



# Diffusion Models and Representation Learning: A Survey

扩散模型与表示学习

2024 PAMI

汇报人：庞媛媛

2024/07/21

# 目 录

---



背景



扩散模型for  
表示学习



表示学习for  
扩散模型



挑战和未来方向

# 背景


- **扩散模型**：生成模型，自监督学习方法，通过逐步添加噪声并再逐步去噪的方式来学习数据分布。
- **适用**：图像数据、序列数据

Computer Science > Machine Learning

[Submitted on 1 May 2023]

**Diffusion Models for Time Series Applications: A Survey**

- **挑战**：最先进生成方法依赖于带注释的数据→与表示学习结合无需注释的引导方法

- **扩散模型与表示学习** 
  - 扩散模型进行表示学习
  - 表示学习改进扩散模型

- 利用预训练的扩散模型学习表示来完成下游识别任务
- 利用表示学习的进步来改进扩散模型本身

# 数学基础

## ■ 扩散过程（加噪声）

□  $\mathbf{x}_0 \sim p(\mathbf{x}) \rightarrow \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$

□  $p(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \forall t \in \{1, \dots, T\}$   $T$ : 扩散时间步数,  $\beta_t$ : 方差表,  $\mathbf{I}$ : 维数等于 $\mathbf{x}_0$ 的特征矩阵

□ DDPMs参数化:  $p(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0; (1 - \bar{\alpha}_t) \mathbf{I}), \alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^t \alpha_i$

□  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{(1 - \bar{\alpha}_t)} \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \mathbf{I})$

## ■ 生成过程（去噪声）

□  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$

由神经网络在扩散时间步 $t$ 和噪声输入图像 $\mathbf{x}_t$ 的条件下预测

□  $p_\theta(x_{0:T}) = p(x_t) \prod_{t=1}^T p_\theta(x_{t-1} | x_t)$

□  $\mu(\mathbf{x}_t, t) := \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\hat{\mathbf{x}}_0}{1 - \bar{\alpha}_t}$  ( $\hat{\mathbf{x}}_0$ 是去噪网络对原始数据的预测)

□ 协方差 $\Sigma_\theta(\mathbf{x}_t, t)$ 固定, 参数化的反向均值改写为附加噪声 $\epsilon_\theta(\mathbf{x}_t, t)$ 的函数  $\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)$

## ■ 目标函数 假设一个维度与数据相当的纯噪声数据分布 $\mathbf{x}_T \sim \pi(\mathbf{x}_T) = \mathcal{N}(0, \mathbf{I})$ , 从 $p(x_0)$ 生成新样本并去噪

□ 保证 $x_0$ 条件下,  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ 与前向过程真实后验之间的距离最小  $\mathcal{L}_{vlb} = -\log p_\theta(\mathbf{x}_0 | \mathbf{x}_1) + D_{KL}(p(\mathbf{x}_T | \mathbf{x}_0) \parallel \pi(\mathbf{x}_T)) + \sum_{t>1} D_{KL}(p(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t))$

□ 简化后 $\mathcal{L}_{simple} = \mathbb{E}_{t \sim [1, T]} \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0)} \mathbb{E}_{\epsilon_t \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2$

# 数学基础

## ■ 速度预测参数化方法—加快采样（反向扩散）效率

$$\mathbf{v} = \bar{\alpha}_t \epsilon - (1 - \bar{\alpha}_t) \mathbf{x}_t$$

## ■ 用连续而非离散的时间步考虑噪声

扩散过程用Itô随机微分方程 (SDE) 表示:  $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}$

矢量漂移系数 $\mathbf{f}(\cdot, t): \mathbb{R}^d \rightarrow \mathbb{R}^d$ 和标量值扩散系数 $g(\cdot): \mathbb{R} \rightarrow \mathbb{R}$ 在实施扩散模型时需要进行选择,  $\mathbf{w}$  是标准的维纳过程

## ■ 两种SDE及变体

□ 保方差 (Variance-Preserving, VP) SDE:  $\mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\beta(t)\mathbf{x}, g(t) = \sqrt{\beta(t)}$ , 当  $T$  变为无穷大时,  $\beta(t) = \beta_t$ , 等同于DDPM 参数化的连续公式。

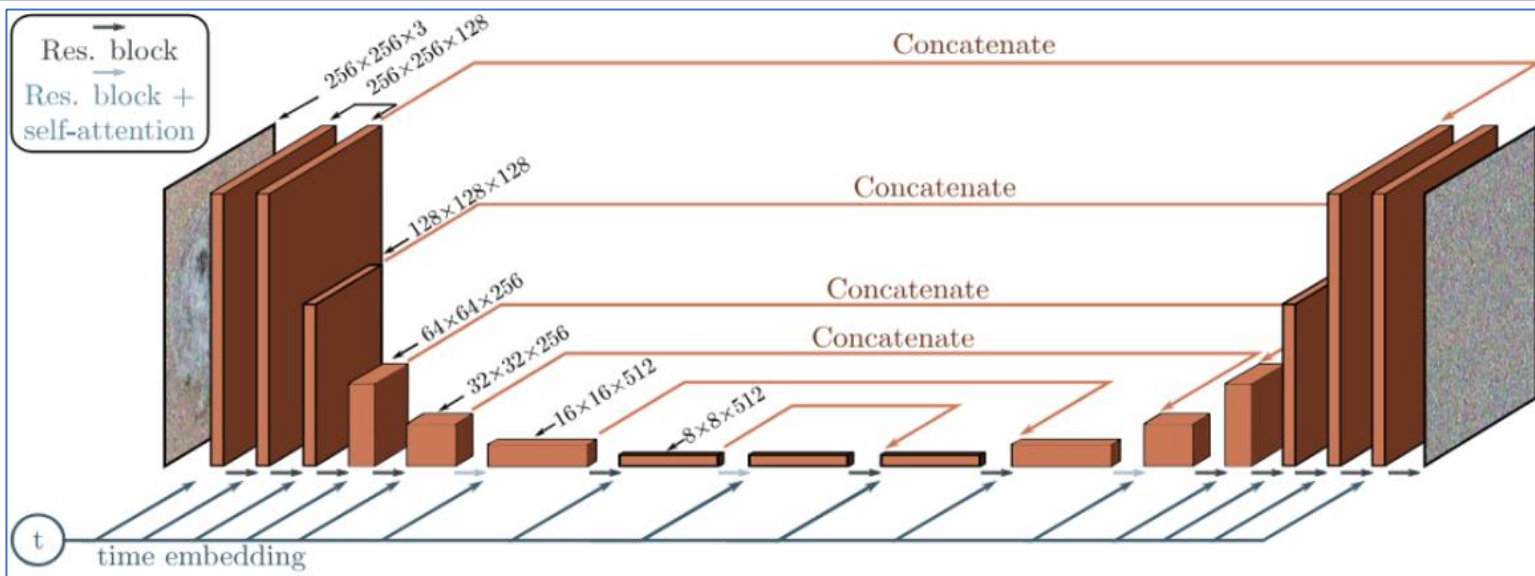
□ 简化后  $\mathcal{L}_{simple} = \mathbb{E}_{t \sim [1, T]} \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0)} \mathbb{E}_{\epsilon_t \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2$

□ 方差解码 (VE):  $\mathbf{f}(\mathbf{x}, t) = 0, g(t) = \sqrt{2\sigma(t) \frac{d\sigma(t)}{dt}}$

□ 能逆转扩散过程的 SDE:  $d\mathbf{x} = -2\sigma(t) \frac{d\sigma(t)}{dt} \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t)) dt + \sqrt{2\sigma(t) \frac{d\sigma(t)}{dt}} d\mathbf{w}$

$\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))$ : 得分函数, 神经网络近似。由于  $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t)) = \frac{D(\mathbf{x}; \sigma) - \mathbf{x}}{\sigma^2}$ , 可以使用最小化 L2 去噪误差的神经网络  $D(\mathbf{x}; \sigma)$  来提取得分函数 “去噪分数匹配”

# 主干架构Backbone



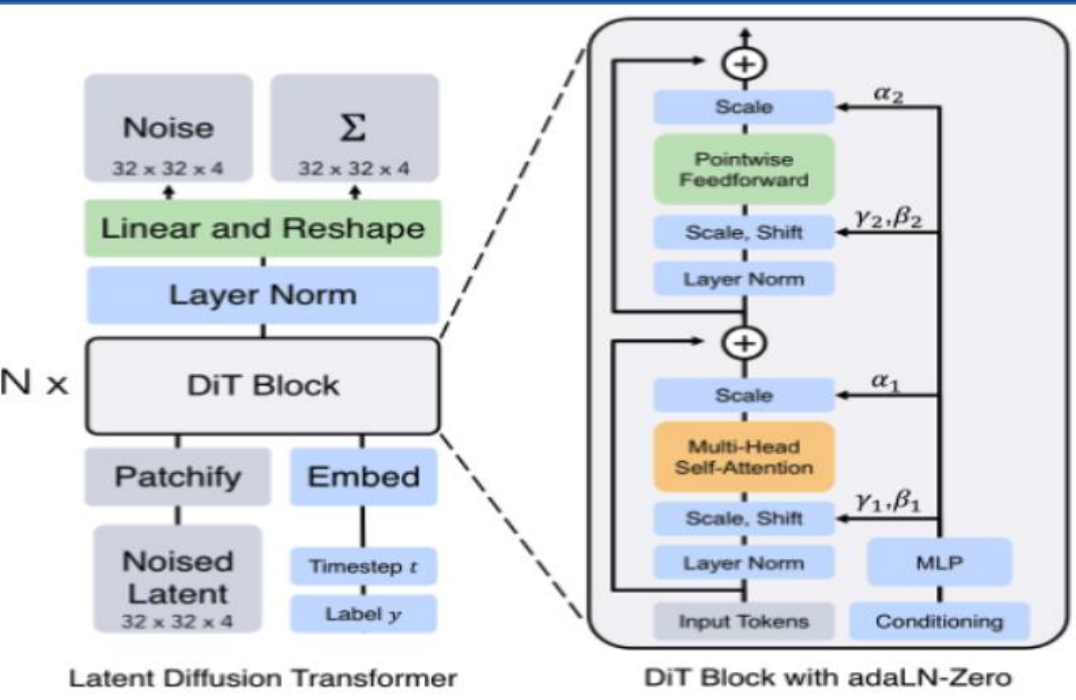
## U-Net网络

基于Wide ResNet，将噪声图像和扩散时间步作为输入，将图像编码为低维表示并输出噪声预测。由编码器和解码器组成，块之间的残差连接可保持梯度流，并有助于恢复压缩表示法中丢失的细粒度细节。编码器由一系列残差块和自注意块组成，将输入图像降采样为低维表示。解码器逐步对低维表示进行上采样，以匹配输入维度。扩散时间步长  $t$  是通过在每个残差块中添加正弦位置嵌入来指定的。

## 潜在扩散模型（LDM）

DDPM 在像素空间中运行，计算成本高。潜在扩散模型（LDM），在预先训练的变分自动编码器的潜在空间中运行。扩散过程应用于生成的表示，而不是直接应用于图像，提高了计算效率，同时保持了生成质量。虽然引入了额外的交叉注意机制以实现更灵活的条件生成，但去噪网络骨干仍然接近 DDPM U-Net 架构。

# 主干架构Backbone

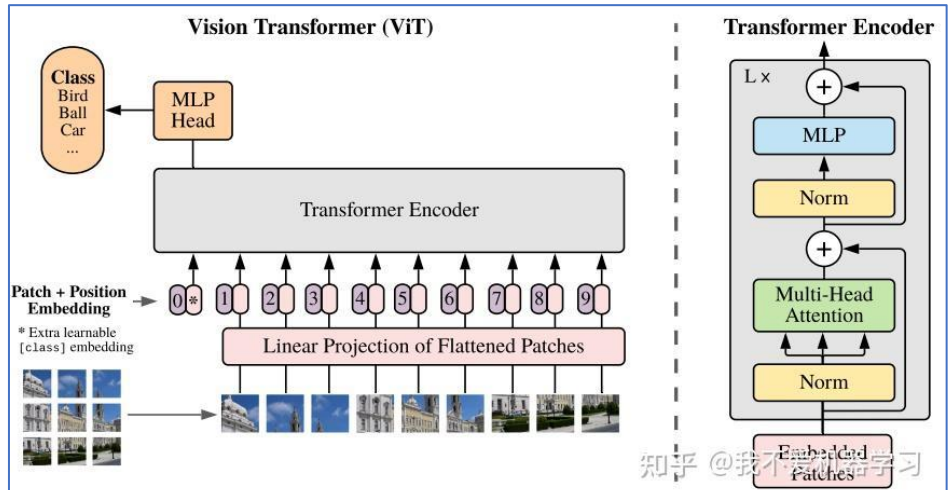


## ■ 基于Transformer的架构

扩散 Transformer (DiT) 是一种基于 ViT 的扩散模型骨干架构，与 LDM 框架相结合，在 ImageNet 上展示了最先进的生成性能。DiT 将输入图像转换为一系列 patches，并通过 "patchify" 层将这些 patches 转换为 tokens。在添加 ViT 风格的位置嵌入后，这些 tokens 送入一系列 Transformer 块，这些块接收扩散时间步  $t$  和调节信号  $c$  等附加条件信息。

## ■ U-ViTs 架构

U-ViTs 将 U-Net 和 ViT 主干网合并为一个统一的主干网。U-ViT 对时间、调节和图像输入进行 tokens 化，但在浅层和深层之间采用长跳接。跳跃连接为低层特征提供了捷径，稳定了去噪网络的训练。研究表明，基于 U-ViT 的骨干网与基于 U-Net CNN 架构的成果不相上下，表明它们有潜力成为其他去噪网络骨干网的可行替代品。







# 扩散模型的引导Guidance

- **引导**：指在生成过程中引入额外信息或条件以引导生成的样本更符合特定目标或约束。
- **实现方式**：
  - 条件扩散模型：条件信息的嵌入（如特定时间、位置或其他上下文信息）
  - 类别指导（分类器）：类别标签嵌入（如交通流量预测中引导模型生成特定时间段或特定区域的交通流量数据）
  - 反馈循环：生成样本与目标样本分布的误差反馈给模型进行调整。
  - 对抗训练：生成对抗样本→对抗训练[结合对抗样本和原始样本，训练生成模型，使模型在面对对抗样本时更加鲁棒]
  - 使用外部模型引导：如预训练的时空图卷积模型生成的特征引导扩散模型的生成过程



# 扩散模型for表示学习的分类

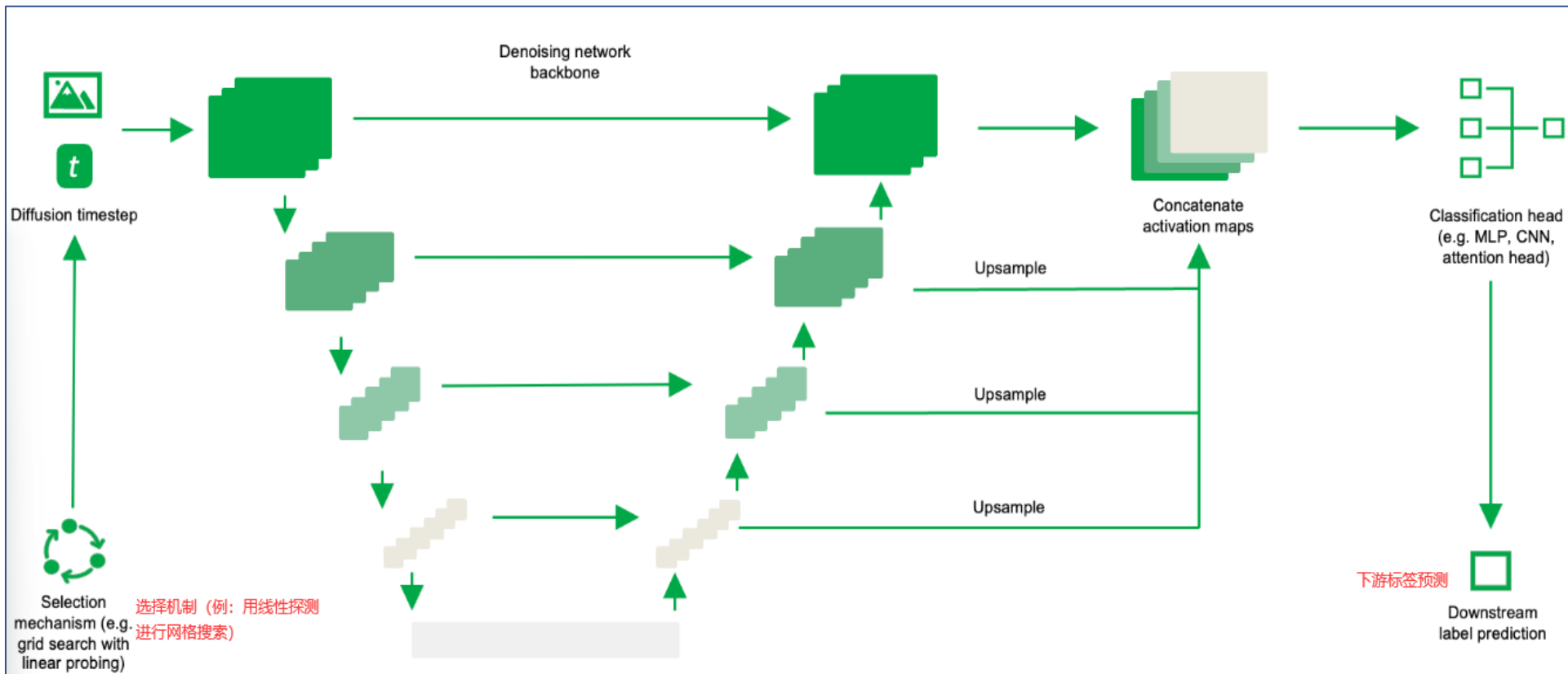
| Paradigm                            | Downstream Task         | Method   |
|-------------------------------------|-------------------------|--|
| Generative Augmentation             | Classification          | Generative Augmentation [10]<br>MA-ZSC [150]   |
|                                     | Semantic Segmentation   | ScribbleGen [148]  |
| Leveraging Intermediate Activations | Classification          | GDC [125]<br>DiffFormer [126]<br>DDAE [169]  |
|                                     | Semantic Segmentation   | DDPM-Seg [15]<br>VDM [187]   |
|                                     | Panoptic Segmentation   | ODISE [170]  |
|                                     | Semantic Correspondence | DIFT [157]<br>SD+DINO [183]<br>Diffusion Hyperfeatures [116]<br>SD4Match [103]<br>USCSD [62] |
|                                     | Depth Estimation        | VDM [187]  |
|                                     | Image Editing           | P2PCAC [65]<br>Plug-and-Play Diffusion Features [160]  |
|                                     |                         |  |
| Diffusion Model Reconstruction      | Classification          | SODA [82]<br>I-DAE [35]<br>DiffMAE [166]   |
|                                     | Semantic Segmentation   | MDM [130]  |
|                                     | Image Editing           | DiffAE [134]<br>PDAE [186]   |
|                                     | Image Interpolation     | InfoDiffusion [165]<br>SmoothDiffusion [58]  |
| Diffusion Model Knowledge Transfer  | Classification          | DiffusionClassifier [95]<br>RepFusion [173]<br>DreamTeacher [96]                             |
| Joint Diffusion Models              | Classification          | JDM [40]<br>HybViT [174]   |
|                                     | Semantic Segmentation   | ADDP [158]   |

# 提取表示的一般框架

- 下游任务中利用中间激活→实现输出理想的扩散时间步输入以及中间层数，其激活表示在上采样和线性探测（评估学习表示的质量）时具有最高的预测性

- 目标：选择时间步  $t \in T$  和一组解码器块编号  $B$ ，使下游任务的预测性能最大化。

$$(t^*, B^*) = \arg \min_{t \in T, B \subseteq B} \mathcal{L}_{\text{discr}}(t, B) \quad (16)$$



网格搜索：通过调整预训练模型的不同参数（如提取特征的层数和噪声水平）进行网格搜索，找到最佳组合。

# 生成增强

- 使用潜在扩散模型生成原始图像的新视图，在采样过程中，ADDP 从纯粹的不可靠标记 $\bar{z}_t$ 开始，通过预测 $\bar{z}_{t-1}$ 对标记序列进行迭代去噪。

$$T_0(\mathbf{x}) = \begin{cases} G(\mathbf{z}; \phi(\mathbf{x})) & \text{if } p \leq p_0 \\ \mathbf{x} & \text{otherwise} \end{cases}$$

$G$  是以噪声向量  $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$  和条件向量  $\phi(\mathbf{x})$  为输入的条件生成模型； $\phi$  是预先训练好的图像编码器， $p \in [0,1]$  是随机数， $p_0$  是超参数，指定应用增强的概率。

- **DiffuMask**：生成增强方法，旨在改进下游语义分割任务。这种方法的理念是利用文本提示和生成图像之间的交叉注意力，将图像合成扩展到语义掩码生成。合成生成的掩码用于数据增强，以提高下游分割性能。使用基于 AffinityNet 的自适应阈值机制，对所有层的单个标记注意力进行平均并转换为二进制掩码。此外，噪声学习模块还能修剪低质量的分割掩码，作者还采用了多种提示技术和静态图像转换技术，以进一步增强生成图像和相应分割掩码的多样性。



# 知识转移

- **知识转移**：一个模型在一个任务上学习到的知识应用（表示）到另一个相关任务的过程。
- **实现方式**：
  - 预训练和微调：大型数据集→较小数据集
  - 迁移学习：源任务→目标任务中（任务有相似性）
  - 域适应：源领域学→目标领域（数据分布不同）（对抗训练、自监督学习减少分布差异）
  - 知识蒸馏：大型模型（教师模型）→较小的模型（学生模型）。学生模型可以在保留教师模型大部分性能的同时，显著减少模型的复杂度和计算成本。
  - 特征重用：模型将从一个任务中学到的特征直接应用到另一个任务中。

# 重构扩散模型

- 通过解构去噪扩散模型：表示学习能力主要由去噪过程驱动而非扩散过程驱动
- 第一阶段：重新定向 DDM 以实现自我监督学习。这包括去除类别条件，重建 DiT 基线中使用的 VQGAN 标记器，并移除依赖注释数据的感知损失项和对抗损失项，实质上将 VQGAN 转换为 VAE。
- 第二阶段：进一步简化 VAE 标记器，用不同的自动编码器变体取代。发现使用更简单的自动编码器变体（如 patch-wise PCA）并不会大幅降低性能。潜在空间每个标记的维度对探测准确性的影响远大于所选的自动编码器。
- 第三阶段：将 DDM 转换为预测去噪输入而非添加的噪声，去除输入缩放，并改变扩散模型以直接在像素空间中运行。这一阶段的结果就是潜在去噪自动编码器（1-DAE）。
- 1-DAE 的灵感来自于扩散模型类似于具有不同噪声尺度的分层自动编码器。
- DiffAE：利用扩散模型进行表示学习，将潜在表示分为紧凑的语义表示和随机表示。DiffAE 包括一个语义编码器和一个条件 DDIM。在推理过程中，通过第二个潜在 DDIM 拟合到语义表示中，从 DDIM 和随机表示中采样，实现高效的语义表示和解码。
- PDAE：发现反向扩散过程中，真实后验均值与预测后验均值之间存在差距。提出预训练 DPM 自动编码，通过编码表示预测均值偏移，确保表示包含尽可能多的信息，显示出更高的训练效率和性能。
- 掩码扩散模型（MDM）：专为自我监督的语义分割设计，采用结构相似性指数（SSIM）损失来缩小重建和分割任务之间的差距。MDM 预训练后，能在较小的标注数据集上取得优异的分割结果。
- SODA：自监督扩散模型，通过新颖的视图生成学习目标和瓶颈层改善表示学习。编码器产生压缩潜在表示，用于生成新视图，训练过程中通过随机清零潜在子向量实现无分类器引导的分层泛化。
- SmoothDiffusion：专注于提高扩散模型潜空间平滑度，通过逐步变化正则化方法，加强潜空间的平滑性。平滑的潜像适用于多种图像插值、反演和编辑任务。



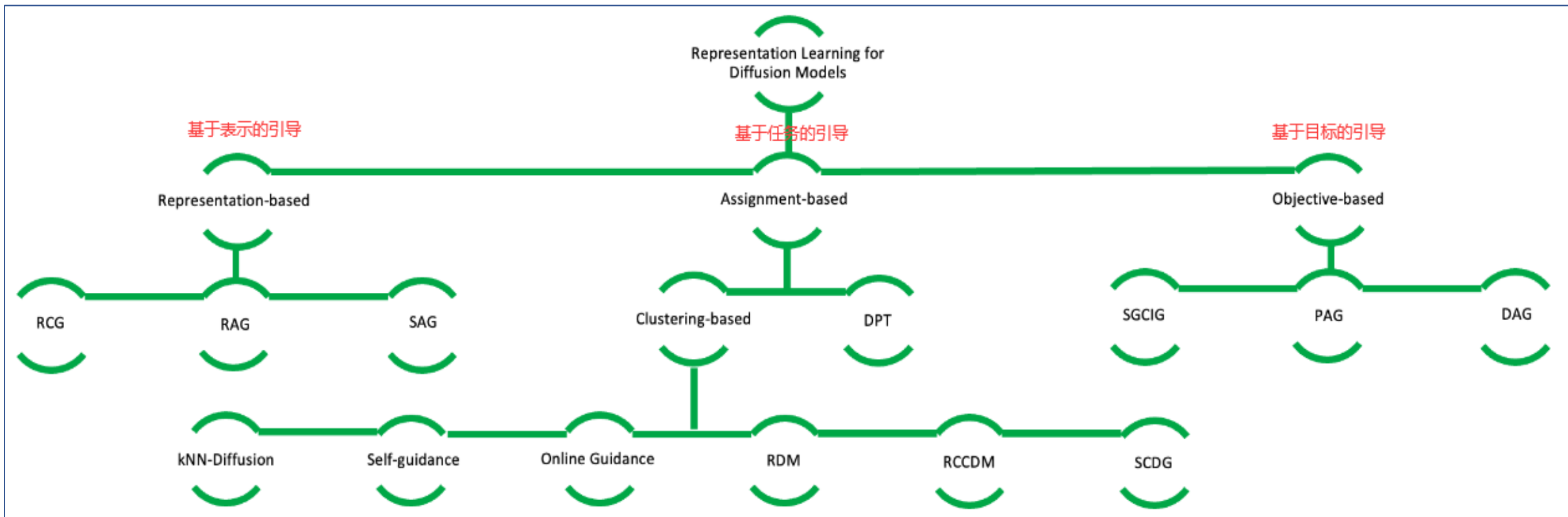
# 联合扩散模型

- 基于扩散的表示学习方法侧重利用扩散模型的潜在变量帮助训练单独的识别网络~混合模型
- 预训练阶段专注合成，后训练阶段/微调阶段专注下游识别，识别网络和扩散去噪网络不共享参数→分离必要吗？
- **HybViT**：扩散模型和视觉Transformer，训练一个混合模型进行图像分类和图像生成。该混合模型使用共享参数进行图像分类和重建，模型的综合损失包括标准交叉熵损失（用于训练分类器）和简单去噪损失（用于训练生成器）。HybViT 使用 ViT 骨干网训练模型，提供了稳定的训练，在生成和判别任务上优于以前的混合模型，但在生成质量上不如纯生成模型。
- **JDM**：JDM 使用 U-Net 骨干网，由编码器、解码器和分类器组成。编码器将输入映射到特征向量，解码器将其重构为去噪样本，分类器预测目标类别。综合训练目标包括交叉熵损失和噪声预测网络的简化目标。JDM 还简化了分类器的引导，通过对噪声图像应用分类器来增强其对噪声的鲁棒性，并通过优化特征向量导向目标标签。JDM 在 CIFAR 和 CelebA 数据集上实现了最佳性能，优于 HybViT。
- **交替去噪扩散过程（ADDP）**：交替对像素和 VQ 标记进行去噪。给定图像  $x_0$ ，预训练的 VQ 编码器将图像映射到 VQ 标记  $z_0$ 。交替扩散过程根据时间步  $t$ ，在采样过程中，ADDP 从纯粹的不可靠标记  $\bar{z}_t$  开始，通过预测  $\bar{z}_{t-1}$  对标记序列进行迭代去噪。



# 用于扩散模型引导的表示学习技术分类

■ 最先进的扩散模型是条件模型，依赖于需要注释数据的引导方法→表示学习为无标记数据分配标签促进引导方法





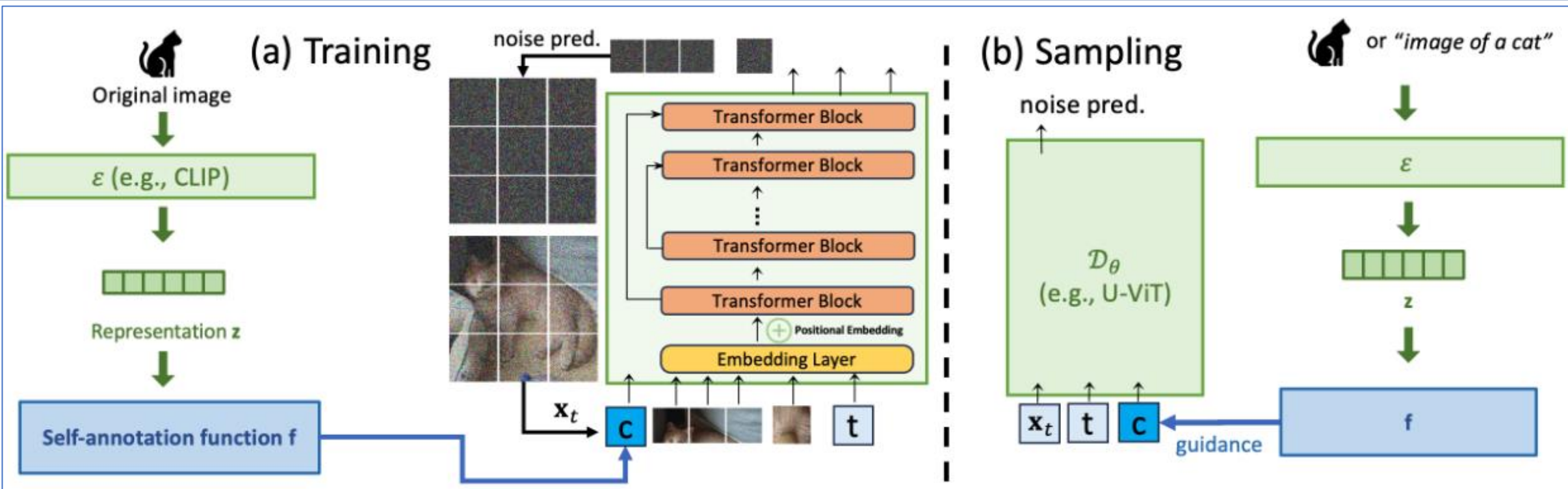
# 基于任务的引导

- **kNN-Diffusion**: 在没有大规模图像文本配对的情况下训练出来高效文本到图像扩散模型。需要一个共享的文本图像编码器，将文本图像对映射到相同的潜在空间中。kNN-Diffusion 利用 kNearest-Neighbors 搜索从检索模型中生成  $k$  个嵌入。检索模型在训练过程中使用输入图像表示法，而文本提示表示法则用于推理。这种方法无需注释数据，但仍需要像 CLIP 这样的预训练编码器，而这又需要一个大规模的文本-图像嵌入数据集进行预训练。
- **检索增强扩散模型 (RDM)**: 它为扩散模型配备了一个图像数据库，用于根据检索到的图像组成新的场景。RDM 包括一个可训练的条件潜在扩散模型  $p_\theta$ 、一个外部图像数据库  $\mathcal{D}$  和一个固定采样策略  $\xi_k$ 。该策略根据查询  $x$  选择  $\mathcal{D}$  的子集  $\mathcal{M}_\mathcal{D}^{(k)}$ 。一种策略是使用距离函数检索  $k$  个最近邻，检索到的数据通过冻结图像编码器  $\phi$  进行处理，并用于调节  $p_\theta$ 。在训练过程中， $\xi_k$  使用 CLIP 图像特征空间中的余弦相似度作为距离函数  $d(x,y)$ ，检索查询图像  $x$  的  $k$  个近邻。这种方法可确保检索到的图像表征对生成任务有用，并可通过 CLIP 共享特征空间进行文本调节。
- **Self-guidance**: 一个包含特征提取函数  $g_\phi$  和自注释函数  $f_\psi$  的框架。特征提取函数是一个自监督特征提取器，它将输入数据  $x \in \mathcal{D}$  映射到特征空间  $\mathcal{H}$ ，其中  $\mathcal{D}$  表示数据集。该特征表示是  $f_\psi$  的输入，它将特征表示  $g_\phi(x; \mathcal{D})$  映射到引导信号  $k$ 。该框架可用于实现自标记引导。虽然这种方法是自我监督的，但它仍然依赖于外部预训练的特征提取器来生成用于聚类的特征表示。

# 基于任务的引导

- **Online-guidance:** 针对self-guidance依赖于外部预训练的特征提取器来生成用于聚类的特征表示而提出在线特征聚类方法，需要在训练过程中从条件扩散模型中获取聚类的条件信号，为解决这个问题，在用于识别聚类信号的条件扩散模型中引入一个向量。对于每个图像示例，以这个零向量为条件的条件扩散模型都要经过一个全连接的特征预测头，用来计算映射到一组可学习原型（用  $\mathbf{M}$  表示）的特征。这种方法结合使用扩散训练损失和 Sinkhorn-Knopp 损失，利用  $\mathbf{M}$  实现基于聚类特征的引导信号  $\mathbf{c}$ 。
- **双重伪训练（DPT）:** 使用在有限标记数据上训练的分类器生成伪标签。然后，利用这些伪标签对扩散模型进行调节，生成伪图像，再利用这些伪图像作为数据增强，对混合了伪图像和真实图像的分类器进行再训练。DPT 包括三个阶段。首先，在部分标注数据上训练半监督分类器，预测所有图像  $x \in \mathcal{X}$  的伪标签  $\hat{y}$ 。其次，在带有伪标签的数据集  $S_1 = \{(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in \mathcal{X}\}$  上训练条件生成模型。最后，在由生成数据扩充的真实数据上重新训练分类器。DPT 在 ImageNet 分类和生成上取得了极具竞争力的性能，每类只需 5 个标签，优于 ADM [42] 和 LDM [138] 等多个有监督扩散模型基准。

# 基于任务引导的扩散模型-通用框架



- 自监督图像编码器：输入映射到低维特征表示 $z$ ，CLIP多模态特征提取器可以针对文本和图像。
- 自标注函数：图像表示为输入生成标注 $c$ （引导信息）
- 去噪网络：将噪声图像 $x_t$ 、扩散时间步 $t$ 和引导信号 $c$ 作为输入，并对图像进行去噪。

# 基于表示、目标的引导

- 表示条件图像生成RCG：利用预训练编码器从图像分布映射出的自监督表示分布来调节扩散模型，核心思路是在预训练编码器生成的表示上训练表示扩散模型（RDM），以生成低维图像表示。然后，再训练以表示为条件的像素生成器，将噪声分布映射到图像分布。
- 读出引导RG：利用在冻结扩散模型基础上训练的辅助读出头，提取生成图像中可用于引导的属性。这些属性可包括人体姿态、深度图、边缘，甚至更高阶的属性，如与另一幅图像的相似性。
- 去噪网络：通过利用中间自我注意力激活图，对包含突出信息的区域进行逆向模糊，并使用残余信息作为引导。这就提高了生成质量，而不需要外部信息或额外的训练。U-Net 和 DiT 扩散骨干网中都包含的自我注意机制允许噪声预测器注意输入中信息量最大的特征。
- 自我引导的可控图像生成（SGCIG）：一种零样本方法，旨在增强用户对文本到图像扩散模型生成的图像中对象的结构和语义元素的控制。工作原理：在去噪网络的目标中添加一系列引导项，定义一系列可用于执行图像处理的属性，通过引导属性的变化进行图像编辑。
- 深度感知引导（DAG）：利用中间去噪网络层的语义信息改进深度感知图像合成。Kim 等人提出使用有限的深度标注数据训练深度预测器，并在生成过程中引入深度一致性引导和深度先验引导两种策略。工作原理：结合使用弱深度预测器和强深度预测器之间的一致性损失，以及在深度域上采用小分辨率扩散 U-Net，增强生成图像的深度语义。
- 扰动注意力引导（PAG）：一种采样引导方法，可提高有条件与无条件设置下的生成质量。PAG 不需要额外的训练或外部预训练模型。工作原理：引入一个隐式判别器  $D$ ，在扩散过程中区分理想样本和不理想样本，通过扰动预先训练好的去噪网络的前向传递估计得分，引导采样过程远离劣化样本。

# 总体挑战

基于扩散模型的表示学习是一个新颖的研究领域，具有很大的理论和实践改进潜力。提高表示学习和生成模型之间的协同作用类似于鸡生蛋、蛋生鸡的问题，即**更好的扩散模型可同时带来更高质量的图像表示，而更好的表示学习方法在应用于自我监督指导方法时可提高扩散模型的生成质量**。在训练过程中为扩散模型提供指导的改进型在线引导方法在此方面大有裨益。

**为了节省扩散模型的计算量**，采样过程已被大大缩减为几步，甚至是一步。然而，在**采样步骤较少的情况下保持表示学习的潜力是一个挑战**。



# 潜在方向

## 扩散模型生成表示-评估的3种方法:

辅助模型评估（辅助模型使用表示作输入，通过辅助任务的性能评估表示的好坏）

- 分类任务：分类器准确率、精确率、召回率等
- 回归任务：回归误差（MSE、MAE）
- 聚类任务：聚类效果（簇内距/簇间距）

可解释性评估（表示是否可以被人类理解和解释）

- 特征重要性分析：使用SHAP值/LIME方法量化每个特征对辅助模型输出的贡献，帮助理解哪些特征在决策过程中最为重要。
- 可视化技术：如热力图或激活图，显示输入特征对预测结果的重要性。

解缠评估（潜在表示的独立性和可分离性）

- 独立性度量：使用互信息或总相关度等度量潜在变量之间的依赖性。较低的互信息或总相关度表示更好的解缠表示。
- 定性评估：人为操控潜在变量，观察生成数据的变化。
- 定量评估：使用如 $\beta$ -VAE（Beta Variational Autoencoder）、FactorVAE等模型，通过调整超参数评估解缠表示的效果。这些模型在损失函数中加入特定的正则项，以鼓励潜在表示的解缠。

表示质量

可解释图：  
特征影像图  
特征依赖图  
特征重要性SHAP

## 潜在研究方向:

- 扩散模型的分离和可解释表示学习
- 基于Transformer的主干网无法进行并行推理以及注意力机制的二次方复杂性限制在高分辨率图像和长视频的应用
- 扩散模型和流匹配模型之间关系，扩散表示学习框架应用于流匹配模型

流匹配模型（Flow Matching Models）是近年来在生成建模领域中提出的一种新型方法，用于高效训练和推断复杂数据分布。流匹配模型将生成过程视为一个连续的、可逆的变换序列，通过流（flow）的匹配优化目标，使模型学习到从简单分布到复杂数据分布的映射。

## 交通预测模型融入扩散模型思路

■ **数据增强**：扩散模型生成额外的船舶交通数据弥补数据稀疏或不平衡，生成的合成数据用来扩展训练数据集，增强训练效果。

■ **对抗训练**：利用扩散模型生成对抗样本，提高时空卷积模型的鲁棒性和性能。

生成对抗样本（与原始数据相似但存在微小扰动）旨在欺骗预测模型→合并对抗样本与原始样本形成增强数据集→训练预测模型

■ **知识转移**：将扩散模型学习到的表示用于基于时空卷积模型的交通预测

➤ 预训练扩散模型：训练一个扩散模型生成交通流量数据的高级表示。

➤ 数据准备→模型训练（CNN作为编码器和解码器训练扩散模型最小化重构误差）→高层表示提取（使用编码器部分提取交通流量数据的高级表示）

➤ 提取预训练模型的表示：表示提取→特征准备（高层→时空图卷积模型输入格式）

➤ 基于时空图卷积的交通预测模型

➤ 整合扩散模型和时空图卷积模型“知识转移”

➤ 模型整合（扩散模型的编码器部分和时空图卷积模型结合，编码器的输出作为时空图卷积模型输入特征的来源）→联合训练（固定扩散模型的参数，只训练时空图卷部分或进行微调使得整个模型协同优化）→评估与优化



谢 谢！