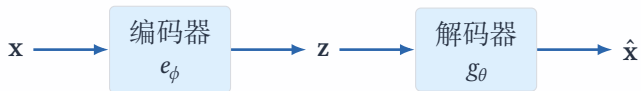


$$\text{从潜变量到图像分布: } p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x} | z) p(z) dz$$

结论

生成模型需要：“从哪里采样” + “怎样形成图像”

1 自编码器 Autoencoder (AE)



$$z = e_{\phi}(x) \quad \hat{x} = g_{\theta}(z) \quad \text{“重构” 损失函数: } L = \|x - \hat{x}\|_2^2$$

- 训练目的: 重建输入图像 x
- 编码器 e_{ϕ} 学习表示 ($x \rightarrow z$), 解码器 g_{θ} 学习重建 ($z \rightarrow x$)
- 通过反向传播共同更新两个网络

思考

问题: 如何生成? 训练后随便取一个高斯随机向量, 解码器能输出合理图像吗?

1 问题：AE 没有指定采样分布

$$\text{AE 损失函数: } L_{\text{AE}} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - g_{\theta}(e_{\phi}(\mathbf{x}_i))\|_2^2$$

- 损失只保证了训练样本的**重建**效果，**不保证**潜变量分布 $\mathbf{z}$
- 生成时，任意采样可能落在训练区域以外，导致无法有效生成

结论

好的重建结果不足以保证从指定先验采样也有好的生成结果

思考

能否在训练同时，让编码分布同时接近一个**已知**、**容易采样**的分布？

1 VAE: 用变分推断近似后验

解决方案: 变分 (variational) 自编码器 VAE

问题: 给定图像, 推断潜变量

$$\text{贝叶斯定理: } p_{\theta}(\mathbf{z} | \mathbf{x}) = \frac{p_{\theta}(\mathbf{x} | \mathbf{z})p(\mathbf{z})}{p_{\theta}(\mathbf{x})}$$

难点: 分母需要对潜变量积分, 难算

思路: 用易计算的分布近似

$$q_{\phi}(\mathbf{z} | \mathbf{x}) \approx p_{\theta}(\mathbf{z} | \mathbf{x})$$

- 一个例子: $q_{\phi}(\mathbf{z} | \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\phi}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}_{\phi}^2(\mathbf{x})))$, $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$
- 编码器学习: 均值 $\boldsymbol{\mu}_{\phi}(\mathbf{x})$ 与方差 $\boldsymbol{\sigma}_{\phi}^2(\mathbf{x})$
- **训练:** 从 $q_{\phi}(\mathbf{z} | \mathbf{x})$ 采样并重建; **生成:** 直接从先验 $p(\mathbf{z})$ 采样

来源: Kingma 与 Welling, Auto-Encoding Variational Bayes

定义 | 重参数化

将依赖参数的随机采样转化为独立噪声与参数的可微函数，并保持采样分布不变

$$\epsilon \sim \mathcal{N}(\mathbf{0}, I) \quad \mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \epsilon \sim \mathcal{N}(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}_\phi^2))$$

```
mu, logvar = encoder(x)
std = exp(0.5 * logvar)
eps = standard_normal(shape=mu.shape)
z = mu + std * eps
logits = decoder(z)
loss = vae_loss(logits, x, mu, logvar)
loss.backward()
```

一次前向采样后

1. 本次反传固定 ϵ
2. 梯度回传：均值与尺度

$$\frac{\partial z_j}{\partial \mu_j} = 1 \quad \frac{\partial z_j}{\partial \sigma_j} = \epsilon_j$$

3. 下次前向重新采样噪声

定义 | ELBO

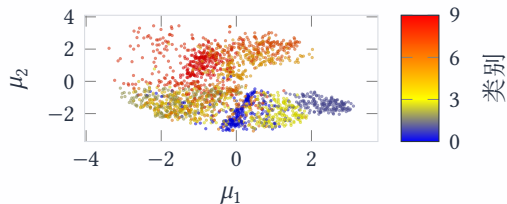
Evidence Lower Bound (证据下界), 是图像对数 (边缘) 似然的可优化下界

$$\log p_{\theta}(\mathbf{x}) = \underbrace{-L_{\text{VAE}}}_{\text{ELBO}} + \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}) \| p_{\theta}(\mathbf{z} | \mathbf{x}))}_{\geq 0}$$

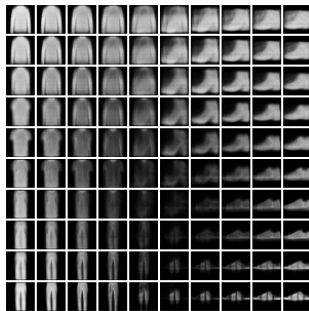
$$L_{\text{VAE}} = \underbrace{-\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x} | \mathbf{z})}_{\text{重建项}} + \underbrace{D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}) \| p(\mathbf{z}))}_{\text{先验约束}}$$

- **重建项**: 从 q_{ϕ} 采样 $\mathbf{z}$, 计算原图的负对数似然, 以一次采样估计期望
- **KL 项**: 衡量编码分布 $q_{\phi}(\mathbf{z} | \mathbf{x})$ 与先验的差异
 - 对角高斯编码分布与标准高斯先验下 $D_{\text{KL}} = \frac{1}{2} \sum_j (\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1)$

1 二维 VAE 潜空间与解码网格



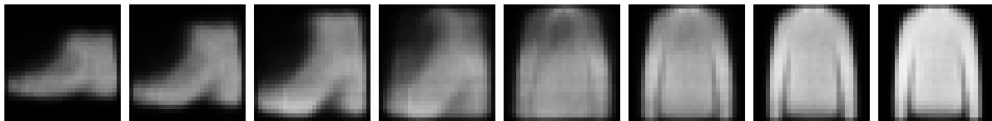
Fashion-MNIST 图像的编码表示，颜色区分类别



$z_1, z_2 \in [-2, 2]$ 解码均值

- 左图：将真实图像送入编码器，得到潜空间表示
- 右图：在二维网格取 z ，通过解码器的结果

$$\mathbf{z}(r) = (1 - r)\mathbf{z}_a + r\mathbf{z}_b, r \in [0, 1]$$

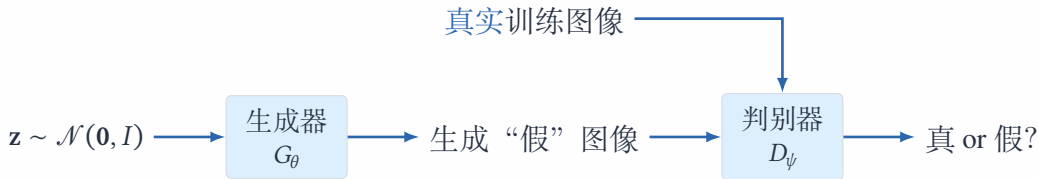


- 先编码两张图像得到潜空间的 $\mathbf{z}_a, \mathbf{z}_b$ ，再把插值点分别送入同一个解码器

思考

另一个方案：能否让另一个网络学习“像不像真实服饰”，用它的反馈指导生成？

GAN: Generative Adversarial Network 结构



- 判别器 D_ψ 区分真实数据与当前生成结果
- 生成器 G_θ 通过判别器的反馈更新参数 θ ，不断对抗，提升生成性能
- 生成时：采样 $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$ ，只用生成器，一次前向传播形成图像

来源：Goodfellow 等，Generative Adversarial Nets

1 GAN 的“对抗”损失函数

$$\min_{\theta} \max_{\psi} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \log D_{\psi}(\mathbf{x}) + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \log(1 - D_{\psi}(G_{\theta}(\mathbf{z})))$$

判别器损失函数

$$L_D = -\mathbb{E}_{\mathbf{x}} \log D_{\psi}(\mathbf{x}) - \mathbb{E}_{\mathbf{z}} \log(1 - D_{\psi}(G_{\theta}(\mathbf{z})))$$

生成器损失函数

$$L_G = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \log(1 - D_{\psi}(G_{\theta}(\mathbf{z}))) \Rightarrow L_G = -\mathbb{E}_{\mathbf{z}} \log D_{\psi}(G_{\theta}(\mathbf{z}))$$

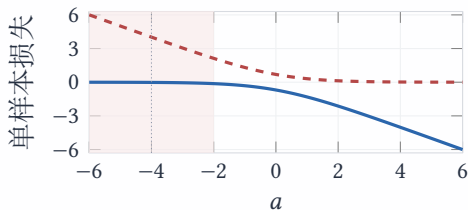
- 判别器：希望真实图像（概率）得分接近 1，生成图像得分接近 0
- 生成器：“混淆”判别器，生成的“假”图像得分增加

生成器损失与梯度：非饱和损失

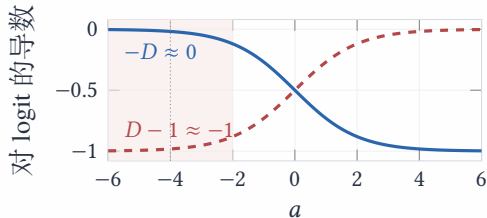
对一个生成样本，记 $D = \text{Sigmoid}(a)$ ， a 为判别器 logit

原始 $L_{\text{sat}} = \log(1 - D)$ 非饱和 $L_{\text{ns}} = -\log D$

损失曲线



反馈梯度



训练初期，得分低，e.g. $a = -4$ 时， $D \approx 0.018$ ：原始导数约 -0.018 ，非饱和导数约 -0.982

结论

低得分区域内，原始损失平坦；非饱和损失仍给出较大的局部梯度

```
fake = G(z)                # 从随机噪声生成假图像
loss_D = BCE(D(real), ones) + BCE(D(fake.detach()), zeros) # detach 截断到 G 的梯度
opt_D.zero_grad()          # 清除 D 的旧梯度
loss_D.backward()           # 计算 D 参数的梯度，真标 1、假标 0
opt_D.step()                # 仅更新 D 参数

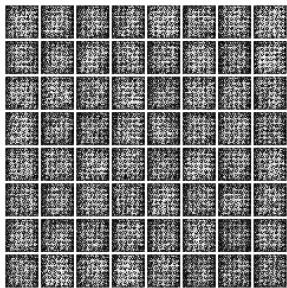
freeze_parameters(D)        # 固定 D 参数，保留对输入的梯度
loss_G = BCE(D(G(new_z)), ones) # 目标标 1，推动生成图像被判真
opt_G.zero_grad()           # 清除 G 的旧梯度
loss_G.backward()            # 梯度经过 D，传回 G
opt_G.step()                 # 仅更新 G 参数
unfreeze_parameters(D)      # 恢复 D 参数求导，供下一轮训练
```

注意

1. 更新判别器 D 时，**截断**生成器 G 的梯度
2. 更新生成器 G 时，梯度需**穿过**判别器 D

1 GAN 生成过程和样本展示

第 1 轮



第 10 轮



第 30 轮



思考

评估：不同样本是否重复？有哪些生成失败？

1 模式坍塌与训练不稳定

模式坍塌：同一时刻看多样性

不同随机输入产生相近图像



20k 步：样本多样性不足

训练不稳定：跨训练时刻看变化

本例在少数重复形态之间切换



10k 步



20k 步



50k 步

更新可能振荡或失衡，未形成稳定覆盖

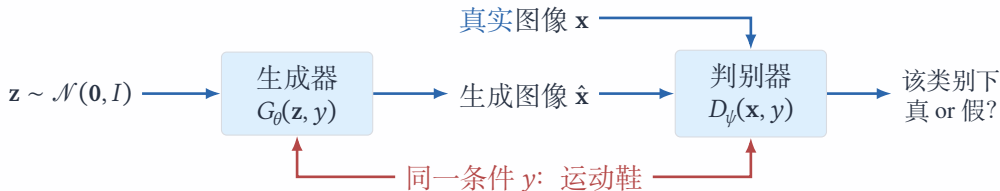
注

两种问题可以同时出现：模式坍塌关注多样性，训练不稳定关注训练动态

来源：Metz 等，Unrolled GAN，图 3 下排局部；MNIST，RNN 生成器 + CNN 判别器，0 步展开

条件 GAN：类别同时输入生成器与判别器

- 目标：希望指定生成“运动鞋”：将类别 y 加入原有 GAN 结构



- 生成器：用 y 控制类别，用随机变量 z 表达同类图像的变化
- 判别器：比较真实与生成的图像—条件配对，学习条件下的真假判断

思考

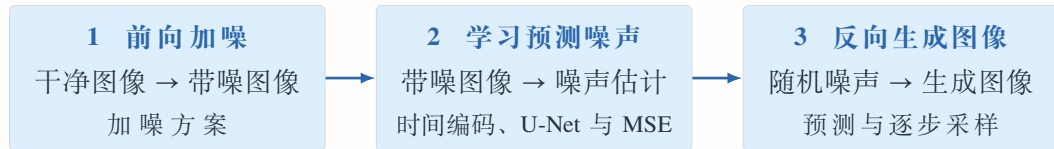
能否只构造一个网络，逐步生成图像？

来源：Mirza 与 Osindero, Conditional Generative Adversarial Nets

2

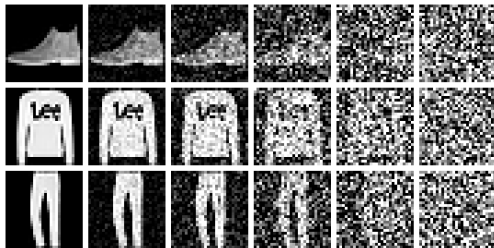
模型怎样从噪声生成图像





学习目标

能够说明训练样本怎样构造、模型如何训练，生成时怎样调用模型



同一图像，按 $t = 0, 10, 25, 50, 75, 100$ ，逐步加入越来越强的噪声

- 随时间的前向加噪过程

真实数据: $\mathbf{x}_0 \sim p_{\text{data}}$ $\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, I)$

- $\beta_t \in (0, 1)$ 是第 t 步新加入的噪声方差
- 核心：每次加噪过程 $q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t I)$

来源：Ho 等，DDPM

2 从一步、两步，到任意步加噪

$$\boxed{\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \epsilon_t} \quad \alpha_t = 1 - \beta_t$$

1. 展开两步加噪

$$\mathbf{x}_2 = \sqrt{\alpha_2 \alpha_1} \mathbf{x}_0 + \sqrt{\alpha_2 \beta_1} \epsilon_1 + \sqrt{\beta_2} \epsilon_2$$

2. 合并独立高斯噪声：线性组合仍为高斯，方差相加

$$\text{Var}(\text{累计噪声}) = (\alpha_2 \beta_1 + \beta_2) I = (1 - \alpha_2 \alpha_1) I$$

3. 推广到任意步： $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, $\bar{\alpha}_0 = 1$

$$\boxed{\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, I) \quad \Rightarrow q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) I)$$

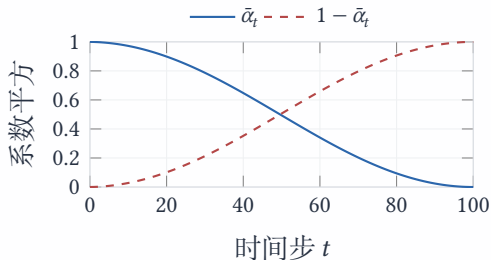
结论

训练时随机选 t ，一步合成带噪图像，无需逐步加噪

ϵ 是累计噪声，须与前向每一步的 ϵ_t 区分

2 噪声设计：余弦加噪方案

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t)I)$$

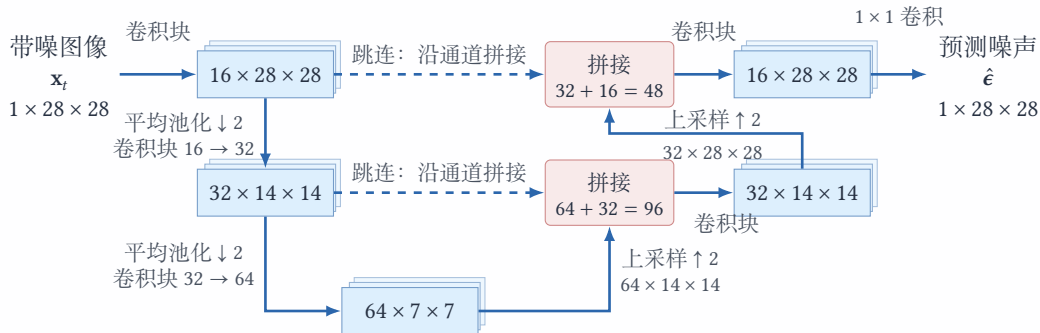


- 要求：逐步减少信号，增加噪声
- 需要 $\bar{\alpha}_T$ 足够接近 0，才能用标准高斯 $\mathcal{N}(0, I)$ 近似终点分布
- 本例采用 **100 步余弦加噪方案**
- 裁切后 $\beta_T = 0.999$ ，实际 $\bar{\alpha}_T \approx 2.43 \times 10^{-7}$

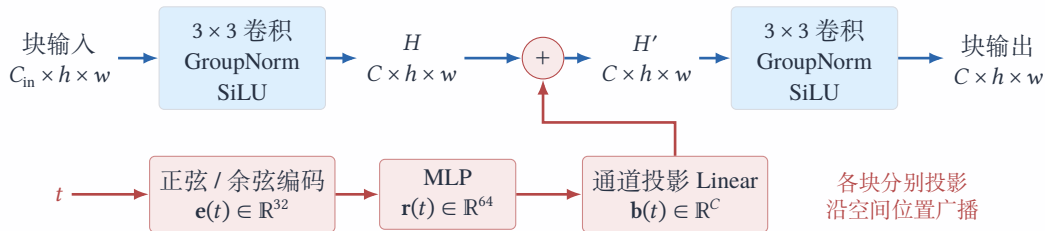
$$\underbrace{(\mathbf{x}_t, t)}_{\text{输入：带噪声图像}} \mapsto \underbrace{\hat{\epsilon} = \epsilon_{\theta}(\mathbf{x}_t, t)}_{\text{预测噪声}}$$

1. 真实数据中取得干净图像 $\mathbf{x}_0$
2. 随机选择时间步 t ，生成独立高斯噪声 ϵ （真实所加噪声已知）
3. 用累计公式合成 $\mathbf{x}_t$ 输入
 - 时间步信息 t 编码后输入
$$e(t) = [\sin(\omega_1 t), \cos(\omega_1 t), \dots, \sin(\omega_m t), \cos(\omega_m t)], H' = H + \text{Linear}(e(t))$$
 - t 作为特征，标明当前信号与噪声的相对强度
 - 多个频率提供不同尺度的时间表示，所有时间步共享参数 θ

用 32 维正弦/余弦编码，经 MLP 得到 64 维时间表示



- 下采样整合较大范围，上采样恢复分辨率
- 跳连将同分辨率的编码特征送入解码端



- H 是第一组卷积、归一化和激活得到的中间特征图
- $H'_{c,i,j} = H_{c,i,j} + b_c(t)$ 加入时间信息：同一通道的所有空间位置加同一个值，相加后特征尺寸不变

2 损失函数：噪声预测损失

$$L(\theta; \mathbf{x}_0, t, \epsilon) = \frac{1}{d} \left\| \epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|_2^2$$

- ϵ : 采样得到的已知噪声
- $\epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)$: 去噪网络针对带噪声输入对应的输出噪声预测
- 随机均匀采样 $t \sim \text{Uniform}\{1, \dots, T\}$
- MSE 平均空间全部像素 ($d = 28 \times 28 = 784$) 和 batch 样本, 反向传播更新模型

来源: Ho 等, DDPM, 简化训练目标

2 扩散模型的一次训练更新

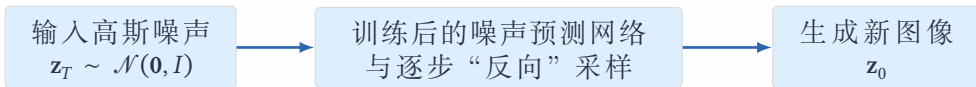
```
x0 = images * 2 - 1          # 像素从 [0, 1] 映射到 [-1, 1]
t = torch.randint(1, T + 1, (B,), device=device) # 每张图随机选时间步
eps = torch.randn_like(x0)   # 已知高斯噪声：用于加噪及监督
a = alpha_bar[t].view(B, 1, 1, 1) # 取累计系数，广播到各通道与像素
xt = a.sqrt() * x0 + (1 - a).sqrt() * eps # 直接合成第 t 步带噪图像
eps_hat = model(xt, t)       # 输入图像与时间步，预测噪声
loss = F.mse_loss(eps_hat, eps) # 对 batch 内全部图像和像素取平均
optimizer.zero_grad()        # 清除上次更新的梯度
loss.backward()               # 反向传播计算参数梯度
optimizer.step()              # 更新去噪网络参数
```

注意区别

$$\alpha_t = 1 - \beta_t \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s \quad \bar{\alpha}_0 = 1$$

- β_t 为单步噪声方差；单步信号乘 $\sqrt{\alpha_t}$ ，从原图直接乘 $\sqrt{\bar{\alpha}_t}$

目标：从随机噪声生成新图像



- 通过噪声预测可以估计干净图像

$$\text{由 } \mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon} \implies \hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}$$

- 从加噪图像中扣除预测的噪声分量，再校正信号缩放，得到干净图像的估计

- 噪声预测有已知监督，可通过神经网络实现

- 训练时自行采样 $\boldsymbol{\epsilon}$ 、构造加噪图像 $\mathbf{x}_t$ ，用已知 $\boldsymbol{\epsilon}$ 作为 MSE 监督目标
- 网络只看到 $(\mathbf{x}_t, t)$ ，需学习图像中的轮廓、纹理等规律以估计噪声

训练加噪状态用 $\mathbf{x}_t$ ，生成状态用 $\mathbf{z}_t$

2 噪声预测如何用于逐步生成？

- 最小化 MSE 学到条件均值

$$\epsilon_{\theta}(\mathbf{x}_t, t) \approx \mathbb{E}[\epsilon | \mathbf{x}_t, t] \implies \hat{\mathbf{x}}_0 \approx \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t, t]$$

- 条件均值是对“平均图”的估计，不等同于从图像分布中采样

- 标准 DDPM：用噪声预测确定反向一步，再重复采样



- 起点：独立采样 $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, I)$ ；经过反向步逐步形成 $\mathbf{z}_t$

- 计算均值：结合状态 $\mathbf{z}_t$ 与系数 α_t, β_t ：

$$\mu_{\hat{\theta}}(\mathbf{z}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\epsilon} \right)$$

- 噪声估计来自神经网络 $\hat{\epsilon} = \epsilon_{\hat{\theta}}(\mathbf{z}_t, t)$
- 系数 $\alpha_t, \beta_t, \bar{\alpha}_t$ 由加噪方案预先确定

- 高斯采样：围绕均值加入扰动

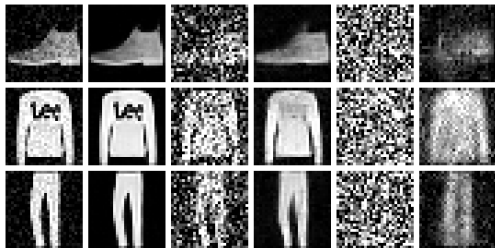
$$\mathbf{z}_{t-1} = \mu_{\hat{\theta}}(\mathbf{z}_t, t) + \sigma_t \boldsymbol{\eta}, \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, I)$$

- σ_t^2 是预设方差， σ_t 控制扰动幅度
- $\boldsymbol{\eta}$ 每步重新采样，与网络预测的噪声不同

将 $\mathbf{z}_{t-1}$ 传入下一轮，重复至 $\mathbf{z}_0$ ；最后一步方差 $\sigma_1 = 0$ ，无随机项

来源：Ho 等，DDPM，第 3.2 节与算法 2

2 单次噪声预测与图像估计



每对列为带噪输入与图像估计 $\hat{x}_0$ ，依次 $t = 10, 50, 90$

思考

为了实现“从噪声中生成近似真实图像”的目标，还需要做什么？

DDPM: Denoising Diffusion Probabilistic Models

去噪扩散概率模型：用训练得到的噪声预测器引导逐步生成

训练输入记为 $\mathbf{x}_t$ ，生成过程的图像状态记为 $\mathbf{z}_t$



前面的训练提供了什么？

- 训练得到噪声预测器 $\epsilon_{\hat{\theta}}(\cdot, t)$, $\hat{\theta}$ 为训练结束后的参数
- 生成时：固定 $\hat{\theta}$ ，每步输入 $(\mathbf{z}_t, t)$ ，用噪声预测计算均值，决定反向更新

$\hat{\mathbf{z}}_0$ 是干净图像估计， $\mathbf{z}_{t-1}$ 是本轮需要生成的下一步状态

来源：Ho 等，Denoising Diffusion Probabilistic Models

2 反向一步：分布与均值

用高斯条件分布建模反向一步过程

$$p_{\hat{\theta}}(\mathbf{z}_{t-1} | \mathbf{z}_t) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\theta}}(\mathbf{z}_t, t), \sigma_t^2 I) \quad t > 1$$

$$(\mathbf{z}_t, t) \xrightarrow{\text{训练后的网络}} \hat{\epsilon} \xrightarrow{+ \text{当前状态与时间步设计}} \boldsymbol{\mu}_{\hat{\theta}}(\mathbf{z}_t, t)$$

$$\boldsymbol{\mu}_{\hat{\theta}}(\mathbf{z}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_{\hat{\theta}}(\mathbf{z}_t, t) \right)$$

- 训练得到 $\boldsymbol{\epsilon}_{\hat{\theta}}$ ，缩放系数由预先设定的时间步给出
- $\boldsymbol{\mu}_{\hat{\theta}}$ 是下一步 $\mathbf{z}_{t-1}$ 的均值

来源：Ho 等，DDPM

基于均值的，从对应高斯分布中抽取下一步状态

$$\mathbf{z}_{t-1} = \mu_{\hat{\theta}}(\mathbf{z}_t, t) + \sigma_t \boldsymbol{\eta} \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, I)$$

- $\epsilon_{\hat{\theta}}$ 使用训练学到的规律预测噪声，决定分布中心
- $\boldsymbol{\eta}$ 是本步新抽取的随机数，用于围绕该中心采样
- 方差： $\sigma_t^2 = \tilde{\beta}_t = \beta_t \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}$

继续迭代直到生成完成

- $t > 1$ ：将 $\mathbf{z}_{t-1}$ 作为新输入，减小 t ，重复预测与采样过程
- $t = 1$ ：不再加入随机项，输出最终图像 $\mathbf{z}_0$

来源：Ho 等，DDPM

2 一个像素的反向更新算例

1. 给定系数、当前状态、噪声预测与新随机数

$$\alpha_t = 0.9 \quad \beta_t = 1 - \alpha_t = 0.1 \quad \bar{\alpha}_t = \sum_{s=1}^t \alpha_s = 0.36 \quad \bar{\alpha}_{t-1} = 0.4$$

当前状态: $z_t = 0.7$ 估计噪声: $\hat{\epsilon} = 0.2$ 新随机噪声: $\eta = -0.5$

2. 计算均值与标准差

$$\mu_{\hat{\theta}} = \frac{0.7 - (0.1/0.8) \times 0.2}{\sqrt{0.9}} \approx 0.7115 \quad \sigma_t = \sqrt{0.1 \times \frac{0.6}{0.64}} \approx 0.3062$$

3. 围绕均值随机采样下一步

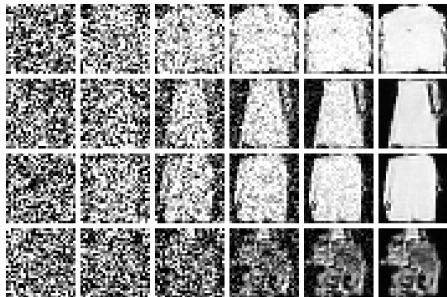
$$z_{t-1} = \mu_{\hat{\theta}} + \sigma_t \eta \approx 0.7115 + 0.3062 \times (-0.5) = 0.5584$$

结论

$\hat{\epsilon}$ 决定均值, η 是本步新采样的随机数

```
model.eval() # 已加载训练参数，生成时保持固定
with torch.no_grad():
    z = torch.randn(N, 1, 28, 28, device=device) # 初态  $z_T \sim N(0, I)$ 
    for step in range(T, 0, -1):
        t = torch.full((N,), step, device=device, dtype=torch.long)
        eps_hat = model(z, t) # 网络预测，不是新随机噪声
        mean = (z - beta[step] / (1 - alpha_bar[step])).sqrt()
            * eps_hat) / alpha[step].sqrt()
        if step > 1:
            eta = torch.randn_like(z) # 本步新采样  $\eta \sim N(0, I)$ 
            z = mean + posterior_var[step].sqrt() * eta
        else:
            z = mean
```

2 反向轨迹如何形成图像？

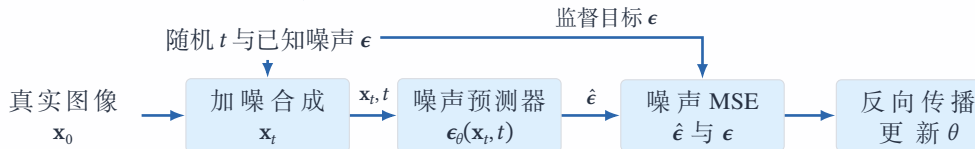


- 每行从独立采样 $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, I)$ 出发
- 六列对应“反向”时间步 $t = 100, 75, 50, 25, 10, 0$
- 过程中固定模型参数 $\hat{\theta}$ ，每步采用新的 $\mathbf{z}_t$ 与 t

思考

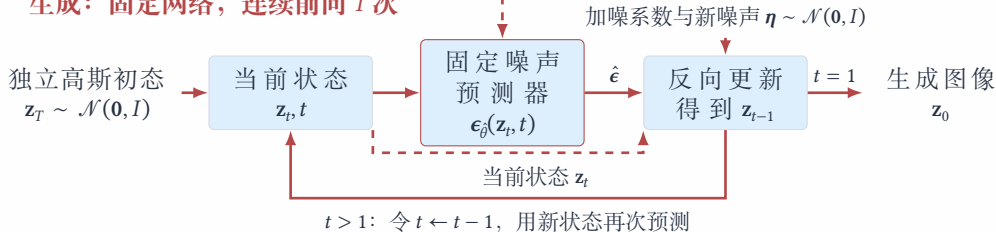
这是恢复一张（训练集中的）原图，还是生成一个新的样本？

训练：每次更新前向一次，学习参数 θ



训练完成：保存并固定参数 $\hat{\theta}$

生成：固定网络，连续前向 T 次



2 无条件生成为何不能指定服饰类别？

$$\mathbf{z}_0 \sim p_{\hat{\theta}}(\mathbf{z}_0) \quad p_{\text{data}}(\mathbf{x}) = \sum_y p_{\text{data}}(\mathbf{x} | y) p_{\text{data}}(y)$$

- 无条件模型模型：学习所有服饰混合形成的图像分布
- 随机改变初始噪声会改变样本，但未能指定类别

思考

前三讲的标签是预测目标，现在能否把它改为生成的**输入条件**？



每行指定同一类别

$$\hat{\epsilon} = \epsilon_{\hat{\theta}}(z_t, t, y)$$

- 类别嵌入与时间表示相加
- 各行依次指定T恤、长裤、凉鞋、运动鞋、包、踝靴
- 训练目标仍是噪声，类别只作条件

3

怎样评价生成结果



- 视觉质量检查：固定随机种子的完整样本网格与失败样本
- 训练、生成过程检查：相同初始噪声，比较训练阶段与反向轨迹
- 分布评价：FID 衡量差异，Precision/Recall 区分质量与覆盖
- 训练样本记忆检查：生成样本与训练集最近邻并排
- 类别覆盖：固定分类器，对相同数量样本统计预测类别
- 采样时间：同设备、同 batch、同样本数，预热后计时

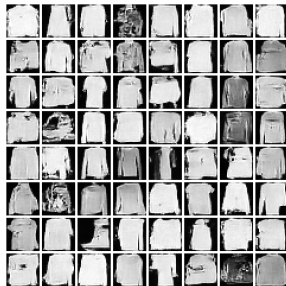
第 1 轮



第 10 轮

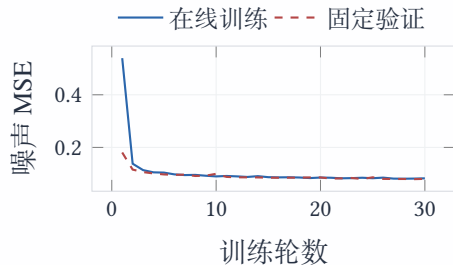
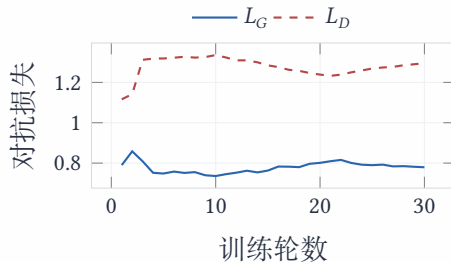


第 30 轮



3

训练曲线：两类损失分别解释

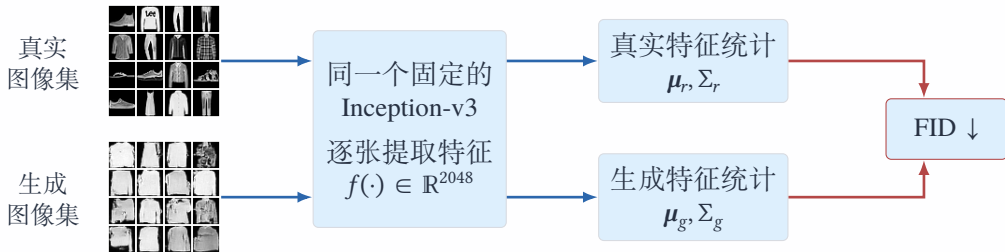


注意

扩散 MSE 与 GAN 对抗损失的数值不能直接比较

定义 | FID

Fréchet Inception Distance，比较真实与生成图像的特征统计



来源：Heusel 等，2017，TTUR 论文第 4 节提出 FID；图示为本讲真实测试图像与实际扩散输出各前 16 张，未计算 FID

$$\text{FID} = \underbrace{\|\mu_r - \mu_g\|_2^2}_{\text{特征中心差异}} + \underbrace{\text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2})^{1/2})}_{\text{特征散布差异}}$$

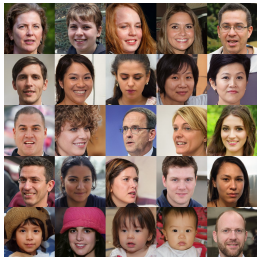
- r/g 对应真实/生成， Tr 为矩阵对角元素之和
- **均值 μ** ：描述特征中心；**协方差 Σ** ：描述变化幅度与维度相关性
- 将两组特征分别**近似**为高斯分布，计算两者的平方 Wasserstein-2 距离
- 相同评价协议下，FID 越低表示这两个高斯近似越接近

注意

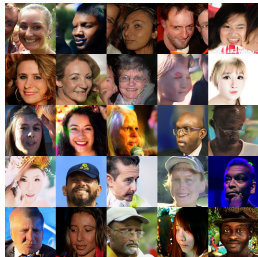
FID 为 0 只说明特征统计指标（均值与协方差）相同，不能保证完整图像分布相同

来源：Heusel 等，2017，FID；Chong 与 Forsyth，2020，有限样本估计偏差

配置 B: FID 16.9



配置 D: FID 16.7



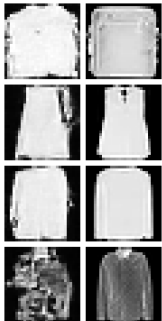
- **Precision ↑: 样本质量**
生成特征落在真实特征邻域内的比例
- **Recall ↑: 分布覆盖**
真实特征被生成特征邻域覆盖的比例
- **具体实现方案:** VGG-16 特征与 k 近邻估计区域

思考

两个 FID 几乎相同，两组图像的质量如何？

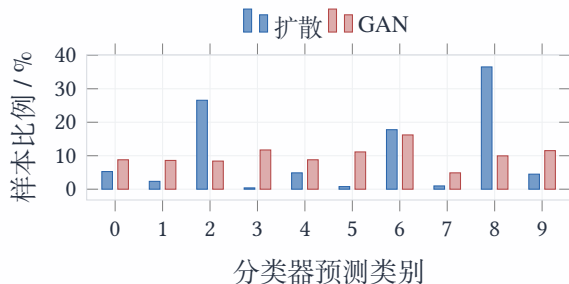
来源: Kynkäänniemi 等, 2019, Improved Precision and Recall Metric for Assessing Generative Models, 图 3

最近邻检查：生成样本是否复现训练图像？

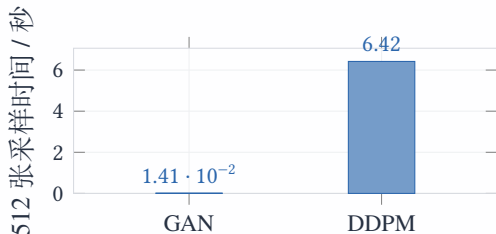


$$\text{NN}(\mathbf{z}_0) = \arg \min_{\mathbf{x}_i \in \mathcal{D}_{\text{train}}} \frac{1}{784} \|\mathbf{z}_0 - \mathbf{x}_i\|_2^2$$

- 左：固定序列的前 4 张生成样本 $\mathbf{z}_0$
- 右：全部 50000 张训练图像中的像素最近邻
- 比较轮廓、细节与像素距离



- 各生成 512 张，使用同一 CNN
- 类别不均衡：需进一步检查



- GAN: 每个 batch 一次生成器调用
- DDPM: 每个 batch 100 次去噪网络调用

	VAE	GAN	扩散模型
学习信号	重建与 KL	对抗判别反馈	多噪声强度的预测
生成起点	潜变量先验	潜变量先验	高斯图像状态
采样方式	一次解码	一次生成	逐步去噪
常见挑战	重建细节与先验约束	优化失衡与覆盖不足	采样代价与累计误差
条件输入	可扩展	生成器与判别器	去噪网络

- 普通 Autoencoder 重建良好，为何仍需说明采样分布？
- GAN 更新生成器时，为何不能切断判别器输入梯度？
- 扩散训练为何无需实际执行全部前向加噪步骤？
- 噪声估计、干净图像估计与反向状态有什么区别？
- FID、最近邻与类别覆盖分别能支持什么结论？
- 缩短采样步数时，为何还需要检查加噪方案和更新公式？

- **线性模型**：建立数据、模型、损失、优化与评价的闭环
- **MLP**：通过非线性表示拓展表达能力，用反向传播学习
- **CNN**：将图像空间结构写入连接与参数共享
- **生成模型**：学习数据分布，定义可采样的生成过程

结论

面对新问题，先定义任务与评价，再建立基线，用受控实验验证模型选择

附录导览

- **A | VAE 概率补充** 变分下界与高斯 KL
- **B | GAN 机制补充** 理想判别器、上采样与扩展方向
- **C | 扩散推导与扩展** 后验与 Score、余弦方案、采样与 CFG
- **D | 实验说明与参考** 复现配置、评价指标、计时与资料

附录 A

VAE 概率补充

变分下界与高斯 KL

A VAE 下界的直觉推导

1. 用编码器分布改写潜变量积分

$$\log p_{\theta}(\mathbf{x}) = \log \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z} | \mathbf{x})}$$

2. 利用对数的凹性，应用 Jensen 不等式

$$\log p_{\theta}(\mathbf{x}) \geq \mathbb{E}_{q_{\phi}} \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q_{\phi}(\mathbf{z} | \mathbf{x})}$$

3. 拆成重建项与先验约束

$$\text{ELBO} = \mathbb{E}_{q_{\phi}} \log p_{\theta}(\mathbf{x} | \mathbf{z}) - D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}) \| p(\mathbf{z}))$$

编码分布需覆盖积分所需的支持范围

A 对角高斯 KL 与重参数化实现

$$D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \| \mathcal{N}(\mathbf{0}, I)) = \frac{1}{2} \sum_j (\mu_j^2 + \sigma_j^2 - 1 - \log \sigma_j^2)$$

```
mu, logvar = encoder(x)
std = torch.exp(0.5 * logvar)
eps = torch.randn_like(std) # VAE 训练时的重参数化噪声
z = mu + std * eps
logits = decoder(z)
recon = F.binary_cross_entropy_with_logits(
    logits, x.flatten(1), reduction="none").sum(1)
kl = 0.5 * (mu.square() + logvar.exp() - 1 - logvar).sum(1)
loss = (recon + kl).mean()
```

重建项与 KL 都先按样本求和再按 batch 平均，不混用不同缩放口径

附录 B

GAN 机制补充

理想判别器、上采样与扩展方向

B 固定生成分布时的理想判别器

$$D^*(\mathbf{x}) = \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x}) + p_G(\mathbf{x})}$$

- 在任意函数、无限容量和精确优化的理想设定下得到
- 若 $p_G = p_{\text{data}}$, 有 $D^* = 1/2$
- 现实中弱判别器也可能给出接近 1/2 的输出
- 有限样本与有限优化无法直接保证理想平衡

注意

判别器无法区分是必要的理论直觉，需结合真实样本与覆盖检查验证

来源：Goodfellow 等，GAN，命题 1 与理论分析

B 转置卷积的输出尺寸

$$H_{\text{out}} = (H_{\text{in}} - 1)s - 2p + d(k - 1) + \text{output_padding} + 1$$

$$k = 4, s = 2, p = 1, d = 1, \text{output_padding} = 0 \implies H_{\text{out}} = 2H_{\text{in}}$$

- 课程生成器中 $7 \rightarrow 14 \rightarrow 28$ ，宽度使用同一公式
- 参数可学习，用于将低分辨率特征映射到高分辨率
- 转置指线性算子的转置，通常不是原卷积的逆
- 也可使用插值上采样后接普通卷积

来源：PyTorch ConvTranspose2d 官方文档

B GAN 的拓展方向

- **改变训练目标**: Wasserstein GAN 用不同分布距离改善反馈机制
- **控制生成**: cGAN 把类别同时输入生成器与判别器
- **规模与结构**: BigGAN、StyleGAN 探索条件生成和外观控制
- **潜变量变化**: 插值与方向编辑提供经验性控制线索

注

这些方法对应不同问题，不替代本讲对抗训练与生成评价的核心机制

来源: Wasserstein GAN

来源: BigGAN

来源: StyleGAN

附录 C

扩散推导与扩展

高斯后验、余弦方案与条件引导

C 累计噪声方差的递推

1. 假设前一步的条件分布，并执行一步加噪

$$\mathbf{x}_{t-1} \mid \mathbf{x}_0 \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0, (1 - \bar{\alpha}_{t-1})I)$$
$$\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}_t \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, I)$$

2. 计算条件均值：新噪声的均值为零

$$\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0] = \sqrt{\alpha_t \bar{\alpha}_{t-1}} \mathbf{x}_0$$

3. 计算条件方差：独立噪声的方差相加

$$\text{Var}(\mathbf{x}_t \mid \mathbf{x}_0) = (\alpha_t(1 - \bar{\alpha}_{t-1}) + \beta_t)I = (1 - \bar{\alpha}_t)I$$

利用 $\alpha_t + \beta_t = 1$ 完成递推， $\boldsymbol{\epsilon}_t$ 属于前向加噪

C 反向高斯后验的配方计算

1. 将后验写成两个高斯因子的乘积

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \propto q(\mathbf{x}_t \mid \mathbf{x}_{t-1})q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)$$

2. 收集平方项，得到精度

$$\text{精度} = \frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} = \frac{1 - \bar{\alpha}_t}{\beta_t(1 - \bar{\alpha}_{t-1})}$$

3. 收集一次项，配方得到均值与方差

$$\tilde{\beta}_t = \text{精度}^{-1} \quad \tilde{\boldsymbol{\mu}}_t = \tilde{\beta}_t \left(\frac{\sqrt{\alpha_t}}{\beta_t} \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} \mathbf{x}_0 \right)$$

本式适用于 $t > 1$ ； $t = 1$ 时，已知 $\mathbf{x}_0$ 的后验退化

C 高斯 KL 到带权噪声目标

1. 同方差高斯 KL 化为均值的平方距离

$$D_{\text{KL}}(\mathcal{N}(\tilde{\mu}_t, \sigma_t^2 I) \| \mathcal{N}(\mu_\theta, \sigma_t^2 I)) = \frac{\|\tilde{\mu}_t - \mu_\theta\|_2^2}{2\sigma_t^2}$$

2. 代入噪声预测参数化

$$\tilde{\mu}_t - \mu_\theta = \frac{\beta_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}}(\epsilon_\theta - \epsilon)$$

3. 整理得到带时间权重的噪声目标

$$D_{\text{KL}} = w_t \|\epsilon_\theta - \epsilon\|_2^2 \quad w_t = \frac{\beta_t^2}{2\sigma_t^2 \alpha_t(1 - \bar{\alpha}_t)}$$

方差选择不同但固定时，另有与 θ 无关的项；简化目标舍弃时间权重

C 条件均值与边缘 Score

$$\mathbf{m}_t = \mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{x}_t, t]$$

$$\mathbb{E}[\|\boldsymbol{\epsilon} - f\|_2^2 \mid \mathbf{x}_t, t] = \mathbb{E}[\|\boldsymbol{\epsilon} - \mathbf{m}_t\|_2^2 \mid \mathbf{x}_t, t] + \|f - \mathbf{m}_t\|_2^2$$

$$\begin{aligned}\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) &= \mathbb{E}_{q(\mathbf{x}_0 \mid \mathbf{x}_t)}[\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{x}_0)] \\ &= -\frac{\mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{x}_t, t]}{\sqrt{1 - \bar{\alpha}_t}}\end{aligned}$$

- 平方风险的可优化部分在条件均值处达到最小
- 在可交换求导与积分的条件下，条件 Score 平均得到边缘 Score

C 课程采样实现：裁切干净图像估计后重算均值

```
eps_hat = model(z, t) # 预测噪声，不是新随机噪声
z0_hat = (z - (1 - alpha_bar[i]).sqrt() * eps_hat)
z0_hat = (z0_hat / alpha_bar[i].sqrt()).clamp(-1, 1)
c1 = beta[i] * alpha_bar[i - 1].sqrt() / (1 - alpha_bar[i])
c2 = alpha[i].sqrt() * (1 - alpha_bar[i - 1]) / (1 - alpha_bar[i])
mean = c1 * z0_hat + c2 * z
z = mean
if i > 1:
    eta = torch.randn_like(z) # 本步新噪声  $\eta \sim N(0, I)$ 
    z = mean + posterior_var[i].sqrt() * eta
```

- 未裁切时，两种均值参数化代数等价
- 裁切改变了估计，因此需用后验系数重新组合
- 最终 $t = 1$ 时输出生成图像 z_0 ，不再加入噪声

C 余弦加噪方案：累计信号的设计

- 先设计 $\bar{\alpha}_t$ ：它是原图信号的**系数平方**， $1 - \bar{\alpha}_t$ 是累计噪声方差
- 在 $[0, \pi/2]$ 上， $\cos^2$ 从 1 平滑降到 0，可用来控制信号衰减

$$f(t) = \cos^2\left(\frac{t/T + s_0}{1 + s_0} \frac{\pi}{2}\right) \quad s_0 = 0.008$$

$$\boxed{\bar{\alpha}_t^{\text{目标}} = \frac{f(t)}{f(0)}} \quad \bar{\alpha}_0^{\text{目标}} = 1 \quad \bar{\alpha}_T^{\text{目标}} = 0$$

- **偏移** s_0 ：避免初始几步的噪声过小，本讲沿用论文取值
- **除以** $f(0)$ ：加入偏移后 $f(0) < 1$ ，归一化保证起点完整保留信号

注

余弦平方形式是人为选择的加噪方案，满足边界要求的曲线并不唯一

来源：Nichol 与 Dhariwal, Improved DDPM, 第 3.2 节、式 16

C 余弦加噪方案：单步噪声方差的推导

1. 从累计乘积解出单步方差： $\alpha_t = 1 - \beta_t$

$$\bar{\alpha}_t = \bar{\alpha}_{t-1} \alpha_t = \bar{\alpha}_{t-1} (1 - \beta_t) \implies \boxed{\beta_t = 1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}$$

2. 代入余弦目标曲线：归一化因子 $f(0)$ 抵消

$$\beta_t^{\text{候选}} = 1 - \frac{f(t)/f(0)}{f(t-1)/f(0)} = \boxed{1 - \frac{f(t)}{f(t-1)}}$$

3. 裁切并重算实际系数

- $f(t)$ 单调递减，候选方差非负； $f(T) = 0$ ，故 $\beta_T^{\text{候选}} = 1$
- 裁切到 $[10^{-5}, 0.999]$ ，再计算 α_t 并连乘得到实际 $\bar{\alpha}_t$

累计信号目标 $\rightarrow$ 相邻比值 $\rightarrow$ 单步噪声方差 $\rightarrow$ 裁切并重算

来源：Nichol 与 Dhariwal, Improved DDPM, 第 3.2 节

C 余弦加噪方案：裁切与采样步数

$$\beta_t = \text{clip}\left(1 - \frac{f(t)}{f(t-1)}, 10^{-5}, 0.999\right) \quad \alpha_t = 1 - \beta_t \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

- clip 将数值限制在给定区间，避免 $\beta_T = 1$ 导致反向更新除以 0
- 100 步时， $\bar{\alpha}_{99} \approx 2.43 \times 10^{-4}$ ， $\beta_{100} = 0.999$

$$\bar{\alpha}_{100} = \bar{\alpha}_{99}(1 - \beta_{100}) \approx 2.43 \times 10^{-4} \times 10^{-3} = 2.43 \times 10^{-7}$$

- 直接训练短链：重新设计加噪方案并检查终点信号
- 对已训练模型跳步采样：需要适配的采样公式，例如 DDIM
- 去掉随机项、随意跳过时间步均会改变原采样协议

来源：Nichol 与 Dhariwal, Improved DDPM Song 等, DDIM

C Classifier-Free Guidance: 加强条件方向

$$\hat{\epsilon}_{\text{CFG}} = \hat{\epsilon}_{\emptyset} + s(\hat{\epsilon}_y - \hat{\epsilon}_{\emptyset})$$

- 训练时随机丢弃部分标签，同一网络学习有条件与无条件预测
- 采样时组合两个预测， $s = 1$ 对应普通条件预测
- $s > 1$ 外推条件差异，可能增强符合度，也可能减少多样性
- 需要两次预测，且较大的 s 未必带来更好结果

来源：Ho 与 Salimans, Classifier-Free Diffusion Guidance

C 生成模型的课后拓展

- **自回归模型**：把联合分布分解为顺序条件分布

$$p(\mathbf{x}) = \prod_j p(x_j | x_{<j})$$

- **能量模型**：用能量函数刻画未归一化密度

$$p_{\theta}(\mathbf{x}) = \frac{\exp(-E_{\theta}(\mathbf{x}))}{Z_{\theta}}$$

- **潜空间扩散**：在压缩表示中进行加噪与去噪
- **文本条件**：用文本表示作为去噪条件，扩大控制接口

拓展阅读，不占本讲无条件扩散主线的推导与实验时间

来源：Prince, Understanding Deep Learning, 生成模型各章

来源：Rombach 等, Latent Diffusion Models

附录 D

实验说明与参考

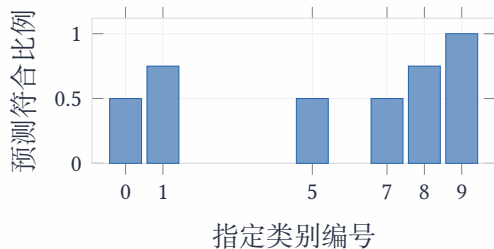
复现配置、评价指标与资料

D 课堂演示的固定配置

项目	本讲协议
数据划分	复用第 4 讲，训练/验证 50000/10000，官方测试 10000
训练子集	从训练划分固定抽取 10000 张，子集种子 602026
扩散主线	无条件，小型 U-Net 类 CNN，100 步余弦加噪方案，30 轮
条件扩展	独立训练 20 轮，丢弃标签概率 0.15，展示 $s = 1$
VAE 桥梁	二维潜变量，隐藏宽度 256，20 轮
GAN 对照	课程简化 DCGAN，随机向量维度 64，30 轮
训练参数	种子 23，batch 128，扩散与 VAE 的 Adam 学习率 10^{-3}
GAN 参数	Adam 学习率 2×10^{-4} ， $\beta_1 = 0.5, \beta_2 = 0.999$

各设置在评价前固定，采用最后一轮 checkpoint，不按生成效果挑选

D 条件符合度与分类器局限



编号 0/1/5/7/8/9 对应 T 恤/长裤/凉鞋/运动鞋/包/踝靴

- 每类固定前 8 张，共 48 张
- 指定类别与 CNN 预测一致的比例
- 样本量很小，仅用于课堂诊断
- 对低质量样本需人工核查

注

符合度不评价类内多样性，也不是严格的人类判断

D 采样计时与复现说明

方法	第 1 次 / 秒	第 2 次 / 秒	第 3 次 / 秒
GAN	0.02	0.011	0.012
DDPM	6.586	6.308	6.364

- 同设备生成 512 张，batch 128，预热后同步计时
- 包含生成样本拷回 CPU，不含加载、训练、分类或绘图
- 精确设备、版本与原始记录保存在协议文件
- 训练与评价脚本支持独立运行，普通课件编译无需训练

课堂材料提供 checkpoint、固定样本数组、配置与来源记录

D FID 的估计协议与本讲评价边界

$$\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{f}_i \quad \Sigma = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{f}_i - \mu)(\mathbf{f}_i - \mu)^\top$$

- 真实与生成特征分别估计统计量；固定网络权重、特征层与真实参考集
- 统一缩放、插值、灰度转 RGB 与数值范围，并报告两组样本数
- 有限样本 FID 存在偏差，重复采样的波动不等于多次训练的波动
- 自然图像 Inception 特征未必适合低分辨率灰度服饰，需检验适用性

本讲约定

- 512 张样本用于已有类别覆盖统计，尚未计算标准 FID
- 若改用服饰 CNN 特征，应注明领域特征距离，不能与标准 FID 直接比较

来源：Parmar 等，CVPR 2022，图像预处理；Chong 与 Forsyth，CVPR 2020，FID / IS 估计偏差

D 其他常见指标：KID 与 Inception Score

定义 | KID ↓

Kernel Inception Distance

$$\text{KID} = \text{MMD}_k^2(P_r^f, P_g^f)$$

- 用三次多项式核比较两组特征
- MMD: Maximum Mean Discrepancy, 最大均值差异
- 可无偏估计, 有限样本仍有波动

P_r^f, P_g^f 为真实与生成特征分布; IS 的 $p(y)$ 为生成样本预测的平均类别分布

定义 | IS ↑

Inception Score

$$\text{IS} = \exp(\mathbb{E}_g D_{\text{KL}}(p(y | z_0) \| p(y)))$$

- 单张图像: 类别预测应集中
- 整组图像: 预测类别应多样
- 不比较真实图像, 可能遗漏类内重复

来源: Bińkowski 等, ICLR 2018, Demystifying MMD GANs, 提出 KID; Salimans 等, NeurIPS 2016, Improved Techniques for Training GANs, 提出 IS

D 参考资料：生成模型与扩散机制

核心论文

[1] Kingma 与 Welling, VAE
编码分布、重参数化与变分下界

[2] Goodfellow 等, GAN
对抗目标与交替训练

[3] Radford 等, DCGAN
卷积生成器与判别器设计

[4] Ho 等, DDPM
前向加噪、噪声目标与反向采样

[5] Nichol 与 Dhariwal, Improved DDPM
余弦加噪方案与采样效率

[6] Ho 与 Salimans, CFG
条件与无条件预测的组合

D 参考资料：教材、数据与实现

课程与实验参考

[7] Understanding Deep Learning
主教材的 VAE、GAN 与扩散模型
章节

[8] Hugging Face Diffusion Course
逐步加噪与从零构建去噪网络

[9] Fashion-MNIST 官方仓库
数据文件、类别编号与许可

[10] PyTorch MSELoss
噪声预测损失的平均口径

[11] PyTorch ConvTranspose2d
可学习上采样与输出尺寸

[12] 2026 课程资料
教师教学主页与课程材料入口