用于图像数据增强的扩散模型的进展：方法、模型、评估指标和未来研究方向综述

Panagiotis Alimisis[1] , Ioannis Mademlis[1] 、
Panagiotis Radoglou-Grammatikis[2,3] , Panagiotis Sarigiannidis[2] ,
Georgios Th.Papadopoulos[1*]

[1]雅典哈罗科皮奥大学信息学与远程信息处理系，地址：Thiseos 70, Athens, GR 17676, Attiki, Greece。

[2]英国伦敦大学电气与计算机工程系
西马其顿，活跃城市规划区，科扎尼，GR 50150，希腊科扎尼。

[3]K3Y, Studentski district, Vitosha quarter, Bl.9, Sofia, BG 1700, Sofia
保加利亚城市省。

*通讯作者。电子邮件：g.th.papadopoulos@hua.gr；投稿作者：csi23301@hua.gr; imademlis@hua.gr; pradoglou@uowm.gr, pradoglou@k3y.bg; psarigiannidis@uowm.gr；

摘要

图像数据扩增是现代计算机视觉任务中的一种重要方法，因为它有助于提高训练数据集的多样性和质量，从而提高机器学习模型在下游任务中的性能和鲁棒性。与此同时，增强方法还可用于以感知上下文和语义的方式编辑/修改给定图像。扩散模型（Diffusion Models，DMs）是生成式人工智能（Artificial Intel-ligence，AI）领域最新且极具前景的一类方法，已成为图像数据增强的有力工具，能够通过学习底层数据分布生成逼真且多样化的

图像。当前的研究对基于 DM 的图像增强方法进行了系统、全面和深入的综述，涵盖了广泛的策略、任务和应用。特别是

首先对 DM 的基本原理、模型架构和训练策略进行了分析。随后，介绍了相关图像增强方法的分类，重点是语义处理、个性化和适应性以及特定应用增强任务方面的技术。然后，分析了性能评估方法和各自的评价指标。最后，讨论了该领域当前面临的挑战和未来的研究方向。

关键词图像数据增强、扩散模型、生成式人工智能、评估指标

1 导言

所谓的深度学习（Deep Learning，DL）范式是现代计算机视觉的主流，它依赖于大规模深度神经网络（Deep Neural Networks，DNNs）的使用。迄今为止，DNNs 在一系列广泛的视觉理解任务中表现出了卓越的性能。然而，这种出色的视觉解释和推理能力的同时，也越来越需要更大和足够多样化的训练数据集。另一方面，随着图像分析任务变得越来越复杂和苛刻，训练数据的数量、多样性和潜在偏差等方面的限制阻碍了 DNN 的稳健泛化能力 Mumuni 和 Mumuni（2022 年）。因此，数据要求已成为一个相当突出的话题，因为足够数量的训练样本对于充分发挥 DNN 的能力至关重要 Zhang 等人（2022 年）。相反，现实世界的图像数据集，尤其是针对特定应用领域的图像数据集，往往在这些方面存在缺陷，甚至包含完全相关的训练图像，而这些图像被证明本质上是冗余的。

图像增强是缓解因数据集限制而产生的问题的一种常见而便捷的方法，即自动创建每个训练图像的额外变体，并利用它们来增强训练集 Xu 等人（2023a）。通常情况下，生成的变体在外观上表现出差异，但保留的语义内容与原始图像相同。用这种合成图像扩展训练数据集，可以增加数据集的多样性，并在许多情况下提高 DNN 的学习和识别性能 Zhou 等人（2023b）。这种行为源于图像增强在训练 DNN 时实质上起到了额外正则化机制的作用，因此有助于防止过拟合 Perez 和 Wang (2017)；Shorten 和 Khoshgoftaar (2019)。

传统的图像增强方法，如几何变换（如图像旋转、翻转、裁剪、缩放、水平/垂直平移、挤压等）和色彩空间调整或光度变换（如模糊、锐化、抖动等），仍然

非常常见 Xu et al (2023b)； Shorten and Khoshgoftaar (2019)； Yang et al (2022)。这种类型的多种变换可以组合在一起，这样就能从原始数据集生成更多的增强图像。这些方法利用领域知识生成的合成示例类似于

的原始图像。最近提出的图像增强方法与此大体一致，是一套系统地破坏原始图像以生成增强变体的策略。这类方法主要包括：a) "mixup" Zhang 等人（2017 年），使用一对训练图像及其标签的凸组合；b) "cutout" DeVries 和 Taylor（2017 年），随机屏蔽输入图像的方形区域；c) "cutmix" Yun 等人（2019 年），通过屏蔽第一张图像与第二张图像的一个区域（反之亦然），随机组合两张训练图像；以及d) "mixup "DeVries 和 Taylor（2017 年）。
d) Kang 等人（2017 年）的 "patchshuffle"，它使用内核滤波器随机交换滑动窗口中的像素值。

由于当代图像分析需求的复杂性和多变性，上述相对简单和直接的增强方法的有效性正受到越来越多的挑战。虽然这些策略可以有效增加简单任务的数据多样性，但它们大多无法捕捉高维图像数据中的潜在结构和复杂关系。此外，它们中的许多都需要特定领域的知识和特定数据集的校准，才能正确应用。此外，DNNs 对大型训练数据集和有效正则化的需求与日俱增，使得图像增强成为现代机器学习的一个极其重要的组成部分。

与处理现有图像以生成变体的传统方法不同，扩散模型（DMs）可以通过合成新的、外观逼真且可信的图像，轻松用于图像增强实践。DMs 是一类复杂的生成式 DNN，擅长隐式建模底层数据生成分布和复杂图像的结构。这种能力使它们能够从训练数据集的分布中采样假冒的新图像，这些图像同时具有多样性、高度现实性和对未见数据场景的代表性，因为它们包含了微妙的细节并保留了原始数据集的固有结构 Zhang et al (2022)；Trabucco et al (2023)。因此，可以直接利用它们对后者进行有意义的增强。

DMs 的学习范式依赖于在训练图像中反复应用噪声，然后学习逆转这一过程，与其他竞争生成模型（如生成对抗网络）相比，DMs 在图像增强方面的前景十分广阔。Rombach 等人（2022 年）还发现，DMs 的最新进展可以通过类标签、文本描述或输入图像对图像合成过程进行调节。这种用户控制水平允许进行有针对性的图像增强，根据手头的任务生成满足特定要求的图像。

最近，通过 DM 和多模式策略（如文本条件图像创建）进行生成式图像合成的进展，得到了在大规模数据集上使用大规模预训练的补充，这与基础模型（

FM）的发展趋势如出一辙：Rombach 等人（2022 年）；Podell 等人（2023 年）；Esser 等人（2024 年）；Saharia 等人（2022 年b）。通过这种方法，可以获得经过预训练的 DM，这些 DM 可以生成具有自然外观变化的图像（例如，改变卡车上的涂鸦设计，如图 1 所示），因此可以直接加以利用

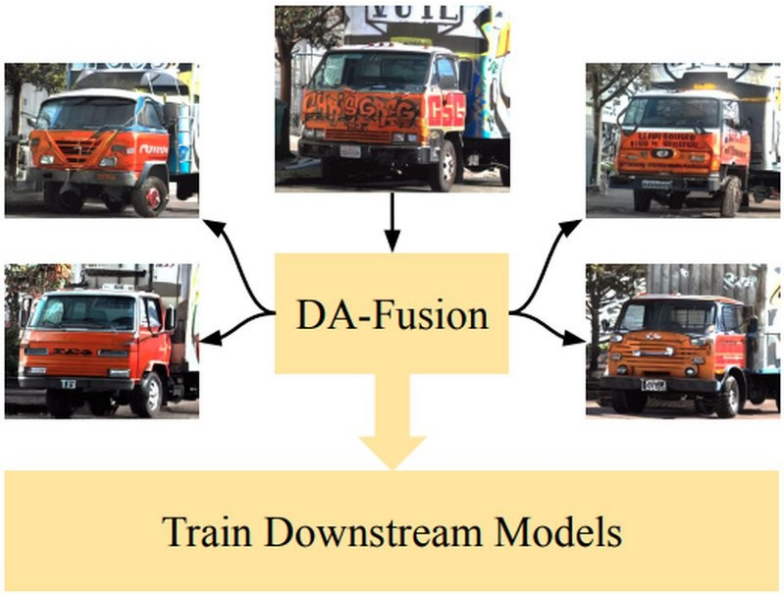

图 1 利用扩散模型改变卡车涂鸦设计的语义。图片来自 Trabucco 等人（2023 年）。

无需大量人力即可进行复杂的图像增强 Dunlap 等人（2023 年）。

尽管最近在应用 DMs 进行高级图像增强方面取得了进展和成就，但相关的文献调查却寥寥无几。现有的调查要么侧重于传统的图像增强，Yang 等人（2022 年），要么只是提供了 DMs 的总体概述，而没有深入研究 DMs 在图像增强中的具体应用 Cao 等人（2024 年）；Yang 等人（2023b）；Song 等人（2024 年）。例如，Mumuni 和 Mumuni（2022 年）的研究全面介绍了图像增强方法，包括输入空间转换、特征空间增强、数据合成和基于元学习的方法。不过，该研究并未涉及使用 DM 的最新进展。相比之下，Croitoru 等人（2023 年）的研究提出了三种通用的扩散建模框架，分别基于去噪扩散概率模型、噪声条件得分网络和随机微分方程，但没有探讨它们在图像增强方面的潜力。一些研究涉及 DMs 的效率方面，Ulhaq 等人（2022 年）或其在特定领域（如医学影像）的应用，Kazerouni 等人（2023 年）；Kebaili 等人（2023 年）。然而，这些著作并未全面概述用于各种计算机视觉任务的增强型 DM。此外，最近的一项调查对基于大型学习模型的增强方法进行了分类，其中包括基于 DMs 的方法，但并没有特别关注这些方法 Zhou 等人（2024 年）。

本文其余部分的结构如下。第 2 节概述了 DM 的基础和基本原理。第 3 节介绍了不同类别的以 DM 为动力的图像增强方法，而第 4 节则介绍了以 DM 为动力的图像增强方法。

第 4 节深入解释了这些类别。第 5 节介绍了用于评估图像增强 DM 性能的各种评价指标。第 6 节讨论了当前使用 DM 进行图像增强所面临的挑战和限制。最后，第 7 节从前面的讨论中总结了一些见解，并对未来的研究方向提出了建议。

2 扩散模型的基础

扩散模型（DM）是一类功能强大的生成模型，在图像合成中获得了显著的应用。受非平衡态热力学的启发，Sohl-Dickstein 等人（2015 年）通过添加高斯噪声的迭代过程（正向扩散），逐步破坏数据的结构，从而将数据分布逐步转化为纯随机噪声分布。然后，通过一个可学习的反向扩散过程来恢复数据的结构，从而产生一个可行的生成模型。因此，可以训练 DM，将随机噪音模式逐步转化为数据生成分布的样本。本节将详细介绍 DM 的基本原理，阐明 DM 为何特别适用于视觉数据增强。

2.1 前向扩散过程

前向扩散（FD）过程是 DM 的基石，因为它通过连续插入高斯噪声来破坏训练数据集。假设初始数据分布为 $q(x_0)$，其中下标 "0 "表示数据集的原始/未修改状态，而 x_0 $q(x_0)$ 是该数据集中的一幅图像。FD 以增量噪声版本 x_1, x_2 ... ... 的序列 q 的形式进行。, x_T，这些版本由一个马尔可夫链。该序列中每一步的条件 d√ 分布$p(x_t|x_{t-1})$、建模为高斯 N $(x_t; 1 - \beta x_{tt-1}, \beta_t \mathrm{I})$，其中 t 范围为 1 到 T，它表示添加到输入图像中的总噪声。T 相当于扩散步骤的总数，β_1, ...β_T 是一串方差参数，定义了每一步的噪音水平；I 是与输入 x 维度相匹配的身份矩阵$_0$；$\mathbf{N}(x; \mu, \sigma)$ 表示均值为 μ、协方差为 σ 的正态分布。

FD 的一个关键特性是，利用重新参数化技巧，可以在任意时间步长 t 以封闭形式对 x_t 进行采样：

$$\text{让 } \alpha_t = 1 - \beta_t, \ \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$$

$$\text{那么 } q(x_t|x_0) = N\left(x; \sqrt{\bar{\alpha}_t}\, x_{t0}, (1 - \bar{\alpha}_t)\mathbf{I}\right)$$

$$x_t = \sqrt{\bar{a}_t}\, x_{t0} + (1 - \bar{a}_t)\, \epsilon, \quad (1)$$

其中整数 $t \in [1, T]$ 和 $\epsilon \sim \mathcal{N}(0, I)$。因此，噪声版本 x_t 可以通过序列 a_i 确定的累积方差调整 β_t 直接获得，其中 $a_t = 1 - \beta_t$。这样，我们就可以从原始版本中计算出任何噪声版本 x_t。

图像 x_0 ，而无需反复生成所有时间步骤的噪声版本。

2.2 反向扩散过程

根据 FD 引入的破坏，迭代反向扩散（RD）过程旨在从噪声版本中恢复原始数据集图像。去噪学习模型（可以是 DNN）不是直接从噪声模式生成图像，而是从最终的 FD 输出开始，迭代预测在 FD 过程的每个步骤中添加到数据中的噪声模式，以便将其去除。渐进式去噪会在连续的 T 个步骤中逐渐完善图像。这就是所谓的去噪扩散概率模型（DDPM）。或者，模型可以学习所谓的 "得分函数"，即数据相对于输入的对数概率密度函数的梯度。然后，模型在每个时间步的预测结果就可以用来按照梯度从分布中迭代采样。这种 DM 变体被称为基于分数的生成模型（SGM），Song et al (2020b)。

采用的预测 DNN 通常是 U-Net CNN Ronneberger 等人（2015 年）。

关于数学公式，研究与发展过程的定义如下：

$$p_\vartheta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\vartheta(x_{t-1} | x_t), \tag{2}$$

其中 $p_\vartheta(x_{t-1} | x_t) = N(x_{t-1}, \mu_\vartheta(x_t, t), \Sigma(x_t, t))$.

在 Ho 等人（2020 年）的研究中，U-Net 是通过以下损失函数进行训练的：

$$L_{simple} = E_{t,x,\epsilon}[||\epsilon - \epsilon_\vartheta(x_t, t)||^2]。 \tag{3}$$

其中，ϵ 表示为获得噪声版本 x_t 而添加到图像 x_0 中的高斯噪声，$\epsilon_\vartheta(x_t, t)$ 表示在给定噪声图像 x_t 和时间步长 t 的情况下，由以 ϑ 为参数的 DNN 预测的噪声。在 SGM 方法中，$\epsilon_\vartheta(x_t, t)$ 是预测得分，因此在训练后，$\mu_\vartheta(x_t, t)$ 可以用 $\epsilon_\vartheta(x_t, t)$ 的函数来近似。尽管 L_{simple} (3) 并未提供学习 $\Sigma_\vartheta(x_t, t)$ 的方法，但 Ho 等人 (2020) 的研究表明将方差定为 σI，可以获得最佳结果，而不是 $\Sigma(x, t)$。[2]

而不是学习。

RD 是一个连续 T 个时间步骤的迭代过程，从噪声模式 x_T 开始，逐步恢复原始图像 x_0。在每个时间步骤 t，计算 μ 和 Σ 并生成一个新版本的输出图像，随后作为下一个时间步骤 t 的输入。需要在 T 个连续迭代中按顺序生成图像是 DM 的

一大局限。一种简单的改进方法是减少采样步数，从 T 步减少到 K 步，取 1 到 T 之间间隔均匀的实数 Nichol 和 Dhariwal (2021)。另外，Song 等人（2020a）的非马尔可夫去噪扩散隐含模型（DDIMs）在生成过程中只对 S 个扩散步骤$[t_1, \ldots, t_S] \subseteq [1, T]$进行采样：

$$x = \sqrt{a} - x_t - \sqrt{1 - a_t \epsilon_\vartheta(x_t)} + \sqrt{1 - a} - \sigma^2 \epsilon(x) + \sigma \epsilon \circ \quad (4)$$

t-1 t-1 t-1 $t\vartheta$ t $t\,t$

2.3 指导

分类器指导 Dhariwal 和 Nichol（2021 年）利用预训练的封闭集分类器，将预训练的无条件 DM 的 RD 过程条件化为所需的类标签。分类器模型 $p_\phi(y\,x_t)$（其中，j 表示其参数）支持与潜在条件类别一样多的不同类别标签 y。采用这种方法并考虑到 SGM 计算公式，RD 过程在每个时间步长都会根据引导采样的对数概率 $\log p_\phi(y\,x_t)$ 的梯度进行调整。因此，$\mu_\vartheta(x_t, t)$ 的近似值为：

$$\vartheta\epsilon(x_t, t) + s * \nabla_{x_t} \log p_\phi(y|x_t)、 \quad (5)$$

其中，s 是控制引导强度的比例因子。这种方法可确保生成的样本符合目标类别分布，而无需重新训练无条件训练的 DNN。

无分类器指导 Ho 和 Salimans（2022 年）在对 DM 进行训练时，直接以类标签为条件，从而省去了明确的单独分类器模型。在训练 DM 时，会同时使用有条件的 ϑ_c 和无条件的 ϑ_u 目标，交替使用标签条件和不使用标签条件。在推理时，通过对条件分数和非条件分数进行内插，从而实现引导，因此 $\mu_\vartheta(x_t, t)$ 的近似值如下：

$$\nabla_{x_t} \log p_\vartheta(x_t|y) = \nabla_{x_t} \log p_{\vartheta c}(x_t|y) + w(\nabla \; log \; p \; (x_{x_t\,\vartheta ct}|y) - \nabla_{x_t} \; log \; p_{\vartheta u}(x_t)),$$

(6) 其中 w 是控制引导强度的权重参数。

2.4 潜空间扩散模型

尽管 DDIMs 的 RD 过程更快，但在像素空间和任意分辨率下生成图像仍然是一个重大瓶颈。为此，Rombach 等人（2022 年）提出了在潜空间运行的潜扩散模型（LDM），以显著加快生成过程。具体而言，LDM 依赖于在大规模数据集上预先训练的外部自动编码器。它的编码器可以学习将图像 x D_x 映射到一个特殊的潜码 $z =$ $D\;E\;(x)$。Van Den Oord 等人（

2017 年）；Agustsson 等人（2017 年）。其解码器　　学会将这种低维潜在表征映射回像素空间，因此　　((*x*))　　*x*.因此，常规 DM 或 DDIM 经过训练后可在潜空间内生成代码。生成的代码可以将通过预训练的图像映射回真实的高维图像。　　*D*.

LDM 可以以类标签、分割掩码甚至文本为条件，从而指导生成过程。假设 c_ϑ *(y)* 是一个模型，它映射了原始条件

输入 y 的条件向量[1].因此，LDM 损失的计算公式为

$$L_{LDM} = \mathbb{E}_{z\in E(x),y,\epsilon\in N(0,1),t}\left[\left|\left|\epsilon - \epsilon_\vartheta\left(z_t, t, c_\vartheta\left(y\right)\right)\right|\right|^2\right], \quad (7)$$

其中，t 是时间步长，z_t 是在步长 t 处去噪的潜在表示，ϵ 是未缩放的噪声样本，ϵ_ϑ 是去噪网络的预测值。直观地说，目标是正确去除添加到图像潜在表示中的噪声。在训练过程中，c_ϑ 和 ϵ_ϑ 将共同优化，以最小化 LDM 损失。在推理时，对随机噪声张量进行采样和迭代去噪，生成新的潜在图像 z_0 。

2.5 用于图像合成的基础扩散模型

条件式 LDM 推动了 "稳定扩散"（SD）基础模型的发展，该模型是一种文本到图像生成器（T2I），在 LAION-5B 数据集 Schuhmann 等人（2021 年）上进行了预训练。近年来，SD 在图像增强的生成式图像合成相关研究中占据了主导地位。尽管如此，人们仍尝试对其进行进一步改进。例如，"稳定扩散 XL"（SDXL）Podell 等人（2023 年）从三个方面对基本 SD 进行了创新：

- 它拥有比原来复杂三倍的 U-Net，并利用双文本编码系统进行文本调节。这个新的文本编码器（OpenCLIP ViT-bigG/14）与原来的编码器一起运行，极大地扩展了机型的容量。
- 它通过将尺寸和剪裁功能结合在一起，增强了对最终图像剪裁的控制。
  在训练过程中进行作物调节。具体做法是通过傅立叶特征嵌入将作物参数作为调节参数输入模型。
- 其推理阶段分为两个步骤："基础 "模型生成初始图像、
  然后将其输入到 "精炼 "模型中，以添加更精细、更高质量的细节。

另一个先进的变体是 *PixArt-α* Chen 等人（2023c），它在三个方面与 SD 有所区别：

- 训练策略分解：设计了三个不同的训练步骤，分别优化像素依赖性、文本图像对齐和图像美学质量。
- 高效 T2I 变压器：在 Diffu- 变压器中加入了交叉注意模块。
  sion Transformer (DiT) 来注入文本条件，并简化计算密集型类条件分支。
- 信息量大的数据：强调文本中概念密度的重要性

在图像对中，大型视觉语言模型（VLM）被用来自动标注密集的伪标题，帮助文本-图像对齐学习。

最近，Esser 等人（2024 年）提出了 "Stable Diffusion 3"（SD3），它依赖于 Peebles 和 Xie（2023 年）的 Transformer 神经架构，而不是 U-Net 卷积神经网络（CNN）。它为两种相关模态（即文本和图像）使用了一组不同的参数，因此可以

[1]如果是文本提示，可以使用任何文本编码器。

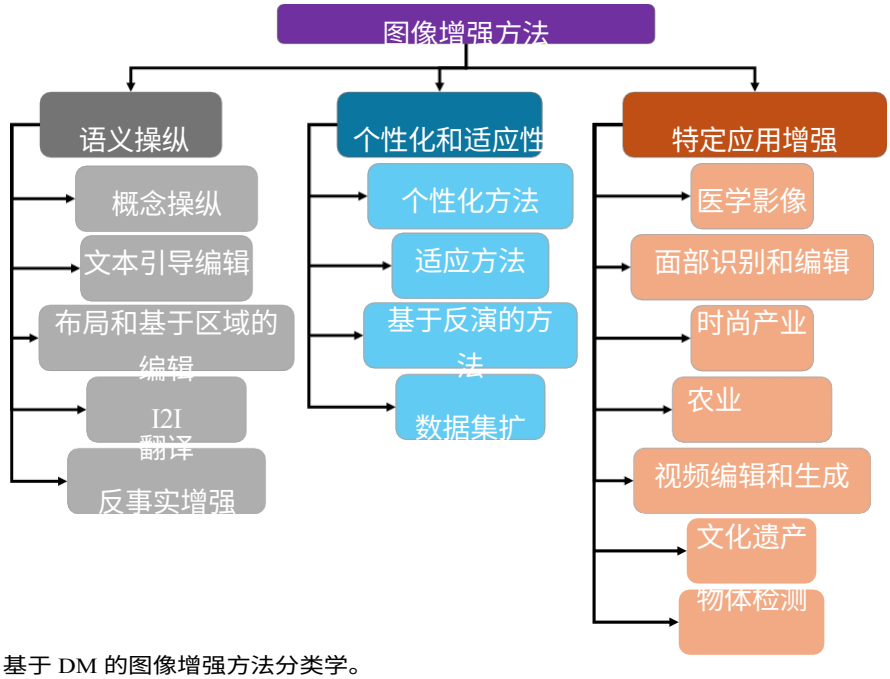

图 2 基于 DM 的图像增强方法分类学。

利用注意力机制，实现图像和文本标记之间的双向信息交换。

3 图像增强扩散模型分类学

本节将概述基于 DM 的图像增强方法。特别是，对各种方法进行了分类，并在图 2 中进行了图解。更具体地说，以每种方法的任务/目标为标准，DM 驱动的图像增强方法可 初步分为以下几大类（而每一类又可进一步分为若干子类，本节稍后将讨论）：

- **语义操作**：目标是对图像进行细粒度的上下文感知修改，同时保持其主要语义内容 Kawar 等人（2023 年）；Kim 等人（2022 年）；Zhang 等人（2023a）。
- **个性化和适应性**：目标是改变
  Ruiz 等人（2023 年）；Gal 等人（2022 年）；Wei 等人（2023 年）。
- **针对具体应用的扩增**：目标是调节增量
  使用特定领域的知识，即引入只对特定应用（如医学成像、面部识别等）有意义的修改 Chambon et al (2022a)；Boutros et al (2023)。

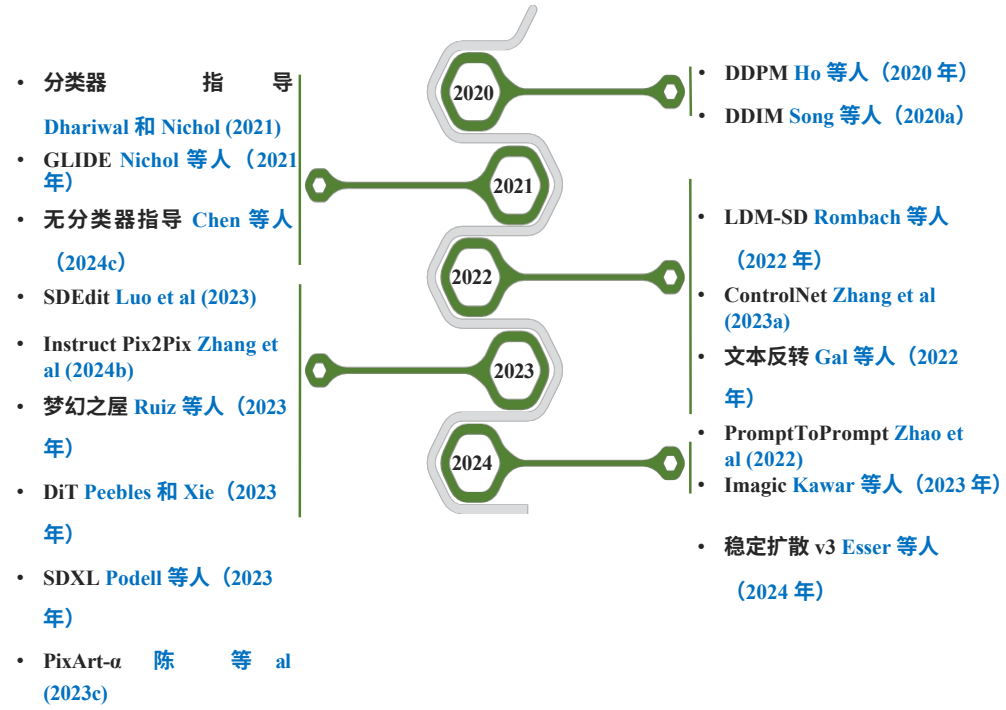

图 3 最近主要的 DM-powered 图像增强方法的时间轴表示。

作为对上述分类的补充，图 3 展示了一个时间轴，其中包含了最近的主要 DMPowered 图像增强方法。每个条目对应的是对研究领域产生重大影响的关键里程碑式工作。当然，较新的工作与更复杂、更先进的 DM 模型/架构有关，也会带来更优越的性能。此外，表 1 对最重要的工作进行了紧凑的总结，同时考虑到了这些工作所属的类别和子类别。

3.1 语义操纵

属于这一类的方法旨在诱导图像发生微妙的、上下文感知的变化，改变图像的解释或传递额外的上下文信息，同时保留核心语义内容，维持连贯而有意义的视觉呈现 Saharia et al (2022a)；Hertz et al (2022)。这些方法有助于生成真实、多样的训练样本，同时保持原始图像的语境。这一类别中更精细、更详细的子类别有：Saharia et al (2022a)；Hertz et al (2022b)：

- **概念操作**：概念处理涉及通过添加、删除或修改对象、属性甚至整个场景来改变图像的语义内涵 Chen 等人（2024c）；Song 等人（2022）。
- **文本引导编辑**：文本引导的编辑利用自然语言描述
  这种方法结合了自然语言处理（NLP）和计算机视觉技术的优势，能够将文本细微解读为视觉变化。这种方法结合了自然语言处理（NLP）和计算机视觉技术的优势，能够将文本细微解读为视觉变化。

表 1 用于图像增强的 DM-powered 方法。

类别	子类别	方法
语义操纵	概念操纵	Chen 等人（2024c）；Luo 等人（2023）；Zhao 等人（2022）；Zhang 等人（2024b）; Song et al (2022); Huang et al (2023a); Brack et al (2022); Rando et al (2022); Schramowski et al (2023); Wasserman et al (2024); Gandikota et al (2023, 2024); Heng and Soh (2024); Kim et al (2023); Kumari et al (2023a); Ni et al (2023); Zhang et al (2024a)
	文本引导编辑	Kawar 等人（2023 年）；Yu 等人（2024 年）；Nichol 等人（2021 年）；Hertz 等人（2022 年）；Chen 等人（2024a）；Lin 等人（2024 年）；Huang 等人（2023b）；Brooks 等人（2023 年）；Wang 等人（2023 年）；杨等人(2024)；金等人(2024)；桑托斯等人(2024)；Av rahami 等人(2023b)；杨等人(2023a)；Kirstain 等人(2023)；Balaji 等人(2022)；耿等人(2024)
	布局和基于区域的编辑	Zeng 等人（2023 年）；Chen 等人（2023d）；Xue 等人（2023 年）；Schnell 等人（2024 年）；Lugmayr 等人（2022 年）；Yu 等人（2023a）；Ackermann 和 Li（2022 年）；Couairon 等人（2022 年）；Av rahami 等人（2022 年，2023a 年）；Sarukkai 等人（2024 年）；Xie 等人（2023b 年）；Xiao 等人（2023 年）；Zhang 等人（2023a 年）；Lev in 和 Fried（2023 年）。
	I2I (图像到图像）翻译	
	反事实增强	
个性化和适应性	个性化方法	Meng et al (2021)；Xu et al (2024)；Kim et al (2022)；Kwon and Ye (2022)；Parmar et al (2023)；Su et al (2022)；Tumanyan et al (2023)；Wang et al (2022a)；Ma et al (2023)；Trabucco et al (2023)；Cao et al (2023)；Saharia et al (2022a)
	适应方法	
	基于反演的方法	Sanchez 等人（2022 年）；Sanchez 和 Tsaftaris（2022 年）；Gu 等人（2023 年）；Madaan 和 Bedathur（2023 年）；Yuan 等人（2022 年）；Parihar 等人（2024 年）；Vendrow 等人（2023 年）
	数据集扩展	
	医学影像	Ruiz 等人（2023，2024）；Gal 等人（2022）；Zhang 等人（2023c）；Kumari 等人（2023b）；Vinker 等人（2023）；Sohn 等人（2023）；Dong 等人（2022）；Chen et al (2024b)；Wei et al (2023)；Gal et al (2023)；Shi et al (2024)；Tewel et al (2023)；Jia et al (2023)；Chen et al (2023b)；Han et al (2023) Hemati 等人（2023）；Wu 等人（2023b）；Dunlap 等人（2022）；Zang 等人（2023）；Zhu 等人（2023）；Qiu 等人（2023） Zhou 等人（2023b,a）；Zhang 等人（2023d）；Li 等人（2023a）；Wallace 等人（2023）；Tang 等人（2024）；Kwon 等人（2022）；Mokady 等人（2023） Zhang et al (2022)；Li et al (2024)；Ye et al (2023)；Wang et al (2022b)；Bansal and Grov er (2023)；Yin et al (2023)；Sheynin et al (2022)；Blattmann et al (2022) Akrout et al (2023)；Sagers et al (2022)；Ali et al (2022)；Pinaya et
	面部识别	

al (2022) ; Hu et	al (2022) ;	Rouzrokh et al (2022)； Wolleb et al (2022)； Chambon et al (2022b,a) ； Guo et al (2023)； Xia et al (2022)； Packhäuser et al (2023)
特定应用增强	and 编辑	布特罗斯等人（2023）；黄等人（2024）
	时尚产业	Li et al (2023b); Kong et al (2023)
	农业	Deng 和 Lu (2023)； Muhammad 等人 (2023)； Chen 等人 (2023a)
	视频编辑 and 生成	Shin et al (2024); Wu et al (2023a)
	文化遗产	Cioni et al (2023)
	物体检测	Fang 等人（2024）；Zhang 等人（2023b）

- **基于布局和区域的编辑**：基于布局和区域的编辑包括修改特定区域或重新排列图像中的元素，以改变其位置或焦点。这些方法对于需要精确控制空间排列和详细修改图像内容的应用来说至关重要 Avrahami 等人（2023a,b）；Zeng 等人（2023）。

- **I2I（图像到图像）翻译**：图像到图像（I2I）翻译方法利用 DM 将源图像转换为不同的目标图像，在保持核心内容的同时改变其风格、纹理或模式特征。这一类方法对于从艺术风格转换到功能医学成像翻译等各种应用至关重要，Saharia 等人（2022a）；Parmar 等人（2023）；Trabucco 等人（2023）。
- **反事实增强**：反事实增强利用 DM 来生成
  生成代表假设情景或假设分析的图像，通常用于增强模型的可解释性和稳健性。这包括生成假设情景，通过改变关键要素来评估潜在结果。这种方法在医学成像和政策制定等领域非常有用，因为在这些领域，了解变量变化的影响至关重要 Sanchez 和 Tsaftaris（2022 年）；Sanchez 等人（2022 年）。

3.2 个性化和适应性

个性化和适应性方法可调整增强过程，以更好地适应特定数据集、任务或用户偏好 Gal 等人（2022 年）；Ruiz 等人（2023 年）。这些方法通过对模型进行微调，使其符合特定要求，从而提高增强数据的相关性和有效性。

- **个性化方法**：个性化方法旨在调整 DM，以生成满足特定用户需求或偏好的内容。这些方法通常涉及对模型进行微调、利用文本或视觉输入，以及优化个性化输出 Gal 等人（2022 年）；Ruiz 等人（2023 年）；Wei 等人（2023 年）。
- **适应方法**：适应性方法：根据不同的领域或不同的应用，定制 DM。
  这些方法对于确保模型在不同的数据集和应用中具有良好的通用性至关重要 Qiu 等人（2023 年）；Hemati 等人（2023 年）。这些方法对于确保模型在不同数据集和应用中具有良好的普适性至关重要 Qiu 等人（2023 年）；Hemati 等人（2023 年）。
- **基于反演的方法**：基于反演的 DM 方法利用以下能力
  莫卡迪等人（2023 年）；张等人（2023 年 d）。
- **数据集扩展**：数据集扩展方法利用 DMs 合成数据集。
  在原始输入数据集的基础上，生成更多图像。其目的是解决小规模数据集的局限性，提高训练图像的多样性。这对提高机器学习模型的泛化和鲁棒性至关重要，尤其是在获取大量标记数据集不切实际的情况下。

3.3 特定应用增强

针对特定应用的扩增方法对扩增过程进行定制，以满足特定应用领域的独特要求和特性，即广泛依赖于使用详细的特定领域知识 Chambon et al (2022a)；Boutros et al (2023)。可以利用这些特性来改进或指导增强过程。具有独特性的典型领域

需求或特性包括医疗成像、面部识别、时尚产业、农业等。

- **医学成像**：用于图像增强的 DMs 被广泛用于生成高保真合成医学图像、增强现有数据集和提高诊断模型的鲁棒性。这些方法可应对数据稀缺、医疗条件多变以及需要匿名训练数据等挑战。
- **其他特定领域的应用**：DM 还有效地应用于
  考虑到每个领域的具体要求，还可以开发其他各种特定领域的应用。例如，面部识别和编辑 Boutros 等人（2023 年）、时装业 Li 等人（2023 年 b）、农业 Deng 和 Lu（2023 年）等。

4 由 DM 驱动的图像增强方法

本节根据第 3 节讨论的方法分类，详细介绍了基于 DM 的图像增强方法的基本原理和机制。重点介绍了每个（子）类别的主要优势，从而为实际应用提供重要启示。

4.1 语义操纵

语义操作改变图像外观，同时部分保留其语义内容 Brooks 等人（2023 年），或操作所描绘的语义概念、潜在文本元素或布局。如第 3 节所述，其子类别包括概念操作、文本引导编辑、基于布局和区域的编辑、图像到图像（I2I）翻译和反事实增强。在许多情况下，编辑的目的是将特定的输入 "参考 "图像转换为增强变体，这一过程一般称为 "编辑"。

4.1.1 概念操纵

有几种概念操作方法专门使用预先训练好的 SD 模型在图像中放置物体，这些方法本身往往不需要训练，如 Chen 等人（2024c）；Luo 等人（2023）；Zhao 等人（2022）；Zhang 等人（2024b）。特别是，他们将 SD 条件设定为由文本提示、网络检索图像或目标对象的高频地图指定的新对象，并在所需位置与场景拼接（由对比语言-图像预训练（CLIP）生成 Radford 等人（2021 年）），以便将其置于输入图像的背景中。可以通过 CLIP 嵌入相似性检查新对象的语义一致

性。例如，Chen 等人（2024c）的方法采用了身份特征提取（使用自监督 DINOv2 模型 Oquab 等人（2023））、细节特征提取和特征注入，通过将 ID 标记和细节映射输入预训练 SD 作为生成最终合成的指导，将目标对象无缝集成到场景中。它还支持额外的控制，如用户绘制的遮罩，以便在推理过程中指示所需的物体形状。

图 4 各种语义操作方法的比较：a) 期望对象（第 1 列），b) 目标图像和期望对象位置（第 2 列），c) 复制粘贴（第 3 列），d) BLIP Li 等人（2022 年）（第 4 列），e) SDEdit Meng 等人（2021 年）（第 5 列），f) ObjectStitch Song 等人（2022 年）（第 6 列）。图片来自 Song et al (2022)。

Song 等人（2022 年）的方法调整了预训练的 SD 模型，以便将现实对象整合到场景中。它使用一个内容适配器模块，将输入图像对象的视觉特征映射到文本嵌入空间，从而为 SD 提供条件。它首先对图像-文本对进行预训练，以学习语义，然后用 SD 进行微调，以保持对象的外观。标清模型吸收背景图像和对象嵌入，生成合成图像。图 4 是一个例子，比较了在目标图像的特定位置添加对象的各种方法。第一列显示所需的对象，第二列显示目标图像和对象的所需位置。

编译器 "Huang 等人（2023a）和 "稳定艺术家 "Brack 等人（2022）的方法利用潜空间中的操作，对图像生成过程进行细粒度控制。合成器 "将图像分解为文本描述、深度图、素描、色彩直方图等代表性因素，并以这些因素为条件训练 DM（用于生成和编辑的图像扩散引导语言（GLIDE）Nichol 等人（2021 年））），从而通过重新组合上述因素实现可定制的内容创建。稳定的艺术家 "Brack 等人（2022 年）使用语义指导（SEGA）引导扩散过程沿着与编辑提示相对应的多个语义方向进行，从而实现微妙的编辑以及构图和风格的变化，而无需遮罩或微调。SEGA 允许用户通过计算原始提示和编辑提示的噪声估计值之间的引导向量来控制潜空间表示，并应用这些引导向量来移动无条件噪声估计值。

像 SD 这样的大型 T2I 模型还可能复制不受欢迎的行为 Birhane 和 Prabhu（

2021 年），或生成不恰当的内容，如受版权保护的艺术作品 Jiang 等人（2023 年）或露骨的图片 Schramowski 等人（2023 年）。为了解决这些问题，人们提出了几种方法，主要分为四类：

图像后处理 Rando 等人（2022 年）：这些方法可在生成图像后过滤掉生成图像中的不恰当内容。

推论 Schramowski 等人（2023 年）：这些方法指导扩散

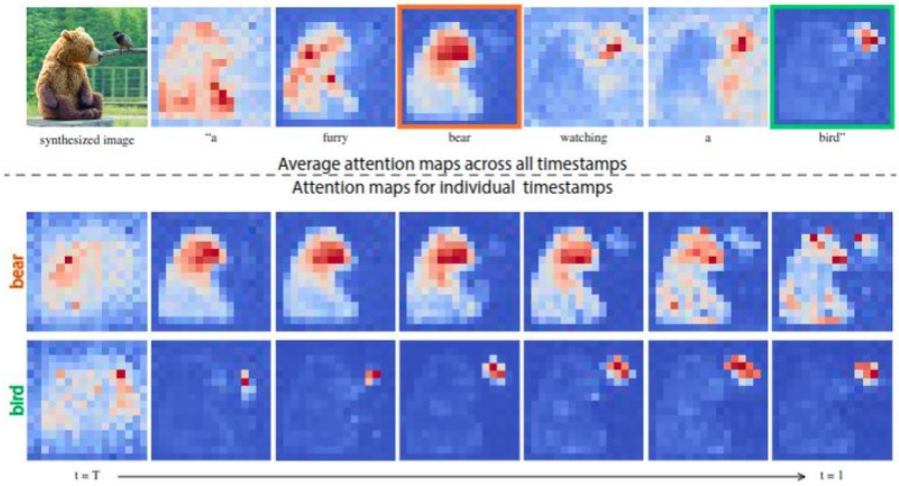

图 5 文本条件扩散图像生成器的交叉注意图。上排显示了合成左侧图像的提示中每个单词的平均注意力掩码。下面几行描绘的是 "熊 "和 "鸟 "这两个词在不同扩散步骤中的注意力图谱。图片来自 Hertz 等人（2022 年）。

在推理过程中避免产生不想要的概念。例如，Safe Latent Diffusion Schramowski 等人（2023 年）定义了一个 "不安全 "的文本概念，并用它来引导扩散过程，避免产生不恰当的内容。

图像涂色 Yu 等人（2023a）；Wasserman 等人（2024）：这些方法可以去除图像中不需要的物体或区域，并用适当的内容对缺失部分进行涂抹。这些方法通常使用遮罩来指定要涂抹的区域。

模型微调 Gandikota 等人（2023，2024）；Heng 和 Soh（2024）；Kim 等人（2023）；Kumari 等人（2023a）；Ni 等人（2023）；Zhang 等人（2024a）：这些方法对预训练的 DM 进行微调，以防止其生成不需要的概念。它们通常使用概念清除 Gandikota 等人（2023 年）、自我蒸发 Kim 等人（2023 年）或退化调整 Ni 等人（2023 年）等方法，从模型的学习表征中删除不需要的概念。

4.1.2 文本引导编辑

基于文本提示的概念操作直接使用文本提示来引导编辑过程，与此不同，文本引导编辑方法首先优化文本嵌入，以重建输入参考图像 Kawar 等人（2023

年）；Yu 等人（2024 年）；Nichol 等人（2021 年）。重构是预训练条件 LDM 的输出，即 RD 过程的结果。此外，目标文本嵌入是文本提示的 CLIP 表示，作为 LDM 的条件。最后，这些方法会在优化的嵌入和目标文本嵌入之间进行插值，从而以插值向量为条件最终生成所需的编辑/增强图像。例如，Imagic Kawar 等人（2023 年）优化了文本嵌入，微调了 DM 以更好地重建输入图像，然后在优化的嵌入和目标文本嵌入之间进行线性插值。

随后，这种插值嵌入将作为微调 DM 的一个条件，从而生成所需的经过编辑的增强图像。

影响逐字生成图像的另一种方法是在扩散过程中加入交叉注意图。如图 5 所示，这些图谱决定了图像的哪些部分（"像素"）在图像创建的不同阶段关注文本提示的特定元素（"标记"）。Hertz 等人（2022 年）和 Chen 等人（2024a）等方法使用交叉注意图将文字描述与图像区域对齐，从而实现细微的图像修改。有些方法侧重于将图像分割成可学习的区域，这些区域可根据文本指令进行单独操作。例如，Lin 等人（2024 年）和 Huang 等人（2023 年b）的方法通过使用预训练模型进行特征提取，将图像划分为多个区域，然后对每个区域分别应用文本指导编辑。这就提高了编辑的粒度，可以更精确地控制图像的特定部分。

一般来说，根据人类指令编辑图像可以对 DM 的操作进行更高级别的控制。这些方法通常利用大语言模型（LLM）通过类似人类的指令来指导编辑过程，如 Brooks 等人（2023 年）；Wang 等人（2023 年）；Yang 等人（2024 年）；Jin 等人（2024 年）；Santos 等人（2024 年）；Geng 等人（2024 年）。例如，InstructPix2Pix Brooks 等人（2023 年）利用 GPT-3 生成编辑指令的合成数据集（例如，输入标题："骑马女孩的照片"，编辑标题："骑龙女孩的照片"），然后利用预训练的 SD 模型生成与前后标题匹配的合成图像。随后，在生成的数据集上训练 SD 模型（通过优化 LDM 损失），从而学会执行图像编辑。无分类器指导也可与文本提示一起用于指导扩散过程，并生成更符合其条件的新图像 Bansal 等人（2023 年）。这样可以提高生成的图像与文本提示之间的一致性，而不需要单独的分类器模型。

Avrahami 等人（2023b）的方法朝另一个方向发展，引入了一种 "空间-文本表示法"，在训练过程中将对象片段的 CLIP 图像嵌入和推理过程中将局部提示的 CLIP 文本嵌入相结合。通过与噪声图像或潜在代码连接，这种表示法被纳入 DM，并相应地对 DM 进行微调（通过最小化 LDM 损失）。多条件无分类器引导被用来控制每个条件的相对重要性，其中包括一个使用单独引导标度的精细变量和一个使用所有条件联合概率的单一标度的快速变量。

Paint-by-Example Yang 等人（2023a）介绍了一种基于范例的编辑方法，

它能自动将参考图像合并到源图像中。自我监督训练利用对象的边界框作为二进制掩码，并利用其中的图像补丁作为参考图像来重建源图像。信息瓶颈利用 CLIP 类标记压缩参考图像，并将解码特征注入扩散过程，以更好地理解内容。强增强功能可减少训练-测试的不匹配，并处理不规则的遮罩形状。通过使用不规则遮罩进行训练，并使用无分类器采样策略来调整编辑图像和参考图像之间的相似度，从而实现可扩展性。

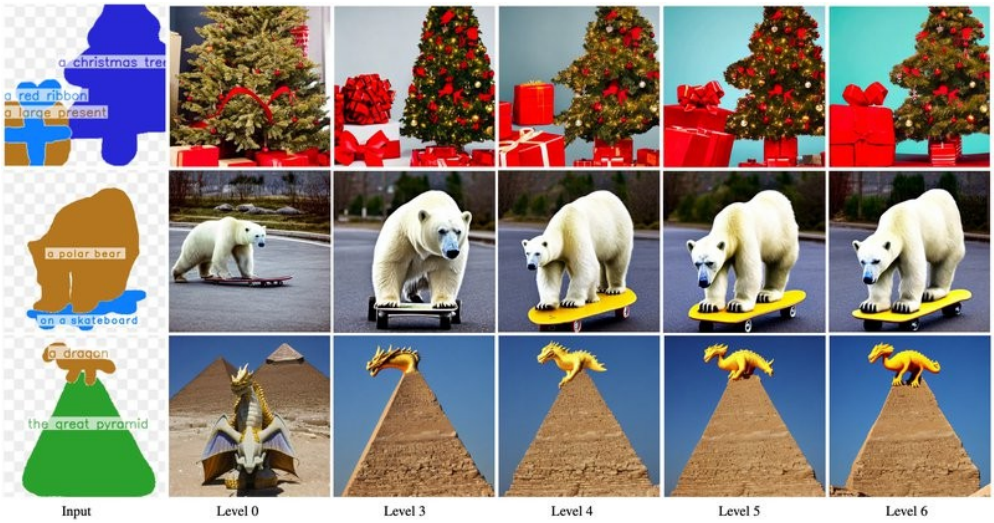

图 6 不同精度水平下的指示性布局和基于区域的编辑结果。对于每个输入布局，图像都是逐步修改的（从相同的噪声开始），因此不同精度水平下生成的图像具有相似的风格。图片来自 Zeng et al (2023)。

X&Fuse Kirstain 等人（2023 年）提出了在 T2I 生成过程中对视觉信息进行调节的一般方法：使用共享的 ResBlocks 分别处理参考图像和输入图像，然后（在注意力块之前）将它们连接起来，以便在两幅图像之间进行交互。eDiff-I Balaji 等人（2022 年）引入了专家去噪器集合，每个专家专门针对生成过程的特定阶段，以捕捉不同的行为，并在不增加推理过程中计算成本的情况下增强模型能力。该研究在训练过程中比较了 T5 文本嵌入 Raffel 等人（2020 年）、CLIP 文本嵌入和可选的 CLIP 图像嵌入。随后，介绍了一种推理方法（Paint-with-words），用户可以从文本提示中选择短语，并创建对象的二进制掩码，作为模型的输入，以控制对象的空间位置。这是通过修改图像和文本特征之间的交叉注意矩阵来实现的。

4.1.3 布局和基于区域的编辑

有几种方法根据布局和区域信息，使用文本提示来指导图像的生成和操作。例如，Zeng 等人（2023 年）和 Chen 等人（2023d）的方法利用文本提示实现了语义图像合成和几何控制。这些方法通常利用单独的布局编码器将空间和语义

信息建模为适合图像生成的格式。Zeng 等人（2023 年）的研究表明，后者包括一个基于精度的掩膜金字塔，它以多种分辨率表示区域形状，并结合文本嵌入。图 6 是这种方法的一个示例。在 Chen 等人（2023d）的研究中，布局编码器通过映射位置将几何布局转化为文本提示、

为了从输入图像的布局中提取空间信息，DM 需要将类别和条件转化为文本标记，然后根据这种布局编码来确定 DM 的条件。

ControlNet Zhang 等人（2023a）是另一种基于布局和区域编辑的强大方法。它引入了一种 DNN 架构，将空间调节控制添加到大型预训练 T2I DM（如 SD）中。ControlNet 创建了模型编码层的可训练副本，并使用 "零卷积 "将其连接到原始模型。这样就可以在小型数据集上针对各种调节任务（如边缘检测、姿势估计和深度映射）进行高效的微调。该架构可将文本提示和调节图像（如边缘映射、姿态映射和深度映射）作为输入进行处理，因此在图像生成和编辑的不同类型空间控制方面具有很强的通用性。

某些方法侧重于从粗略布局或涂鸦中生成图像。例如，Xue 等人（2023 年）的方法通过将语义文本嵌入与空间布局整合到 SD 中，利用 SD 生成自由式图像。它使用文本概念来表示输入布局中的每个语义类别，然后将其编码为文本嵌入。随后，引入一个矫正交叉注意（RCA）模块，将这些文本语义注入扩散模型 U-Net 交叉注意层中的相应布局区域。通过对 U-Net 与布局-图像对上的集成 RCA 进行微调，预训练模型可以根据用户指定的布局和自由形式的文本提示生成图像，从而实现为对象绑定新属性和生成未见对象类别等功能。Schnell 等人（2024 年）利用 ControlNet Zhang 等人（2023a 年）的方法生成了以涂鸦标签和文本提示为条件的合成图像。该方法采用无分类器引导（10% 的条件涂鸦输入被随机丢弃并替换为可学习的嵌入），并引入编码比率来调整生成图像的多样性和逼真度（通过执行较少的 FD 步骤），从而在模式覆盖率和样本保真度之间进行权衡。

Inpainting 方法 Lugmayr 等人（2022 年）；Yu 等人（2023a）根据给定的遮罩利用 DM 来填充图像的缺失区域。这些方法通常将遮罩图像和遮罩本身作为 DM 的条件，然后生成内容来填充遮罩区域。Ackermann 和 Li（2022 年）、Couairon 等人（2022 年）、Avrahami 等人（2022 年、2023 年a）等方法的一个子集采用了带有遮罩引导的多阶段扩散过程，以实现高分辨率图像编辑和内绘。

拼贴扩散 Sarukkai 等人（2024 年）的方法将用户定义的图层序列作为输入，称为拼贴。它由一个全图像文本字符串组成，其中描述了要生成的整个图像，以

及从后向前排列的图层序列；每个图层由一个 RGBA 图像（alpha 遮罩输入图像）和一个描述它的文本组成。该方法修改了 DM 中的文本-图像交叉注意力，以实现空间保真度，同时扩展 ControlNet 以保持每层的外观保真度。拼贴扩散技术还允许用户通过指定所需的噪音水平来控制每一层的协调性和保真度之间的权衡，从而实现逐层图像编辑。与此类似，SmartBrush Xie 等人（2023b）利用文本和形状指南对 DMs 进行对象涂色，使用户能够根据文本描述和对象遮罩控制涂色内容。FastComposer Xiao 等人（2023）是一种无需调整的多主体图像生成方法，它增强了文本提示

使用图像编码器从输入图像中提取视觉特征。它使用预训练的 CLIP 编码器将文本提示和输入图像作为嵌入模型，然后使用多层感知器 (MLP) 用视觉特征增强文本嵌入。该方法使用主体增强的图像-文本配对数据集对图像编码器、MLP 模块和 U-Net 进行去噪损失训练，同时使用参考主体的分割掩码定位交叉注意图，以防止多主体生成中的身份混淆。它还在迭代去噪中采用了延迟主体条件，以平衡身份保护和可编辑性。

Levin 和 Fried（2023 年）设计的方法允许使用 "变化图 "对应用的修改量进行像素级的编辑控制。后者是一个矩阵，其维度与原始输入图像的空间分辨率相同，描述了在每个位置应用的编辑强度。可以使用不同的方法生成 "变化图"，如 Segment-Anything Kirillov 等人（2023 年）、MiDas Ranftl 等人（2020 年），甚至是手动绘制的变化图。该方法可在潜空间中运行，并已在 SD、SDXL、Kandinsky Razzhigaev 等人（2023 年）和 StabilityAI 的 DeepFloyd IF（2023 年）预训练 LDM 上进行了评估，使用文本提示（手动创建，或将输入图像反转为 CLIP 和 BLIP Li 等人（2022 年））以及变化图来指导推理过程。

4.1.4 I2I（图像到图像）翻译

有几种方法利用条件 DM 进行 I2I 翻译。例如，SDEdit Meng 等人（2021 年）利用随机微分方程指导的条件 DM 进行图像编辑。CycleNet Xu 等人（2024 年）和 DiffusionCLIP Kim 等人（2022 年）利用基于 CLIP 的损失对条件 DM 进行微调，以确保生成的图像与目标图像的文本描述相匹配。Palette et al. Saharia 等人（2022a）采用条件 DM 学习各种任务的分布 $p(y\,x)$，如着色、内画、JPEG 还原和剪裁。该模型采用具有自我关注层的 U-Net 架构，通过连接对输入图像进行调节。在训练过程中，它会预测添加到原始图像中的噪声，最大限度地减少预测噪声与实际噪声之间的 L2 或 L1 损失。然后，推理过程涉及 1000 个时间步的迭代去噪，从高斯噪声开始，逐步完善输出。该方法的优势在于它能用单一架构处理多项任务，无需针对具体任务进行定制（图 7）。

另一类方法侧重于将风格和内容分离，以实现更有针对性的 I2I 翻译。例如，Kwon 和 Ye（2022 年）的方法利用单独的编码器对风格和内容进行编码，然后将这些表示注入 DM 以生成翻译后的图像。其他方法则探索少镜头或

零镜头的 I2I 翻译。例如，pix2pix-zero Parmar 等人（2023 年）首先使用 DDIM 和 BLIP 反转输入图像以获得噪声图。然后利用自相关目标对这些噪声图进行正则化处理，以提高可编辑性。通过为源域和目标域生成不同的句子，并计算它们的 CLIP 嵌入之间的平均差，自动发现编辑方向。为了在编辑过程中保持图像结构，一种新颖的交叉注意力引导技术

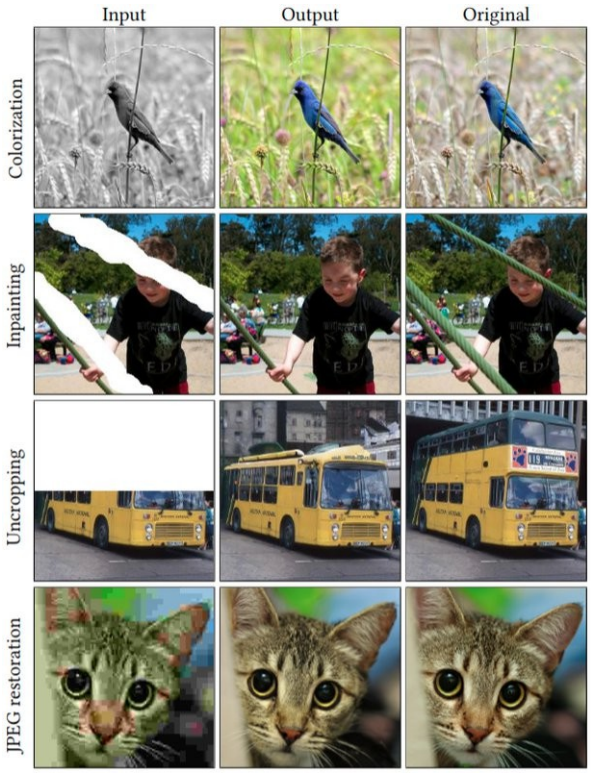

图 7 图像到图像 DM 的指示性示例，该 DM 可在各种任务中生成高保真输出，而无需结合特定任务进行定制。图片来自 Saharia 等人（2022a）。

该方法将编辑后的交叉注意图与与原始结构相关的参考图进行匹配。该方法以 SD 模型为基础，利用其 CLIP 调节层和交叉注意层。另一方面，双扩散隐含桥（Dual Diffusion Implicit Bridges，DDIBs）苏等人（2022 年）利用两个独立训练的 DM（一个用于源域，一个用于目标域）来执行 I2I 翻译，而无需在配对数据上进行联合训练。

Tumanyan 等人（2023 年）的方法将从源图像扩散过程中提取的空间特征和自我注意图注入到翻译图像的生成过程中，从而实现了对生成结构的精细控制，同时符合目标文本提示。此外，Wang 等人（2022a）的方法利用预训练的 T2I DM（GLIDE）作为生成先验，包括一个 64x64 分辨率的基本模型和一个 256x256 分辨率的上采样模型。该框架采用编码器-解码器架构，其中特定任务编码器将输入条件（如分割掩码、草图）映射到预先训练好的扩散解码器的语义潜空间。扩散

解码器在不同的文本图像对上进行预训练，使潜在空间以高度语义化的文本嵌入为条件。该管道包括预训练 DM、使用两阶段微调方案训练特定任务编码器、使用对抗训练微调扩散上采样器，以及采样

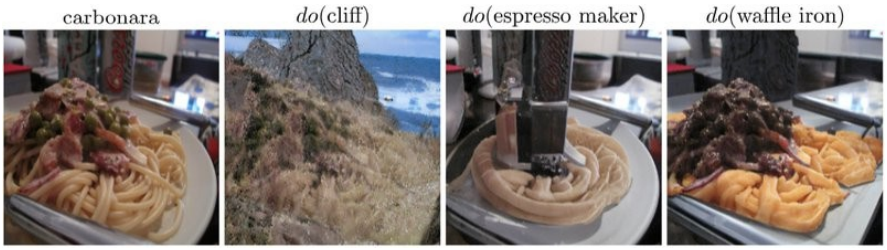

图 8 Diff-SCM 在 ImageNet 上生成的反事实。从左至右：从数据分布中抽取的随机图像及其反事实 do(class)，对应于 "图像应如何变化才能被分类为另一个类别？图片来自 Sanchez 和 Tsaftaris (2022)。

此外，Ma 等人（2023 年）还提出了另一个相关框架，旨在联合生成主题和文本条件图像。此外，Ma 等人（2023 年）还提出了另一个相关框架，旨在联合生成以主题和文本为条件的图像。DM（SD）可生成与输入文本语义一致的高质量图像，同时保留输入图像的主题。它利用 CLIP 编码器将文本和图像映射到统一的多模态潜空间。融合采样策略平衡了统一条件和纯文本条件之间的噪声预判定，以避免过度拟合背景信息，即最终噪声估计使用给定的比率融合两种预测。

DA-Fusion Trabucco 等人（2023 年）是另一种旨在改变图像语义内容的方法。它依赖于预训练的 SD，并根据类标签指导修改输入参考图像。通过在文本编码器中插入和微调新的标记嵌入，使用文本反转方法（见第 4.2 节）来处理 DM 训练数据集之外的新视觉概念。输入图像会以随机的时间间隔拼接到扩散过程中，以引导生成，而不是从头开始生成。随机化的拼接时间步提供了不同的增强强度。

MasaCtrl Cao 等人（2023 年）在 U-Net 中引入了一种相互自我关注机制，允许模型从源图像中查询相关的局部结构和纹理，同时与目标编辑提示保持一致。该架构修改了标准的 U-Net 模块，将自我关注转换为相互自我关注，其中查询特征源自当前的去噪过程，而关键特征和值特征则源自输入图像的扩散过程。为了防止前景和背景元素之间的混淆，MasaCtrl 利用从交叉注意图中提取的掩码，采用了掩码引导的互自注意策略。

4.1.5 反事实增强

许多方法利用条件 DM 生成反事实图像。例如，Sanchez 等人（2022 年）的方法利用条件 DM，通过对肿瘤掩膜的模型条件化，生成带有肿瘤的大脑 MRI 图像的 "健康 "反事实。Diff-SCM Sanchez 和 Tsaftaris（2022 年）采用条件 DM 估算因果框架中干预措施的效果，具体方法是

在干预变量 *do*(*class*)（即 "图像应如何变化才能被归类为另一个类别"）上对模型进行调节。例如，对于给定的公园里一只狗的照片，如果 *do*(*cat*)成立，则狗应被猫取代，而图像的其他部分保持不变（图 8）。

很多方法都侧重于为特定应用生成反事实。例如，MEDJOURNEY Gu 等人（2023 年）通过将 DM 条件化为期望变化的文本描述来生成反事实医学图像，而 Madaan 和 Bedathur（2023 年）以及 Yuan 等人（2022 年）的方法则生成反事实图像来解释模型的行为并提高其鲁棒性。另一方面，一组不同的方法利用反事实增强来解决数据集和模型中的偏差和公平性问题。例如，Parihar 等人（2024 年）的研究利用条件 DM 生成反事实图像，以平衡数据集中敏感属性（如性别、种族）的分布。该模型以所需的属性分布为条件，生成与之相匹配的图像，同时保留图像的其他方面。

另一组方法侧重于用生成的反事实或分布外（OOD）示例扩展给定数据集，以提高在扩展数据集上训练的模型（如分类器）的鲁棒性。例如，Vendrow 等人（2023 年）的方法利用条件 DM 生成反事实示例，改变输入图像的特定属性（如物体位置、颜色、纹理），同时保留整体场景结构。这些反事实例子可用于诊断和缓解 OOD 数据的模型故障。

4.2 个性化和适应性

个性化和适应性是图像增强任务中非常常见的方法。在本小节的其余部分，我们将根据第 3 节中介绍的分类方法，详细介绍基于 DM 的各种方法。

4.2.1 个性化方法

个性化意味着必须对 DM 进行调整，以生成满足特定用户需求或偏好的内容。这些方法通常涉及微调预训练模型、利用文本或视觉输入以及优化个性化输出。很多方法都利用微调来个性化预训练的 T2I DM，以生成主题驱动的内容。例如，DreamBooth Ruiz 等人（2023 年）使用一小组（3-5 张）单一主体的图像，配以包含唯一标识符和主体所属类别名称（如 "A [V] dog"）的文本提示，在少量学习设置中对 DM 进行了微调。这些图像包含一个特定的主题，是 DM 需要学习的训练数据集，而文本提示只是描述新的主题。特别是，该方法应用了类保存损失，

以确保生成的图像保持主体的身份。同样，HyperDreamBooth Ruiz 等人（2024 年）也是一种高效个性化 DM 的方法，它引入了三个关键组件，即轻量级 DreamBooth（LiDB）、用于快速个性化的超级网络以及秩松弛快速微调。LiDB 利用一个秩-1 LoRA Hu 等人（2021 年）权重空间来分解秩-1 LoRA Hu 等人的权重空间。

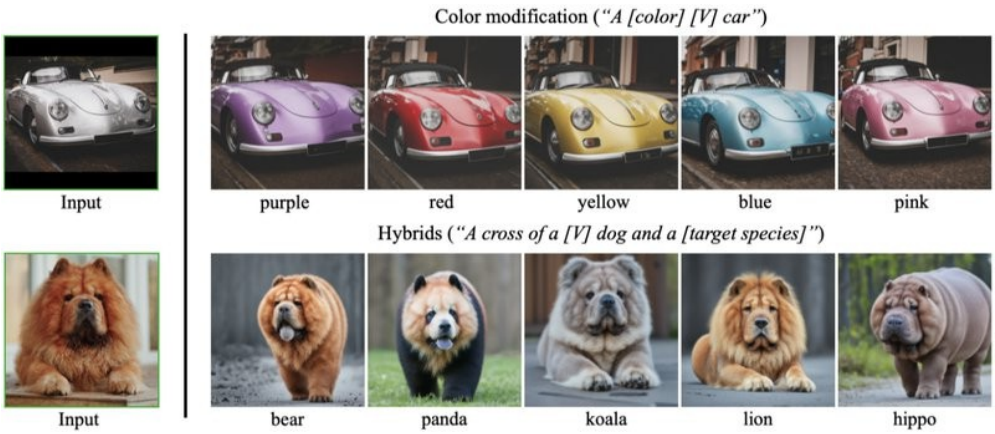

图 9 以个性化为导向的图像增强示例，在此示例中，需要对给定对象的属性进行修改，同时保留其关键的视觉特征，以塑造其身份。图片来自 Ruiz 等人（2023 年）。

随机正交不完整基础，从而产生更小的个性化模型。此外，由 ViT 编码器和变压器解码器组成的超网络利用扩散去噪和权重空间损失，预测输入图像的 LiDB 残差。最后，等级松弛快速微调通过松弛 LoRA 等级和使用预测的 HyperNetwork 权重进行微调来捕捉精细级别的细节，从而在保持快速个性化的同时提高受试者的保真度（图 9）。

另一种方法是利用所谓的 "文本反转 "Gal 等人（2022 年）来生成个性化的 T2I。这一过程利用一小组描述特定视觉概念的图像，学习代表该概念的新文本嵌入（称为 "伪词"）。这样，模型就能使用包含伪词的自然语言描述生成概念图像。ProSpect Zhang 等人（2023c）对这一想法进行了扩展，学习了一系列伪词（称为 "提示词谱"），这些伪词捕捉了概念的不同视觉属性（如材料、风格、布局）。

在不同的基础上，另一组方法侧重于结合多个概念进行个性化生成。例如，Kumari 等人（2023b）和 Vinker 等人（2023）的方法采用了一种组合机制，不同的概念由独立的文本嵌入表示，这些嵌入可以组合生成新颖的图像。StyleDrop Sohn 等人（2023 年）和 DreamArtist Dong 等人（2022 年）采用了一种基于风格的方法，生成图像的风格由学习到的嵌入控制。

另一种概念化方法包括利用基于编码器的方法来个性化 T2I 生成。例如，ELITE Wei 等人（2023 年）利用图像编码器将用户提供的输入图像映射到文本嵌

入，然后将文本嵌入用于预训练 SD 的条件。这样，该模型就能生成与输入的风格和内容相匹配的个性化图像。同样，Gal 等人（2023 年）的方法采用了交叉关注机制，将图像编码器中的视觉特征注入不同层的 DM 中。

SuTI Chen 等人（2024b）是一个学徒学习框架，它在图像-文本集群上为每个科目训练专家 DM（Imagen Saharia 等人（2022b）），每个集群包含 3-10 幅图像-文本。然后，在 K 个群组上对 K 个专家 DM 进行微调，以学习属于群组的特定主题。随后，在这些专家 DM 的输出上训练一个 DM。通过这种方法，SuTI 可以将众多专家模型中的专业知识整合到一个通用模型中。在推理过程中，它可以为新主题生成定制图像，而无需进行某种优化。

InstantBooth Shi 等人（2024 年）采用了另一种方法，无需在部署时对 DM 本身进行微调，即可实现个性化生成。它引入了可学习的图像编码器，将输入图像转换为文本嵌入，并引入适配器层，注入丰富的视觉特征，以更好地保存身份信息。原始模型权重被冻结，只对新的组件进行训练。在文本提示中，唯一标识符标记 V 代表输入概念，同时对输入图像进行裁剪和背景屏蔽。

Perfusion Tewel 等人（2023 年）是一种基于键锁定秩-1 编辑的紧凑而高效的架构，它解决了个性化生成中的关键难题，如过拟合和保持高视觉保真度，同时允许创造性控制。键锁定秩-1编辑法通过在交叉注意层中锁定所学概念与其超类别的键来防止过拟合，从而确保概念继承生成先验而不会偏离太多。该方法通过在扩展潜空间中学习特定概念值来捕捉概念的独特外观，从而保持了较高的视觉保真度。此外，秩-1 更新是根据当前编码与目标概念之间的相似性进行控制的，从而允许在推理时对每个概念的影响进行细粒度控制。

Taming Encoder Jia 等人（2023 年）是一种生成用户指定的定制对象图像的方法，而无需像以前的方法那样对每个对象进行冗长的优化。具体来说，给定输入图像和所需输出的文本描述， 图像对象编码器（CLIP）计算对象嵌入，而文本编码器计算文本嵌入。然后将这两部分传递给 Imagen Saharia 等人（2022b），通过一次前向传递生成最终输出。训练数据使用字幕模型（PaLI Chen 等人（2022 年））和二进制掩码来隔离对象。该框架在特定领域和通用领域数据集上进行联合训练，使用交叉引用正则化和对象嵌入去除。除对象编码器外，整个网络都采用与预训练 DM 相同的目标进行调整。

DisenBooth Chen 等人（2023b）是一个主体驱动 T2I 生成框架，它利用文本身份保留嵌入和视觉身份无关嵌入。其中，身份保留分支使用 CLIP 文本编

码器将主体身份映射到特殊文本标记，而身份无关分支则使用预训练的 CLIP 图像编码器和可学习掩码来提取身份无关特征，然后使用带跳过连接（Adapter）的 MLP 将这些特征与文本特征空间对齐。在微调过程中，DisenBooth 利用去噪、弱去噪和对比嵌入目标，并采用了

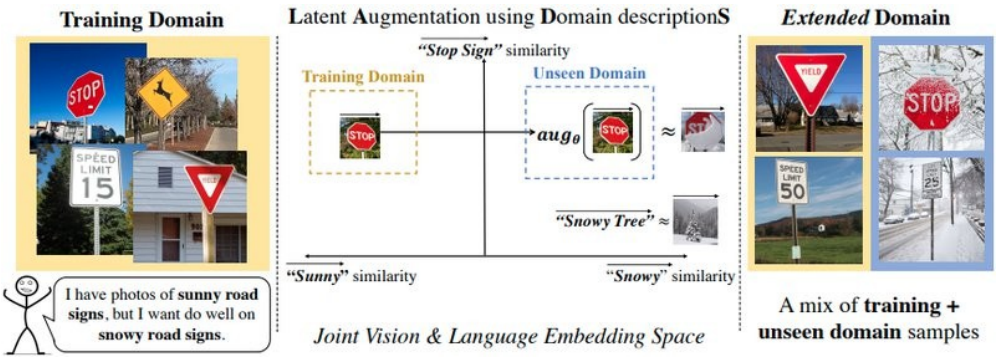

图 10 基于 DM 的图像增强领域适应性示例，其中在源领域图像上训练的模型可应用于目标领域的图像。图片来自 Dunlap 等人（2022 年）。

使用 LoRA 进行参数高效微调。经过微调后，DisenBooth 可将主体的身份标记与其他文本描述相结合，并可选择继承参考图像的特征，从而实现灵活可控的生成。

HiPer Han 等人（2023 年）使用单张输入图像，将预训练的 SD 模型个性化，用于文本驱动的图像处理。它将 CLIP 文本嵌入空间分解为初始标记和结束标记，以保留主体身份。特别是，在给定输入图像、输入图像的提示和目标所需图像的提示后，HiPer 会优化 DM 潜在空间中的嵌入。在推理过程中，目标提示的嵌入会与经过优化的 HiPer 嵌入进行连接，从而调节预训练的 SD 模型，生成既能保留主体身份又能包含目标提示语义的操作图像。HiPer 的主要贡献在于它对文本嵌入空间进行了分解，以分别控制语义和身份，只需要一张源图像，而无需对 DM 进行微调。

4.2.2 适应方法

相当一部分适应方法依赖于微调来调整预训练的 DM 以适应新的领域。例如，Hemati 等人（2023 年）的方法利用预训练的 SD 模型生成合成图像，以弥补各领域之间的差距，减少训练数据的非一致性。它从一个领域接收图像作为输入，从另一个同类领域接收引导属性（文本提示或图像），利用 LDM 创建插值合成图像。Wu 等人（2023b）的方法引入了一种优化方法，以在每个去噪步骤中找到输入图像描述和目标图像描述文本嵌入的最佳软组合权重。

通过基于 CLIP 的损失对权重进行优化，以匹配目标属性，同时使用感知损失保留其他内容。Dunlap 等人（2022 年）的方法利用了特定领域的 CLIP 模型，引导扩散过程生成与目标领域相匹配的图像（图 10）。此外，Zang 等人（2023 年）的方法利用特定领域的判别器，在训练过程中过滤掉低质量或领域外的样本。

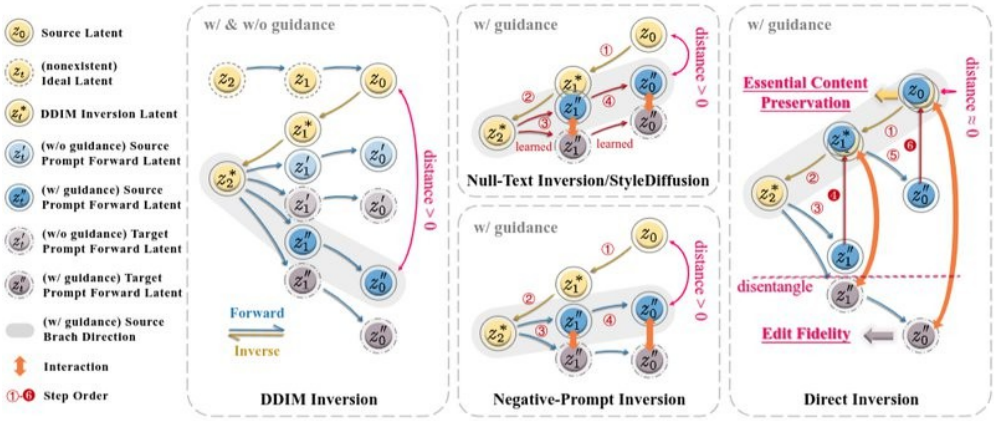

图 11 各种基于反演的扩散图像编辑方法。图片来自 Ju 等人（2023 年）。

另一类方法侧重于根据特定任务或应用调整预训练 DM。具有代表性的例子有细胞核分割 Yu 等人（2023b）或膝关节骨关节炎严重程度分类 Chowdary 等人（2023），他们分别在细胞核图像或 X 光图像数据集上进行了微调。

DomainStudio Zhu 等人（2023 年）提出了一种最新方法，通过引入成对相似性损失来保持域适应过程中生成样本之间的相对距离，并设计一种高频细节增强方法来提高生成质量，从而利用有限的数据将预训练 DM 适应目标域。此外，Qiu 等人（2023 年）提出的正交微调法（OFT）是一种使 DM 适应新任务的方法，它将可训练的正交矩阵注入模型的注意层，以乘法转换神经元权重。这些矩阵初始化为同一性，并保持正交，从而实现稳定的微调，预服务于模型的语义知识。OFT 接收带有文本提示的主体图像或控制信号作为输入，并根据主体身份或控制信号生成图像，同时与提示相匹配。神经元的正交变换可保持它们的成对角度，从而加快收敛速度，提高质量和可控性。

4.2.3 基于反演的方法

基于反演的方法通常会优化潜码，即用户提供的输入图像的潜空间代表，该图像包含 DM 要学习的对象，这样就可以使用 DM 重构图像。随后，对潜码进行处理，以引导 DM 执行所需的编辑或增强。图 11 总结了各种基于反转的图像增强方法。

有许多方法利用反演生成高质量的合成数据，用于训练其他 DNN。例如，

Zhou 等人（2023b）和 Zhou 等人（2023a）的方法通过将真实图像反转到潜空间，然后在反转图像周围采样新的 潜码，从而生成合成图像。这有助于

对原始数据集进行多样化和现实化的扩充，从而显著提高下游分类模型的性能。

另一组方法则采用反转技术来实现精确的图像编辑和风格转换。例如，空文本反转 Mokady 等人（2023 年）的方法由两个主要部分组成，即枢轴反转（使用 DDIM 反转估计初始扩散轨迹）和 "空文本 "优化（仅优化每个扩散时间步的非传统 "空文本 "嵌入）。给定一幅输入图像及其相关的文字说明，该方法会对图像进行反演，以获得潜码和优化的空文本嵌入。然后就可以利用这些信息，在保留模型权重和条件嵌入的情况下，通过修改文本提示，使用诸如 P2P Hertz 等人（2022 年）的方法直观地编辑图像。

按照类似的概念，Zhang 等人（2023d）的方法利用预训练的 SD，实现了从单幅参考画到输入图像的艺术风格转换。具体而言，该方法引入了一个基于注意力的文本反转模块，通过多层交叉注意力，从参考绘画的 CLIP 图像嵌入中学习风格表示（"[C]"）。然后将学习到的文本嵌入编码成 DM 标题调节格式，以指导生成将输入图像的内容与参考绘画的艺术风格相结合的新图像。具体来说，该方法包括提取参考画作的图像嵌入，通过反转模块学习相应的文本嵌入，将其编码为 DM 调节格式，对内容图像进行随机反转以获得初始潜噪声图，并根据文本嵌入和反转噪声图（输入图像的噪声版本）生成输出图像。

与此密切相关的是，Li 等人（2023a）的方法可以学习将输入图像反转为值嵌入，同时保留原始模型中的键和注意力映射。各自的架构由一个冻结的 CLIP 图像编码器、一个可学习的映射网络和 SD 模型组成。键控制输出图像的结构，而值则决定对象的风格。编辑过程包括 DDIM 反转以生成潜码和注意力图，训练映射网络以重建这些潜码和注意力图，同时保留类似对象的注意力，并使用训练好的网络和目标提示嵌入进行编辑。

按照不同的研究思路，一组特殊的方法侧重于改进反演过程本身。例如，EDICT Wallace 等人（2023 年）利用耦合 DM 进行反演，其中一个模型学习将图像映射到潜在空间，另一个模型学习将潜在代码映射回原始图像空间。与直接优化潜码相比，这可以实现更准确、更高效的反演。LocInv Tang 等人（2024 年）利用局部感知反转过程，优化潜码以匹配输入图像的空间注意图，这有利于在编辑过程中保留图像的局部结构和细节。此外，Kwon 等人（

2022 年）的方法引入了非对称反向过程（Asyrp），在冻结的预训练 DM 中发现了一个名为 "h 空间 "的语义潜空间。它通过移动 U-Net 架构瓶颈特征图中的预测噪声来修改 RD 过程，同时保留指向当前时间步的方向。这打破了破坏性干扰和

图 12 基于 DM 的数据集扩展图像增强示例。图片来自 Bansal 和 Grover (2023)。

可对生成的图像进行语义操作。h 空间具有编辑所需的特性，如同质性、线性、可合成性、鲁棒性和跨时间步的一致性。

4.2.4 数据集扩展

条件 DM 通常用于生成从给定数据集的基本分布中采样的合成图像。例如，Zhang 等人（2022 年）的方法在一个小型数据集上训练一个类条件 DM，然后用它生成每个类的额外图像。生成的示例使用分类器进行过滤，以确保它们符合实际情况，并与目标类别分布相匹配。同样，Li 等人（2024 年）的方法利用语义指导的 DM 来生成与原始数据集的语义布局和对象类别相匹配的合成图像。Wang 等人（2022b）的方法包括一种单尺度像素级 DDPM，可学习单张自然图像的内部斑块分布。它采用了一个 U-Net 去噪网络，具有受限的斑块级感受野，使其能够捕捉图像斑块统计数据，而无需记忆整个图像。该模型在单一尺度上进行训练，避免了渐进式增长方法中的误差累积。Ye 等人（2023 年）的方法利用时间条件、U-Net 驱动的 LDM 和两阶段训练过程来生成逼真的合成图像并提高分类性能。第一阶段是在无标记数据上进行大规模预训练，以学习无条件图像合成的通用特征。第二阶段是在小型标注数据集上对模型进行微调，从而在潜在分类器的指导下进行有条件合成。同样，Bansal 和 Grover（2023 年）的方法生成的合成示例在特征空间上与训练数据接近，但在 图像空间上却相去甚远，这有助于提高模型对新视觉概念的泛化能力（图 12）。

另一类方法则利用DM的能力，从文本描述中生成高质量图像。例如，TTIDA Yin 等人（2023 年）对预先训练好的文本到文本（T2T）模型（GLIDE Nichol 等

人（2021 年））进行微调，以生成物体和场景的各种文本描述，然后使用 T2I DM 生成相应的图像。这样就可以生成具有丰富注释的大规模合成数据集，用于训练更准确、更强大的视觉模型。

KNN-Diffusion Sheynin 等人（2022 年）包括一个多模式 CLIP 编码器、一个不可训练的检索索引和一个以检索嵌入为条件的可训练 DM。在训练过程中，模型接收图像作为输入，同时使用 CLIP 图像嵌入及其 k 个近邻作为生成条件。在推理过程中，模型接收一个文本提示作为输入，而 CLIP 文本嵌入及其 k 个近邻被用来作为生成条件。在输入嵌入和检索到的邻域的指导下，DM 对噪声矢量进行迭代去噪，以生成输出图像。检索到的邻域可以弥补图像和文本嵌入之间的分布差距。

检索增强扩散模型（RDM） Blattmann 等人（2022 年）将可训练的条件扩散模型（DM）与包含各种视觉示例的固定外部数据库和不可训练的检索函数相结合。在训练过程中，该方法使用 CLIP 嵌入从数据库中检索每幅图像的 k 个近邻，并使用预训练编码器对其进行编码。随后，它将生成式解码头置于这些编码表征的条件下，生成目标图像。在推理过程中，数据库和检索功能可以灵活交换，以实现无条件采样、类别条件采样、文本条件采样或风格转移，具体方法是根据不同的标准或使用具有不同视觉风格的数据库检索邻近图像。这种方法为相对较小的生成模型增加了一个大的外部存储器，使其能够根据检索到的相关信息组成新的图像，而不是记忆全部的训练数据；从而缩小了模型的大小，同时提高了性能和灵活性。

4.3 特定应用增强

本小节讨论针对特定应用的 DM-powered 图像增强方法，即考虑到仅存在于所研究应用领域的特定事实和特征的方法。目前已研究过的常见应用领域包括医疗成像、面部识别、物体检测和农业等。

4.3.1 医学影像

在医学成像领域，图像增强方法通常采用条件DM生成合成医学样本。例如，Akrout 等人（2023 年）和 Sagers 等人（2022 年）的方法在皮肤病变数据集上训练类条件 DM，并利用它们生成每种病变类型的额外示例。生成的图像用于增强训练数据，提高皮肤病分类模型的性能。同样，Ali 等人（2022 年）和 Packhäuser 等人（2023 年）的方法利用 DMs 生成合成胸部 X 光图像，并

利用这些图像训练更强大的胸部异常检测模型。还有一组特殊的方法专注于生成特定类型的医学图像，如脑部磁共振成像（MRI）或视网膜光学断层扫描（OCT）。例如，Pinaya 等人（2022 年）的方法在脑核磁共振成像图像数据集上训练 DM，并用它生成具有以下特征的合成图像

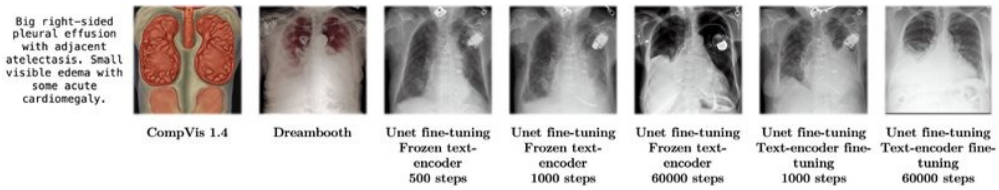

图 13 基于特定应用的 DM 图像增强示例（医学成像），给定特定文本提示。图片来自 Chambon 等人（2022a）。

不同的神经状况。Hu 等人（2022 年）的方法是在退行性 OCT 图像上训练 DM，并用它生成去噪和超分辨版本的图像。

还有一些方法利用 DM 进行图像着色和异常检测。例如，Rouzrokh 等人（2022 年）的方法通过在肿瘤掩膜上调节模型，训练 DM 来对带有肿瘤的脑部 MRI 图像进行涂抹。这样就能从部分或损坏的扫描图像中生成完整逼真的大脑图像。相比之下，Wolleb 等人（2022 年）的方法采用 DDIM 从患病个体生成合成健康图像。具体而言，U-Net 架构经过训练，可以迭代地对噪声表示进行去噪，而二元分类器则会引导去噪过程向健康类别发展。该方法假设健康图像和患病图像的训练数据不成对且只有图像级标签。对于未见过的测试图像，DDIM 去噪过程将其解剖信息编码为噪声表示，然后利用分类器的引导进行迭代去噪，生成相应的健康合成图像。异常图是根据输入图像和合成图像之间的像素差异计算得出的，突出显示了异常区域。噪声水平和分类器梯度规模控制着保留输入细节和转换到健康类别之间的权衡。

另一组方法则侧重于将预先训练好的 DM 调整到医疗领域。例如，Chambon 等人（2022b）使用文本反转和交叉注意控制，在胸部 X 光图像和放射报告数据集上对预先训练的 T2I DM 进行了微调。这样就能根据自然语言描述生成具有特定异常和属性的合成胸部 X 光图像。同样，Chambon 等人（2022a）在胸部 X 光片和放射学报告数据集上对预训练 SD 进行了微调，使其能够根据文本提示生成合成医学图像。其关键部件是一个用于压缩输入的冻结变量自动编码器（VAE）编码器、一个以来自 CLIP 编码器的文本嵌入为条件的 U-Net 去噪器和一个 VAE 解码器。在推理过程中，U-Net 在文本提示的引导下对随机噪声进行编码和逐步去噪，最终生成反映所述成像结果的合成胸部 X 光片（图 13）。

最近，有几种方法利用更先进、更复杂的 DM 架构，旨在进一步提高生成图像的增强性能和质量。Guo 等人（2023 年）的方法引入了 PD-DDPM，这是一种用于医学图像分割的加速 DM，它使用预分割网络生成有噪声的分割预测，然后将预测结果用于图像分割。

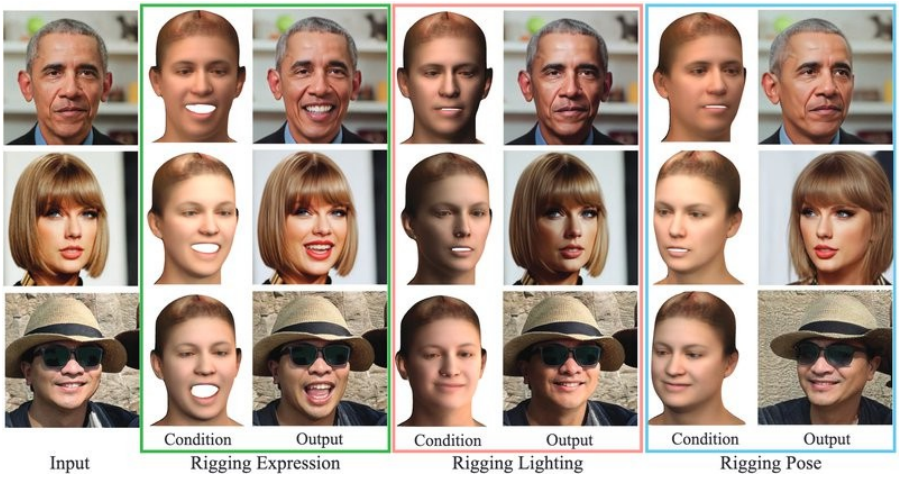

图 14 基于 DM 的特定应用图像增强（面部识别和编辑）示例，涉及不同的表情、光线和姿势。图片来自 Ding 等人（2023 年）。

与普通 DM 相比，去噪步骤更少。Xia 等人（2022 年）的方法采用的 DDPM 是在成对的低剂量计算机断层扫描（CT）和正常剂量 CT 图像上进行训练的。在 FD 过程中，高斯噪声会根据时间变化计划逐渐添加到正常剂量图像中，从而产生符合标准正态分布的噪声图像。在 RD 过程中，U-Net 以相应的低剂量图像为条件，通过预测和去除每个时间步的噪声，学会从噪声图像中恢复干净的正常剂量图像。为了提高采样效率，在 RD 过程中集成了一个快速常微分方程（ODE）求解器，称为 DPM 求解器 Lu 等人（2022a），从而可以用更少的步骤实现更快的采样，同时与原始 DDPM 相比保持或提高去噪性能。

4.3.2 其他特定领域的应用

除医学成像外，DM 还被成功应用于各种特定领域，详情如下。

人脸识别与编辑：Boutros 等人（2023 年）和 Huang 等人（2024 年）的方法侧重于生成高保真合成人脸，以提高识别准确率。这些方法通常使用在大规模人脸数据集上训练的条件 DM，并通过从所学分布中采样生成新的人脸。有些方法还利用多级文本相关增强吴等人（2023c）或分离表示丁等人（2023）等方法，对生成的人脸进行更精细的控制，以改变特定属性或表情（图 14）。

时尚产业：Li 等人（2023b）和 Kong 等人（2023）的方法可实现逼真的虚拟试

穿体验，让用户可视化不同体型和姿势的服装。这些方法通常使用在以下基础上训练的条件 DM

图 15 使用不同输入图像的基于 DM 的特定应用图像增强示例（时尚行业）。图片来自 Li 等人（2023b）。

服装图像和相应身体姿势的数据集，并通过对所需服装和姿势的调节生成新的试穿图像（图 15）。

农业 Deng 和 Lu（2023 年）、Muhammad 等人（2023 年）以及 Chen 等人（2023a）的方法侧重于增强植物病害检测和杂草识别的数据集，从而提高农业模型的准确性。这些方法通常采用在具有不同病害或杂草类型的植物图像数据集上训练的条件 DM，并通过从所学分布中采样生成新示例。

视频编辑和生成：Shin 等人（2024 年）和 Wu 等人（2023a 年）的方法能够根据文本提示进行高质量的视频操作和生成。这些方法通常依赖于三维或视频扩散模型的使用，这些模型可以根据文字描述或输入视频生成连贯的视频帧序列。

文化遗产：Cioni 等人（2023 年）采用 LDM 方法，以文字描述为条件，生成多种语义一致的艺术品变体。具体来说，LDM 以原始艺术品图像及其编码标题为输入，在卷积自动编码器学习的潜在空间中执行条件去噪扩散过程。通过改变生成种子，LDM 可以输出艺术品的多种变体，这些变体保留了标题中描述的内容。这些合成图像-标题对增强了原始艺术数据集，缩小了自然图像和艺术作品之间的领域差距，同时提高了艺术概念的视觉基础。扩充数据集包含真实数据和合成数据，有助于更有效地训练下游视觉语言模型，如图像标题和跨模态检索模型。

物体检测：Fang 等人（2024 年）和 Zhang 等人（2023 年 b）的方法利用 DM 生成合成数据，大大提高了模型训练和性能。这些方法利用在物体检测数据集上训练的条件 DM，通过对所需的物体类别和边界框进行条件化，生成新的样本。

5 评估指标

要了解 DM 驱动的图像增强方法对视觉图像分析任务的影响，评估这些方法的性能至关重要。本节概述了用于评估此类框架的功效和效率的方法。

5.1 定量评估

定量评估包括测量模型性能的数值改进，以及增强图像的 感知质量和多样性。这类评估提供了可直接比较和分析的客观指标。

5.1.1 提高模型性能

主要的定量评估包括测量学习到的模型（如分类器）的下游任务性能指标的（潜在）改进，这些模型是在包含由 DM 生成的增强图像的情况下训练的。与仅使用原始非增强数据集实现训练的基线相比，下游模型的性能提高越多，则认为进行增强的 DM 越好。 最常见的 情况如下：

- **分类任务：**准确率、精确度、召回率和 F1 分数等指标用于评估分类模型的性能。这些指标全面概述了模型区分不同类别的能力 Trabucco et al (2023)；Chowdary et al (2023)；Packhäuser et al (2023)；Sagers et al (2022)；Bansal and Grover (2023)。
- **分割任务：**诸如 "交集大于联合" （IoU）和 "骰子"（Dice）等指标系数用于评估分割模型。IoU 衡量的是预测分割与地面实况分割之间的重叠程度，而 Dice 系数评估的是这两者之间的相似程度 Xie 等人（2023a）；Schnell 等人（2024）；Sanchez 等人（2022）；Valvano 等人（2024）。
- **物体检测任务：**平均精度 (mAP)、精确度 (preci-precision)、平均mAP 计算不同召回级别的平均精确度，是评估物体检测器精确度的标准指标。另一方面，精确度和召回率还能帮助我们深入了解检测器正确识别物体并将误报率降至最低的能力 Fang 等人（2024 年）；Zhang 等人（2023b）。

分析这些指标有助于确定增强图像在多大程度上提高了下游模型的预测能力。

5.1.2 感知质量和多样性

要评估增强图像的质量和多样性，最常用的指标如下：

- **弗雷谢特入门距离（FID）：** FID Heusel 等人（2017 年）测量生成的图像与真实图像在预训练 Inception 网络 Szegedy 等人（2016 年）的特征空间中的分布之间的距离。它同时反映了生成图像的质量和多样性。FID 分数越低，表明生成的图像质量越高、种类越多 Esser 等人（2024 年）；Ho 等人（2020 年）；Rombach 等人（2022 年）；Xie 等人（2023b）；Avrahami 等人（2023b）；Gandikota 等人（2024 年）；Zhang 等人（2023a）；Pinaya 等人（2022 年）；Couairon 等人（2022 年）；Fu 等人（2024 年）。
- **入门得分（IS）：** IS Salimans 等人（2016 年）根据预训练的入门模型对类别预测的置信度来评估生成图像的质量。高 IS 值意味着生成的图像是多样化的，而且每幅图像都能被高置信度地识别为属于某个特定类别 Blattmann 等人（2022 年）；Gafni 和 Wolf（2020 年）；Luo 等人（2023 年）；Chen 等人（2023a）。
- **内核起始距离（KID）：** KID Bińkowski 等人（2018 年）是 FID 的替代方法，它估算生成的图像样本与真实图像样本的特征表示之间的最大平均差异。KID 分数越低，表明质量越好，与真实图像的相似度越高 Muhammad et al (2023)；Li et al (2023b)；Kumari et al (2023b)；Meng et al (2021)。
- **感知度量：** 学习感知图像补丁相似度（LPIPS）Zhang 等人（2018）和结构相似性指数测量（SSIM）Wang 等人（2004）等指标用于评估生成样本与真实样本之间或图像编辑任务的输入与输出之间的感知相似性Kulikov 等人（2023）；Wang 等人（2022b）；Li 等人（2023a）；Zhang 等人（2024c）；Qiu 等人（2023）；Chambon 等人（2022b）；Xu 等人（2024）。

通常情况下，对上述所有指标进行估算并将其结合在一起，就能全面了解增强图像的质量和可变性。

5.2 定性评估

定性评估涉及对增强图像的视觉逼真度和相关性的主观评估。此类评估对于确保 DM 生成的合成数据不仅在技术上可靠，而且在感知上令人信服、在语境上合适至关重要。

5.2.1 增强图像的视觉质量

评估增强图像的视觉质量涉及专家检查，目的是确定合成图像与真实图像的相似程度。专家评估员需要分析以下主要方面的因素：

- **逼真度：** 评估增强图像是否与真实图像难以区分，重点关注纹理、光照和颜色

一致性等特征 Sanchez 和 Tsaftaris (2022)； Kwon 和 Ye (2022)； Dong 等人 (2022)。

- **细节保留：** 必须确保增强图像保留
  任务所需的关键细节，如细粒度纹理和结构完整性 Ruiz 等人（2023 年）； Kawar 等人（2023 年）； Dong 等人（2022 年）。

- **编辑一致性：**专家们会评估增强图像与原始输入和编辑说明在语义上保持一致性的程度。这包括评估编辑是否准确地应用于预期区域，同时保留图像的整体背景和结构 Wang 等人（2023 年）；Kawar 等人（2023 年）；Tang 等人（2024 年）；Li 等人（2023a）。

总之，上述主观评估指标有助于验证生成图像的视觉真实性和可用性。

5.2.2 相关性和背景适宜性

对增强图像的相关性和语境适宜性进行评估，可确保这些图像保持语义连贯，并适合手头的具体应用。这主要涉及以下几个方面：

- **语境一致性：**专家审查增强图像是否符合任务的预期背景。例如，在医学成像中，合成图像应准确反映被检查疾病或被建模条件的特征 Gandikota 等人（2023 年）；Han 等人（2023 年）；Ruiz 等人（2024 年）；Wang 等人（2023 年）；Zhang 等人（2023d）；Tewel 等人（2023 年）。
- **语义准确性：**增强图像应传达正确的语义信息，避免任何误导或无意义的变化。这对于准确呈现物体和场景至关重要的应用来说至关重要，Huang 等人（2023b）；Li 等人（2024）。
- **特定任务特征：**还可评估增强图像的相关性基于其对手头特定任务的实用性。例如，在物体检测中，图像应包含正确标注和定位的物体，以提高模型训练的效率 Chambon 等人（2022a）；Valvano 等人（2024）。

上述定性评估可确保增强图像不仅看起来逼真，而且能令人满意地达到预期目的。

6 挑战与未来研究方向

使用 DM 驱动的方法进行图像增强是提高训练数据集的多样性和质量的一种有前途的方法。然而，尽管该领域在近期取得了快速进展和重大进步，但仍有一些公开挑战有待解决，这些挑战同时也是未来的研究方向。这些挑战大致可分为一般挑战和图像增强挑战，前者涉及 DM 的一般应用，与特定应用

情况无关；后者则对图像增强尤为重要。

6.1 计算成本和效率

与任何基于 DM 的架构的应用相关的一个重要因素是其计算成本非常高。特别是，DM 需要大量的计算机资源，而且训练和推理都很耗时，这可能会延误其开发和部署。相反，经过训练的生成对抗网络（GAN）生成器可直接合成输出图像，即只需一次前向传递。DM 的迭代去噪过程使得这种方法相对难以扩展到现实世界的应用、大型数据集和复杂任务中。例如，Rombach 等人（2022 年）用一个 A100 40GB GPU 在 LAION-5B 数据集上训练 SD 需要 200,000 多个 GPU 小时。DDIM和DPM-Solver++ Lu等人（2022b）等方法旨在加快采样过程并提高效率。然而，还需要进一步研究开发更高效的架构和采样方法，以便在减少计算开销的同时生成高质量的图像。

6.2 缺乏精细控制和可解释性

DMs 的可解释性和可控性普遍存在问题，因此很难理解它们是如何生成输出结果的（例如，在特定条件下）。用户精确控制生成图像中特定属性、对象或区域的能力有限，这可能会妨碍这些方法在某些应用中的实用性。开发解释基于 DM 模型输出的方法是一个重要的研究领域。已经提出了无分类器引导、交叉注意控制或空间条件（如遮罩、区域等）等方法来缓解这些问题，并实现对采样过程的控制。不过，这一领域还需要进一步研究。

6.3 生成数据的多样性和真实性有限

最普遍的挑战之一是 DM 生成的合成图像的（相对）多样性和真实性受到限制。虽然 DM 在生成高质量图像方面表现出令人印象深刻的能力，但它们往往难以捕捉真实世界数据分布的全部多样性和复杂性。这种局限性会导致合成数据与真实数据之间存在领域差距，从而影响使用生成的数据训练下游任务（如图像分类和物体检测）的效果。迄今为止，已经有几种方法旨在解决这一问题，它们采用了一些方法来提高生成样本的多样性和真实性，如使用语言增强和后过滤（如基于 CLIP 的过滤）。然而，要开发出更先进的方法，在各个领域生成真正多样化和真实的合成数据，还需要进行更多深入的研究。

6.4 模型过拟合和灾难性遗忘

本研究涉及的许多方法面临的最常见挑战之一是模型过拟合和灾难性遗忘问题。特别是

过拟合是指模型在学习时过于贴近训练数据，从而影响了对新的未见数据的泛化。这种情况在使用有限的训练数据时尤其容易出现问题，医学影像等领域通常就是这种情况。另一方面，灾难性遗忘指的是可训练模型在对新数据或任务进行微调时，容易遗忘之前学习过的表征。对于那些旨在将预训练模型适应特定领域或风格的方法来说，这是一个主要障碍。在有限的数据集上直接对大型模型进行微调，会导致快速丢失最初学习的知识结构。最近一些旨在缓解这一问题的方法包括：Zeng 等人（2024 年）；Zhong 等人（2024 年）；Zając 等人（2023 年）利用各种方法，包括选择性参数更新、双流架构和对去噪过程的精心管理，在适应新任务或新概念的同时保持模型的泛化。

6.5 评估指标和基准

评估由 DM 驱动的图像增强的有效性本身仍然是一项挑战。FID 和 KID 等传统指标可能无法全面评估/捕捉生成样本的质量和多样性，尤其是在复杂或专业领域。此外，由于缺乏标准化的基准和评估协议，因此很难对不同的方法进行比较，也很难评估它们的泛化和增强能力。为此，开发更全面、更通用以及针对特定领域的评估指标，同时建立通用基准和数据集，是未来研究的一个重要领域，以便进行更严格、更一致的评估。

6.6 伦理考虑和偏见

大规模 T2I DM 通常是在网络抓取的数据集上进行训练的，这些数据集可能包含有害的刻板印象、攻击性内容以及与性别、种族、年龄和其他敏感属性相关的偏见。这些偏见会在生成的图像中被放大，导致不公平的表述和社会成见的延续。另一个道德方面的考虑因素是，在训练数据集中使用受版权保护或私人的数据时，没有得到适当的同意或归属。这引起了人们对生成图像的所有权和公平使用以及侵犯隐私的可能性的极大关注。

7 结论

图像增强是现代计算机视觉的一项基本任务，因为它可以用逼真的合成样本来增

强训练数据集，允许对给定的参考图像进行上下文和语义感知的自动编辑等。扩散模型（DM）在生成逼真、多样的图像，捕捉高维图像数据中的复杂关系和结构方面显示出了巨大的潜力。此外，利用类标签、文本描述或视觉提示对生成过程进行调节的能力，可以有针对性地进行增强，根据手头的任务生成满足特定要求的图像。

本研究全面概述了以 DM 为动力的图像增强技术的最新进展，对现有方法的主要类别进行了分类，深入分析了以 DM 为动力的技术在语义操作、个性化和适应性以及特定应用图像增强方面的实际应用，并回顾了相关的性能评估指标。未来研究的重要方向包括提高DM的效率，即降低计算成本和提高可扩展性，增强生成图像的可解释性和可控性，以及提高同步数据的多样性和真实性。最后，开发新的、可靠的评估指标和解决伦理方面的问题，是在图像增强领域应用 DMs 并取得进展的关键所在。

致谢。本文的研究成果得到了欧盟地平线欧洲研究与发展计划（Horizon Europe Research and Development Programme）的资助，资助协议编号为 101073876（停火）。

声明

资助产生这些成果的研究得到了欧盟委员会第 101073876 号资助协议（停火）的资助。

利益冲突 作者没有与本文内容相关的利益冲突需要声明。

作者贡献 Panagiotis Alimisis、Ioannis Mademlis 和 Georgios Th.帕帕佐普洛斯进行了文献综述并撰写了手稿草稿。Panagiotis Radoglou-Grammatikis 和 Panagiotis Sarigiannidis 对手稿进行了审阅和编辑。Georgios Th.帕帕佐普洛斯负责获得实施研究的资金。

参考资料

ArXiv preprint arXiv:221012965

Agustsson E, Mentzer F, Tschannen M, et al (2017) Soft-to-hard vector quantization for end-to-end learning compressible representation.神经信息处理系统进展 30

Akrout M, Gyepesi B, Holló P, et al (2023) 基于扩散的 皮肤病分类数据增强：从原

始医学数据集到全合成图像的影响。In：医学影像计算和计算机辅助干预国际会议，施普林格，第 99-109 页

Ali H, Murad S, Shah Z (2022) Spot the fake lungs：利用神经扩散模型生成合成医学图像。In：爱尔兰人工智能和认知科学会议，施普林格，第 32-39 页

Asperti A, Evangelista D, Marro S, et al (2023) 去噪生成模型的图像嵌入。人工智能评论》56（12）：14511-14533

Avrahami O、Lischinski D、Fried O (2022) 自然图像文本编辑的混合扩散。In：IEEE/CVF 计算机视觉和模式识别会议论文集》，第 18208-18218 页

Avrahami O、Fried O、Lischinski D (2023a) 混合潜在扩散。ACM 图形事务（TOG）42（4）：1-11

Avrahami O, Hayes T, Gafni O, et al (2023b) Spatext：用于可控图像生成的空间文本表示法。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 18370-18380 页

Balaji Y, Nah S, Huang X, et al (2022) ediff-i：带有专家去噪器集合的文本到图像扩散模型。arXiv 预印本 arXiv:221101324

Bansal A, Chu HM, Schwarzschild A, et al (2023) 扩散模型通用指南。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 843-852 页

Bansal H, Grover A (2023) Leaving reality to imagination： arXiv preprint arXiv:230202503

Bińkowski M, Sutherland DJ, Arbel M, et al (2018) Demystifying mmd gans. arXiv preprint arXiv:180101401

Birhane A, Prabhu VU (2021) 大型图像数据集：计算机视觉的惨胜？In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1536-1546

Blattmann A, Rombach R, Oktay K, et al (2022) Semi-parametric neural image synthesis.

Boutros F, Grebe JH, Kuijper A, et al (2023) Idiff-face：通过模糊身份条件扩散模型进行基于合成的人脸识别。In：IEEE/CVF 计算机视觉国际会议论文集》，第 19650-19661 页

Brack M, Schramowski P, Friedrich F, et al (2022) The stable artist： arXiv preprint

arXiv:221206013

Brooks T, Holynski A, Efros AA (2023) Instructpix2pix：学习遵循图像编辑指令。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 18392-18402 页

Cao H, Tan C, Gao Z, et al (2024) A survey on generative diffusion models.IEEE 知识与数据工程论文集

Cao M, Wang X, Qi Z, et al (2023) Masactrl：用于一致图像合成和编辑的无调谐互自注意控制。In：IEEE/CVF 计算机视觉国际会议论文集》，第 22560-22570 页

Chai J, Zeng H, Li A, et al (2021) 计算机视觉中的深度学习：新兴技术与应用场景评述。机器学习与应用 6:100134

Chambon P, Bluethgen C, Delbrouck JB, et al (2022a) Roentgen：胸部 X 射线生成的视觉语言基础模型。

Chambon P, Bluethgen C, Langlotz CP, et al (2022b) Adapting pretrained vision-language foundational models to medical imaging domains. arXiv preprint arXiv:221004133

Chen D, Qi X, Zheng Y, et al (2023a) Deep data augmentation for weed recognition enhancement：基于扩散概率模型和迁移学习的方法。In: 2023 ASABE Annual International Meeting, American Society of Agricultural and Biological Engineers, p 1

Chen H, Zhang Y, Wu S, et al (2023b) Disenbooth：用于主题驱动文本到图像生成的保全身份不纠缠调谐。arXiv 预印本 arXiv:230503374

Chen J, Yu J, Ge C, et al (2023c) *Pixart-α*：ArXiv preprint arXiv:231000426

Chen K, Xie E, Chen Z, et al (2023d) Geodiffusion：用于对象检测数据生成的文本提示几何控制。arXiv 预印本 arXiv:230604607

Chen M, Laina I, Vedaldi A (2024a) 通过交叉注意引导实现免训练布局控制。In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 5343-5353 页

Chen W, Hu H, Li Y, et al (2024b) 通过学徒学习实现主体驱动的文本到图像生成。神经信息处理系统进展 36

Chen X, Wang X, Changpinyo S, et al (2022) Pali：联合缩放的多语言语言图像模型。arXiv preprint arXiv:220906794

Chen X, Huang L, Liu Y, et al (2024c) Anydoor：零镜头对象级图像校正。In：

IEEE/CVF 计算机视觉与模式识别会议论文集》，第 6593-6602 页

Chowdary PN, Vardhan GV, Akshay MS, et al (2023) Enhancing knee osteoarthri- tis severity level classification using diffusion augmented images.

Chung H, Lee ES, Ye JC (2022) 使用正则化反向扩散对先生图像进行去噪和超分辨率处理。电气和电子工程师学会医学影像论文集》42（4）：922-934

Cioni D, Berlincioni L, Becattini F, et al (2023) Diffusion based augmentation for captioning and retrieval in cultural heritage.In：IEEE/CVF 计算机视觉国际会议论文集》，第 1707-1716 页

Couairon G, Verbeek J, Schwenk H, et al (2022) Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:221011427

Croitoru FA, Hondru V, Ionescu RT, et al (2023) Diffusion models in vision：概览。电气和电子工程师学会模式分析与机器智能论文集

Deng B, Lu Y (2023) 稳定扩散用于可可和杂草数据集的数据增强。

DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout.

Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis.神经信息处理系统进展 34:8780-8794

Ding Z, Zhang X, Xia Z, et al (2023) Diffusionrig：为面部外观编辑学习个性化先验。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 12736-12746 页

Dong Z, Wei P, Lin L (2022) Dreamartist： Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. arXiv preprint arXiv:221111337

Dunlap L, Mohri C, Guillory D, et al (2022) Using language to extend to unseen domains.In：第十一届学习表征国际会议

Dunlap L, Umino A, Zhang H, et al (2023) 利用基于自动扩散的增强技术实现视觉数据集的多样化。神经信息处理系统会议

Esser P, Kulal S, Blattmann A, et al (2024) Scaling rectified flow transformers for high-resolution image synthesis.In：第四十一届机器学习国际会议

Fang H, Han B, Zhang S, et al (2024) 通过可控扩散模型进行物体检测的数据增强。

In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 1257-1266 页

Fu Y, Chen C, Qiao Y, et al (2024) Dreamda：用扩散模型生成数据增强。arXiv 预印本 arXiv:240312803

Gafni O, Wolf L (2020) Wish you were here：情境感知人类生成。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 7840-7849 页

Gal R, Alaluf Y, Atzmon Y, et al (2022) An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:220801618

Gal R, Arar M, Atzmon Y, et al (2023) 基于编码器的领域调整，实现文本到图像模型的快速个性化。ACM 图形学论文集（TOG）42（4）：1-13

Gandikota R, Materzynska J, Fiotto-Kaufman J, et al (2023) Erasing concepts from diffusion models.In：IEEE/CVF 计算机视觉国际会议论文集》，第 2426-2436 页

Gandikota R, Orgad H, Belinkov Y, et al (2024) 扩散模型中的统一概念编辑。In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 5111-5120 页

Geng Z, Yang B, Hang T, et al (2024) Instructdiffusion：视觉任务的通用建模界面。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 12709-12720 页

Gu Y, Yang J, Usuyama N, et al (2023) Medjourney：通过多模态患者旅程的指导学习生成反事实医学图像

Guo X, Yang Y, Ye C, et al (2023) Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation.In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, pp 1-5

Han I, Yang S, Kwon T, et al (2023) 通过稳定扩散为图像处理提供高度个性化的文本嵌入。

Hemati S, Beitollahi M, Estiri AH, et al (2023) Cross domain generative augmentation：arXiv preprint arXiv:231205387

Heng A, Soh H (2024) 选择性遗忘：深度生成模型中遗忘的持续学习方法。神经信息处理系统进展 36

Hertz A, Mokady R, Tenenbaum J, et al (2022) Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:220801626

Heusel M, Ramsauer H, Unterthiner T, et al (2017) 通过双时间尺度更新规则训练的 Gans 收敛到局部纳什均衡。神经信息处理系统进展 30

Ho J, Salimans T (2022) Classifier-free diffusion guidance.

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models.神经信息处理系统进展 33:6840-6851

Hu D, Tao YK, Oguz I (2022) 利用扩散概率模型对视网膜 oct 进行无监督去噪。In：医学成像 2022：图像处理》，SPIE，第 25-34 页

Hu EJ, Shen Y, Wallis P, et al (2021) Lora：arXiv preprint arXiv:210609685

Huang L, Chen D, Liu Y, et al (2023a) Composer：ArXiv preprint arXiv:230209778

Huang N, Tang F, Dong W, et al (2023b) Region-aware diffusion for zero-shot text-driven image editing. arXiv preprint arXiv:230211797

Huang Z, Geng H, Wang H, et al (2024) 利用扩散模型进行人脸识别的数据增强。In：CVPR 2024 Workshop SyntaGen：利用生成模型合成视觉数据集，URL https://openreview.net/forum?id= GXmlanJ6rC

Jia X, Zhao Y, Chan KC, et al (2023) 使用文本到图像扩散模型的零微调图像定制驯服编码器。

Jiang HH, Brown L, Cheng J, et al (2023) 艾艺术及其对艺术家的影响。In：Proceed-ings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society.美国计算机协会，纽约州纽约市，AIES '23, p 363-374, https://doi.org/10.1145/3600211.3604681, URL https://doi.org/10.1145/3600211.3604681

Jin Y, Ling P, Dong X, et al (2024) Reasonpix2pix：用于高级图像编辑的指令推理数据集。arXiv 预印本 arXiv:240511190

Ju X, Zeng A, Bian Y, et al (2023) Direct inversion：用 3 行代码提升基于扩散的编辑。arXiv 预印本 arXiv:231001506

Kang G, Dong X, Zheng L, et al (2017) Patchshuffle regularization.

Kawar B, Zada S, Lang O, et al (2023) Imagic：基于文本的真实图像编辑与扩散模型。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 6007-6017 页

Kazerouni A, Aghdam EK, Heidari M, et al (2023) Diffusion models in medical

imaging：全面调查。医学图像分析》第 102846 页

Kebaili A, Lapuyade-Lahorgue J, Ruan S (2023) 医学影像数据增强的深度学习方法：综述。影像学杂志》9（4）：81

Kim G, Kwon T, Ye JC (2022) Diffusionclip：用于稳健图像处理的文本引导扩散模型。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 2426-2435 页

Kim S, Jung S, Kim B, et al (2023) Towards safe self-distillation of internet-scale text-to-image diffusion models.

Kirillov A, Mintun E, Ravi N, et al (2023) Segment anything.In：IEEE/CVF 计算机视觉国际会议论文集》，第 4015-4026 页

Kirstain Y, Levy O, Polyak A (2023) X&fuse: Fusing visual information in text-to-image generation.

Kong C, Jeon D, Kwon O, et al (2023) 利用现成的扩散模型进行多属性时尚图像处理。In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 848-857 页

Kulikov V, Yadin S, Kleiner M, et al (2023) Sinddm：单一图像去噪扩散模型。In：国际机器学习会议，PMLR，第 17920-17930 页

Kumari N、Zhang B、Wang SY 等 (2023a) 文本到图像扩散模型中的消融概念。In：IEEE/CVF 计算机视觉国际会议论文集》，第 22691-22702 页

Kumari N、Zhang B、Zhang R 等 (2023b) 文本到图像扩散的多概念定制。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 1931-1941 页

Kwon G, Ye JC (2022) Diffusion-based image translation using disentangled style and content representation.

Kwon M, Jeong J, Uh Y (2022) Diffusion models already have a semantic latent space.

Levin E, Fried O (2023) Differential diffusion：让每个像素都有自己的力量。arXiv 预印本 arXiv:230600950

Li B, Xu X, Wang X, et al (2024) 利用扩散模型的语义引导生成图像增强方法用于图像分类。In：AAAI 人工智能会议论文集》，第 3018-3027 页

Li J, Li D, Xiong C, et al (2022) Blip：引导语言图像预训练，实现统一的视觉

语言理解和生成。In：国际机器学习会议，PMLR，第 12888-12900 页

Li S, van de Weijer J, Hu T, et al (2023a) Stylediffusion：基于文本编辑的提示嵌入反转。arXiv 预印本 arXiv:230315649

Li Z, Wei P, Yin X, et al (2023b) 利用姿态-服装关键点引导涂色的虚拟试穿。In：IEEE/CVF 计算机视觉国际会议（ICCV）论文集，第 22788-22797 页

Lin Y, Chen YW, Tsai YH, et al (2024) 通过可学习区域进行文字驱动的图像编辑。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 7059-7068 页

Lu C, Zhou Y, Bao F, et al (2022a) Dpm-solver：用于扩散概率模型采样的快速ode求解器（约10步）。神经信息处理系统进展 35:5775-5787

Lu C, Zhou Y, Bao F, et al (2022b) Dpm-solver++：扩散概率模型引导采样的快速求解器 arXiv preprint arXiv:221101095

Lugmayr A, Danelljan M, Romero A, et al (2022) Repaint：使用脱色扩散概率模型进行涂色。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 11461-11471 页

Luo XJ, Wang S, Wu Z, et al (2023) Camdiff：arXiv preprint arXiv:230405469

Ma Y, Yang H, Wang W, et al (2023) Unified multi-modal latent diffusion for joint subject and text conditional image generation. arXiv preprint arXiv:230309319

Madaan N, Bedathur S (2023) Diffusion-guided counterfactual generation for model explainability.In：XAI in Action：过去、现在和未来的应用

Meng C, He Y, Song Y, et al (2021) Sdedit: Guided image synthesis and editing with stochastic differential equations.

Mokady R, Hertz A, Aberman K, et al (2023) 利用引导扩散模型编辑真实图像的空文本反演。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 6038-6047 页

Muhammad A, Salman Z, Lee K, et al (2023) 利用扩散模型的力量进行植物病害图像增强。https://doi. org/10.3389/fpls.2023.1280496

Mumuni A, Mumuni F (2022) Data augmentation：现代方法综述。Array 16:100258. https://doi.org/https://doi.org/10.1016/j.array.2022.100258, URL https://www.sciencedirect.com/science/article/pii/ S2590005622000911

Ni Z, Wei L, Li J, et al (2023) Degeneration-tuning：使用加扰网格屏蔽稳定扩散中不需要的概念In：第 31 届 ACM 国际多媒体会议论文集》，第 8900-8909 页

Nichol A, Dhariwal P, Ramesh A, et al (2021) Glide：利用文本引导的扩散模型实现逼真图像的生成和编辑。arXiv 预印本 arXiv:211210741

Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models.In：国际机器学习会议，PMLR，第 8162-8171 页

Oquab M, Darcet T, Moutakanni T, et al (2023) Dinov2： arXiv preprint arXiv:230407193

Packhäuser K, Folle L, Thamm F, et al (2023) 使用潜在扩散模型生成匿名胸部放射图，用于训练胸部异常分类系统。In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, pp 1-5

Parihar R, Bhat A, Basu A, et al (2024) Balancing act：扩散模型中以分布为导向的除杂。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 6668-6678 页

Parmar G、Kumar Singh K、Zhang R 等人 (2023) 零镜头图像到图像翻译。In：ACM SIGGRAPH 2023 会议论文集，第 1-11 页

Peebles W, Xie S (2023) 利用变压器的可扩展扩散模型。In：IEEE/CVF 计算机视觉国际会议论文集》，第 4195-4205 页

Perez L, Wang J (2017) 深度学习在图像分类中的数据增强效果。arXiv preprint arXiv:171204621

Pinaya WH, Tudosiu PD, Dafflon J, et al (2022) 利用潜在扩散模型生成脑成像。In：MICCAI 深度生成模型研讨会，施普林格出版社，第 117-126 页

Podell D, English Z, Lacey K, et al (2023) Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:230701952

Qiu Z, Liu W, Feng H, et al (2023) 通过正交微调控制文本到图像的扩散。神经信息处理系统进展 36:79320-79362

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision.In：国际机器学习会议，PMLR，第 8748-8763

页

Raffel C, Shazeer N, Roberts A, et al (2020) 用统一的文本到文本转换器探索迁移学习的极限。机器学习研究期刊》21（140）：1-67

Rando J, Paleka D, Lindner D, et al (2022) Red-teaming the stable diffusion safety filter.

Ranftl R, Lasinger K, Hafner D, et al (2020) Towards robust monocular depth estimation：混合数据集实现零镜头跨数据集传输。电气和电子工程师学会模式分析和机器智能交易 44(3):1623-1637

Razzhigaev A, Shakhmatov A , Maltseva A, et al (2023) Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. ArXiv preprint arXiv:231003502

Rombach R, Blattmann A, Lorenz D, et al (2022) 利用潜在扩散模型进行高分辨率图像合成。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 10684-10695 页

Ronneberger O, Fischer P, Brox T (2015) U-net：用于生物医学图像分割的卷积网络。In：医学影像计算与计算机辅助干预-MICCAI 2015：第 18 届国际会议，德国慕尼黑，2015 年 10 月 5-9 日，论文集，第三部分 18，施普林格，第 234-241 页

Rouzrokh P, Khosravi B, Faghani S, et al (2022) Multitask brain tumor inpainting with diffusion models：arXiv preprint arXiv:221012113

Ruiz N, Li Y, Jampani V, et al (2023) Dreambooth：微调文本到图像的扩散模型，实现主题驱动生成。In：IEEE/CVF 计算机视觉与模式识别会议论 文 集 》，第 22500-22510 页

Ruiz N, Li Y, Jampani V, et al (2024) Hyperdreambooth：用于文本到图像模型快速量化的超网络。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 6527-6536 页

Sagers LW, Diao JA, Groh M, et al (2022) Improving dermatology classifiers across populations using images generated by large diffusion models. arXiv preprint arXiv:221113352

Saharia C, Chan W, Chang H, et al (2022a) Palette：图像到图像扩散模型。In：ACM SIGGRAPH 2022 会议论文集，第 1-10 页

Saharia C, Chan W, Saxena S, et al (2022b) Photorealistic text-to-image diffu- sion

models with deep language understanding.神经信息处理系统进展 35:36479-36494

Salimans T, Goodfellow I, Zaremba W, et al (2016) Improved techniques for training gans.神经信息处理系统进展 29

Sanchez P, Tsaftaris SA (2022) Diffusion causal models for counterfactual estimation.

Sanchez P, Kascenas A, Liu X, et al (2022) What is healthy?In：MICCAI 深度生成模型研讨会、

Springer, pp 34-44

Santos R, Silva J, Branco A (2024) Pix2pix-onthefly： ArXiv preprint arXiv:240308004

Sarukkai V, Li L, Ma A, et al (2024) 拼贴扩散。In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 4208-4217 页

Schnell J, Wang J, Qi L, et al (2024) Scribblegen：生成数据增强改进了涂鸦监督语义分割。2311.17121

Schramowski P, Brack M, Deiseroth B, et al (2023) Safe latent diffusion：减轻扩散模型中的不当退化。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 22522-22531 页

Schuhmann C, Vencu R, Beaumont R, et al (2021) Laion-400m：经剪辑过滤的 4 亿图像-文本对的开放数据集。arXiv 预印本 arXiv:211102114

Sheynin S, Ashual O, Polyak A, et al (2022) Knn-diffusion： arXiv preprint arXiv:220402849

Shi J, Xiong W, Lin Z, et al (2024) Instantbooth：无需测试时间微调的个性化文本到图像生成。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 8543-8552 页

Shin C, Kim H, Lee CH, et al (2024) Edit-a-video：具有对象感知一致性的单一视频编辑。In：亚洲机器学习会议，PMLR，第 1215-1230 页

Shorten C, Khoshgoftaar TM (2019) 深度学习图像数据增强调查。大数据期刊》6:1-48。URL https://api.semanticscholar.org/ CorpusID:195811894

Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures.IEEE access 7:53040-53065

Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al (2015) Deep unsupervised learning using nonequilibrium thermodynamics.In：国际机器学习会议，PMLR，第

2256-2265 页

Sohn K, Ruiz N, Lee K, et al (2023) Styledrop：ArXiv 预印本 arXiv:230600983

Song J, Meng C, Ermon S (2020a) Denoising diffusion implicit models.

Song W, Ma W, Zhang M, et al (2024) Lightweight diffusion models: a survey.人工智能评论》57（6）：161

Song Y, Sohl-Dickstein J, Kingma DP, et al (2020b) Score-based generative modeling through stochastic differential equations.

Song Y, Zhang Z, Lin Z, et al (2022) Objectstitch：arXiv preprint arXiv:221200932

at StabilityAI DL (2023) DeepFloyd IF: a novel state-of-the-art open source text-to-image model with the high degree of photorealism and language understanding. https://www.deepfloyd.ai/deepfloyd-if, retrieved on 2023-11-08

Su X, Song J, Meng C, et al (2022) Dual diffusion implicit bridges for image-to-image translation.

Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the inception architecture for computer vision.In：电气和电子工程师学会计算机视觉和模式识别会议论文集》，第 2818-2826 页

Tang C, Wang K, Yang F, et al (2024) Locinv：ArXiv preprint arXiv:240501496

Tewel Y、Gal R、Chechik G 等人（2023 年）用于文本到图像个性化的按键锁定一级编辑。In：ACM SIGGRAPH 2023 会议论文集，第 1-11 页

Trabucco B, Doherty K, Gurinas M, et al (2023) Effective data augmentation with diffusion models.

Tumanyan N, Geyer M, Bagon S, et al (2023) 文本驱动图像到图像翻译的即插即用扩散特征。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 1921-1930 页

Ulhaq A, Akhtar N, Pogrebna G (2022) Efficient diffusion models for vision：arXiv preprint arXiv:221009292

Valvano G、Agostino A、De Magistris G 等人（2024 年）利用稳定扩散对工业数据进行可控图像合成。In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 5354-5363 页

Van Den Oord A, Vinyals O, et al (2017) 神经离散表征学习。神经信息处理系统研究进展 30

Vendrow J, Jain S, Engstrom L, et al (2023) Dataset interfaces： ArXiv preprint

arXiv:230207865

Vinker Y, Voynov A, Cohen-Or D, et al (2023) Concept decomposition for visual exploration and inspiration.ACM 图形学论文集（TOG）42（6）：1-13

Wallace B, Gokul A, Naik N (2023) Edict: Exact diffusion inversion via coupled transformations.In：IEEE/CVF 计算机视觉与图像会议论文集

模式识别》，第 22532-22541 页

Wang Q, Zhang B, Birsak M, et al (2023) Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. arXiv preprint arXiv:230518047

Wang T, Zhang T, Zhang B, et al (2022a) 图像到图像翻译只需预训练。

Wang W, Bao J, Zhou W, et al (2022b) Sindiffusion：从单张自然图像学习扩散模型。arXiv 预印本 arXiv:221112445

Wang Z, Bovik AC, Sheikh HR, et al (2004) Image quality assessment: from error visibility to structural similarity.IEEE 图像处理交易 13(4):600- 612

Wasserman N, Rotstein N, Ganz R, et al (2024) Paint by inpaint： arXiv preprint arXiv:240418212

Wei Y, Zhang Y, Ji Z, et al (2023) Elite：将视觉概念编码为文本嵌入，用于定制文本到图像的生成。In：IEEE/CVF 计算机视觉国际会议论文集》，第 15943-15953 页

Wolleb J, Bieder F, Sandkühler R, et al (2022) Diffusion models for medical anomaly detection.In：医学影像计算和计算机辅助干预国际会议，施普林格出版社，第 35-45 页

Wu JZ, Ge Y, Wang X, et al (2023a) Tune-a-video：用于文本到视频生成的图像扩散模型的一次性调整。In：IEEE/CVF 计算机视觉国际会议论文集》，第 7623-7633 页

Wu Q, Liu Y, Zhao H, et al (2023b) 挖掘文本到图像扩散模型的解缠能力。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 1900-1910 页

Wu R, Wang Y, Shi H, et al (2023c) Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework. arXiv preprint arXiv:230503980

Wu T, Ye S, Chen S, et al (2023d) Detail reinforcement diffusion model： arXiv preprint arXiv:230908097

Xia W, Lyu Q, Wang G (2022) 使用去噪扩散概率模型的低剂量 CT，提速 20 *倍*。

Xiao G, Yin T, Freeman WT, et al (2023) Fastcomposer： ArXiv preprint arXiv:230510431

Xie J, Li W, Li X, et al (2023a) Mosaicfusion：扩散模型作为大词汇量实例分割的数据增强器。arXiv 预印本 arXiv:230913042

Xie S, Zhang Z, Lin Z, et al (2023b) Smartbrush：利用扩散模型进行文字和形状引导的对象涂色。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 22428-22437 页

Xu M, Yoon S, Fuentes A, et al (2023a) 深度学习图像增强技术综述。模式识别 137:109347

Xu M, Yoon S, Fuentes A, et al (2023b) A comprehensive survey of image augmentation techniques for deep learning.模式识别 137:109347. https://doi.org/10.1016/j.patcog.2023.109347, URL http://dx.doi.org/10.1016/j.patcog.2023.109347

Xu S, Ma Z, Huang Y, et al (2024) Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation.神经信息处理系统进展 36

Xue H, Huang Z, Sun Q, et al (2023) Freestyle layout-to-image synthesis.In：Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14256-14266

Yang B, Gu S, Zhang B, et al (2023a) Paint by example：基于范例的扩散模型图像编辑。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 18381-18391 页

Yang L, Zhang Z, Song Y, et al (2023b) Diffusion models：方法与应用综述。ACM Computing Surveys 56(4):1-39

Yang L, Zeng B, Liu J, et al (2024) Editworld：模拟遵循指令的图像编辑世界动态。 arXiv 预印本 arXiv:240514785

Yang S, Xiao W, Zhang M, et al (2022) Image data augmentation for deep learning：arXiv preprint arXiv:220408610

Ye J, Ni H, Jin P, et al (2023) 利用大规模非传统预训练进行合成增强。In：医学影像计算和计算机辅助干预国际会议，施普林格出版社，第 754-764 页

Yin Y, Kaddour J, Zhang X, et al (2023) Ttida: Controllable generative data

augmen- tation via text-to-text and text-to-image models.

Yu H, Luo H, Wang F, et al (2024) Uncovering the text embedding in text-to-image diffusion models.

Yu T, Feng R, Feng R, et al (2023a) Inpaint anything: Segment anything meets image inpainting.

Yu X, Li G, Lou W, et al (2023b) Diffusion-based data augmentation for nuclei image segmentation.In：医学影像计算和计算机辅助介入国际会议，施普林格出版社，第 592-602 页

Yuan J, Pinto F, Davies A, et al (2022) Not just pretty pictures：使用文本到图像生成器实现介入数据增强。arXiv 预印本 arXiv:221211237

Yun S, Han D, Oh SJ, et al (2019) Cutmix：利用可定位特征训练强分类器的正则化策略。In：IEEE/CVF 计算机视觉国际会议论文集》，第 6023-6032 页

Zając M, Deja K, Kuzina A, et al (2023) Exploring continual learning of diffusion models.

Zang Z, Luo H, Wang K, et al (2023) Boosting unsupervised contrastive learning using diffusion-based data augmentation from scratch. arXiv preprint arXiv:230907909

Zeng W, Yan Y, Zhu Q, et al (2024) Infusion：防止定制文本到图像扩散的过度拟合。arXiv 预印本 arXiv:240414007

Zeng Y, Lin Z, Zhang J, et al (2023) Scenecomposer：任意级别语义图像合成。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 22468-22478 页

Zhang G, Wang K, Xu X, et al (2024a) Forget-me-not：在文本到图像扩散模型中学习遗忘。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 1755-1764 页

Zhang H, Cisse M, Dauphin YN, et al (2017) mixup：ArXiv preprint arXiv:171009412

Zhang L, Rao A, Agrawala M (2023a) 为文本到图像扩散模型添加条件控制。In：IEEE/CVF 计算机视觉国际会议论文集》，第 3836-3847 页

Zhang M, Wu J, Ren Y, et al (2023b) Diffusionengine：扩散模型是用于物体检测的可扩展数据引擎。arXiv 预印本 arXiv:230903893

Zhang R, Isola P, Efros AA, et al (2018) 深度特征作为感知度量的不合理有效性。In：IEEE 计算机视觉与模式识别会议论文集》，第 586-595 页

Zhang Y, Zhou D, Hooi B, et al (2022) Expanding small-scale datasets with guided

imagination.

Zhang Y, Dong W, Tang F, et al (2023c) Prospect：扩散模型属性感知个性化的提示谱。ACM 图形学论文集（TOG）42（6）：1-14

Zhang Y, Huang N, Tang F, et al (2023d) 基于差异模型的反演式风格转移。In：IEEE/CVF 计算机视觉与模式识别会议论文集，第 10146-10156 页

Zhang Z, Han L, Ghosh A, et al (2023e) Sine：使用文本到图像扩散模型进行单张图像编辑。In：IEEE/CVF 计算机视觉与模式识别会议论文集》，第 6027-6037 页

Zhang Z, Lin M, Ji R (2024b) Objectadd：通过免训练扩散修改方式在图像中添加对象。 arXiv 预印本 arXiv:240417230

Zhang Z, Zheng J, Fang Z, et al (2024c) 通过去除图像信息进行文本到图像编辑。In：IEEE/CVF 计算机视觉应用冬季会议论文集》，第 5232-5241 页

Zhao H, Sheng D, Bao J, et al (2022) X-paste：重新审视使用剪辑和滞后扩散进行实例分割的可扩展复制粘贴。arXiv 预印本 arXiv:221203863

Zhong J, Guo X, Dong J, et al (2024) Diffusion tuning： arXiv preprint arXiv:240600773

Zhou Y, Sahak H, Ba J (2023a) Training on thin air：利用生成数据改进图像分类。arXiv preprint arXiv:230515316

Zhou Y, Sahak H, Ba J (2023b) 使用合成数据进行数据扩增以提高分类准确性

Zhou Y, Guo C, Wang X, et al (2024) A survey on data augmentation in large model era.

Zhu J, Ma H, Chen J, et al (2023) Domainstudio：利用有限数据微调领域驱动图像生成的扩散模型