联系电话：18613801095
联系邮件：qimao@cuc.edu.cn

基于扩散模型的图像编辑研究现状

毛琪[1*]，方镇[1]，陈澜[1]，陈浩坤[1]
（1.中国传媒大学信息与通信工程学院，北京 100024）

摘要：随着扩散模型（Diffusion models）的提出与迅速发展，依托其高度可解释的数学特性及高质量和多样性的结果，逐渐打破对抗生成网络(Generative Adversarial Network, GANs)在图像生成和图像编辑领域的垄断地位，基于扩散模型的图像编辑逐渐成为计算机视觉领域的研究热点。本文首先介绍了图像编辑的任务定义和扩散模型的基本原理；然后重点分类依次介绍了基于扩散模型的图像编辑技术的发展历程；总结了图像编辑领域常用的评价指标和数据集，同时定性定量比较了经典方法在不同数据集上的效果；最后对基于扩散模型的图像编辑现状进行总结和展望。

关键词：图像编辑；计算机视觉；扩散模型
中图分类号：TP391.4　　**文献标识码：**A

An overview of image editing based on diffusion models

Qi Mao[1*], Zhen Fang[1] , Lan Chen[1], Haokun Chen[1]
（ 1.Communication University of China, Beijing 100024, China）

Abstract: With the introduction and rapid development of diffusion models, these frameworks have begun to challenge the dominance of Generative Adversarial Networks (GANs) in the realms of image generation and editing, thanks to their highly interpretable mathematical properties and the high quality and diversity of their outputs. Image editing based on diffusion models is emerging as a research hotspot in the field of computer vision. This paper first introduces the task definition of image editing and the basic principles of diffusion models. It then categorizes and details the developmental trajectory of image editing techniques based on diffusion models. Furthermore, the paper reviews common evaluation metrics and datasets used in the image editing domain, and provides both qualitative and quantitative comparisons of classical methods across various datasets. Finally, it summarizes the current state and prospects of image editing based on diffusion models.

Keywords: image editing; Computer Vision; diffusion model

基金项目： 国家自然科学青年基金项目（62201522）；国家重点研发计划子课题（2022YFF0902402）
作者简介(*为通讯作者)：毛琪(1995-)，女，博士，副教授，从事图像 / 视频生成方向研究。Email：qimao@cuc.edu.cn；
方镇（2003-），男，本科生，从事图像生成方向研究。Email: faziizhen@cuc.edu.cn；陈澜（2001-），女，本科生，从事图像 / 视频生成方向研究。Email: eva_cl@cuc.edu.cn；陈浩坤（2004-），男，本科生，从事图像 / 视频生成方向研究。Email: chenhaokun@cuc.edu.cn

1 引言

你是否曾幻想过现实生活中有这样一支神奇的画笔，只需肆意一抹便有：光色流转，视角腾挪，生灵变换，斗转星移……得益于计算机视觉的不断发展，这种“神奇画笔”已经慢慢由幻想照进现实----图像编辑技术应运而生并不断发展成熟。图像编辑指的是给定一张图像与一些额外的条件如编辑指令、参考图像、掩码、点拖动操作等，期望得到一张图像在原图的基础上又体现出额外条件中的信息。而随着扩散模型的提出与迅速发展[1-3]错误!未定义书签。，图像编辑又迎来一波研究热潮，涌现出许多不同的图像编辑技术[4-6]。如图 1 所示，这些图像编辑技术极大丰富了人们的日常生活，并仍具有巨大的应用潜力：从简单的颜色、纹理编辑到更为复杂的形状、非刚性编辑、风格编辑、物体移除等。

除了将图像编辑技术如上述根据应用进行分类外，还可以根据输入的不同将图像编辑技术分为基于文本（指令和提示）的方法、基于图像（掩码和参考图像）的方法、基于点拖动操作的方法、基于布局的方法等。这些输入并不是完全独立的，很多方法支持组合输入，这可以帮助实现更精确复杂的图像编辑操作。由于扩散模型本身的特性，主流使用的

图 1 图像编辑应用举例

扩散模型是预训练大模型，生成图像只需进行推理。根据这一过程，可以将这些方法分为需要额外训练网络、需要微调模型和无需训练（微调）的方法。

尽管当前研究在图像编辑领域已取得显著进展，但作为人工智能生成内容(artificial intelligence generated content，AIGC)领域的一个新兴议题，图像编辑仍面临众多尚未解决的挑战[7]。这些挑战包括如何在更低的计算资源消耗下，缩短处理时间，同时实现更优的编辑效果；如何扩展图像编辑的应用范围，利用多种模态的输入执行更精确、更复杂的编辑操作；以及如何建立一个统一和规范的图像编辑量化评价框架。目前，针对这些问题的研究仍处于持续探索阶段。

为了深入挖掘模型的潜力并探索未来可能的改进方向，本文对基于扩散模型的图像编辑方法的研究现历程及现状进行综述。首先在第 2 节对扩散模型的原理进行简要介绍，接着在第 3 节，会根据需要额外训练的方法，以及无需训练的方法，依次地介绍基于扩散模型的图像编辑算法研究现状。然后在第 4 节介绍了在图像编辑任务中常用的数据集和评价指标。其次在第 5 节对一些经典模型进行了定量和定性的对比评估。最后在第 6 总结了这一领域模型的发展历程，并对未来的研究方向提出了深入的思考和展望。

2 扩散模型简介

扩散模型是实现从噪声到数据样本的转换的模型，其基本思想是先通过正向扩散过程来系统地扰动数据中的分布，然后通过学习反向扩散过程恢复数据的分布。扩散模型彻底改变了图像生成的格局，打破了对抗生成网络[8]对于图像生成和编辑领域的垄断地位。

扩散模型提出，存在一系列的噪声将输入图片变为纯高斯噪声，而模型能够借助已知的添加噪声过程，去逆向学习去除噪声生成图片的过程。扩散模型包含两个过程：前向过程及反向过程，两者都是一个参数化的马尔可夫链。前向过程又称为扩散过程，即对真实图片通过数次添加噪声的过程，最终得到纯高斯噪声图片，如图 1 所示。

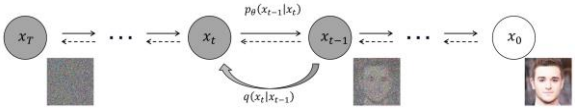


图 2 扩散过程示意图

扩散过程是对数据逐渐增加高斯噪声直至数据变成随机噪声的过程。由$t-1$时刻的结果计算t时刻的结果，随着t的增大，图片将会越来越接近纯噪声。对原始数据，扩散过程的每一步都是对上一步得到的数据按如下方式增加高斯噪声：

$$q(x_{1:T}\mid x_0)=\prod_{t=1}^{T}q(x_t\mid x_{t-1}) \quad (1)$$

$$q(x_t\mid x_{t-1})=N(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I) \quad (2)$$

x_t表示时间步数t的数据，$x_{1:T}$表示从时间步数1到时间步数T的数据序列，β_t表示时间步数t的噪声系数，控制噪声的大小。而反向过程是一个去噪过程，即扩散过程的逆过程，通过近似后验分布从高斯噪声中恢复图像。在此过程中，训练一个神经网络来拟合这个分布，这里采用 U-net 网络结构。将反向过程定义为一个马尔可夫链，且由一系列用神经网络参数化的高斯分布组成：

$$p_\theta(x_{0:T})=p(x_T)\prod_{t=1}^{T}p_\theta(x_{t-1}\mid x_t) \quad (3)$$

$$p_\theta(x_{t-1}\mid x_t)=N(x_{t-1};\mu_\theta(x_t,t),\sum_\theta(x_t,t)) \quad (4)$$

μ_θ表示均值。由此可明确，Diffusion 模型的训练过程可以视为：获取输入图片，从 1 至T时间中采样一个t；从标准高斯分布中采样一个噪声；最终最小化该噪声与高斯噪声构成的损失函数。

Denoising Diffusion Implicit Models[3] (DDIM)在 Denoising Diffusion Probabilistic Models (DDPM)的基础上不再限制扩散过程必须是马尔可夫链，这使得模型可以使用更少的采样步骤来加速生成过程。DDIM 的这一特点使其在处理大规模图像数据时具有显著的速度优势。Dhariwal 和 Nichol[9]在其研究中通过实验和架构优化，展示了扩散模型在样本质量和生成速度方面超越生成对抗网络 Generative Adversarial Networks[8] (GANs)，证明了扩散模型在图像生成领域的适配性。Latent Diffusion Models[1] (LDM)通过将扩散过程应用于预训练的自编码器的潜在空间，而不是直接在像素空间中进行，从而提高训练效率和推理效率。同时，LDM 通过引入交叉注意力机制(Cross-Attention)，使得扩散模型能够处理文本或边界框等一般条件输入，并以卷积方式进行高分辨率合成。

Stable Diffusion[1] (SDM)是一种基于 Laion-5B[10]数据集训练的潜在扩散模型(LDMs)，在基于文本生成图像(Text-to-Image，T2I)领域取得了显著成就。SDM 通过冻结的 Contrastive Language-Image Pre-training Model[11] (CLIP) ViT-L/14 文本编码器处理文本输入，能够根据简短的文本描述生成高质量、逼真的图像，极大地简化和优化了图像创作流程。Laion-5B 数据集有 58 亿对图像-文本配对，覆盖多种语言和高分辨率图像，为模型提供了丰富的训练素材，从而显著提升了图像生成的质量和准确性。

SDM 的高效设计使其能在较低的硬件要求下运行，如仅需 10GB 显存的 GPU，使其成为广受欢迎的开源文生图模型。该模型已广泛应用于艺术创作、广告设计及游戏开发等领域，展示了其在多个创意行业中的广泛适用性和强大潜力。

如图 3(a)所示，SDM 包含三个模块：自编码器、条件编码器和去噪 U-net。SDM 首先使用自编码器对图像编码操作，将图像映射到潜在空间中，变成 64*64 分辨率的潜在噪声特征z_0。接着对z_0进行前向扩散过程得到z_t。接着使用去噪 U-net 对z_t进行采样操作。这里的z_t是经过编码的潜在噪声特征，前文中的x_t是未经过编码的噪声图像，后文使用z_t统一指代噪声图像和潜在噪声特征。在每一个采样步中，条件信息通过条件编码器被传输到 U-net 中，结合交叉注意力引导去噪。对于文本条件，使用 CLIP 的预训练模型作为文本编码器。在采样过程结束后使用解码器将图像从潜在空间映射回像素空间。将扩散过程从像素空间映射到潜在空间的过程称为感知压缩，它加速了扩散模型并降低对计算资源的要求。其中交叉注意力涉及到条件信息的注入，是一系列编辑方法的关键。

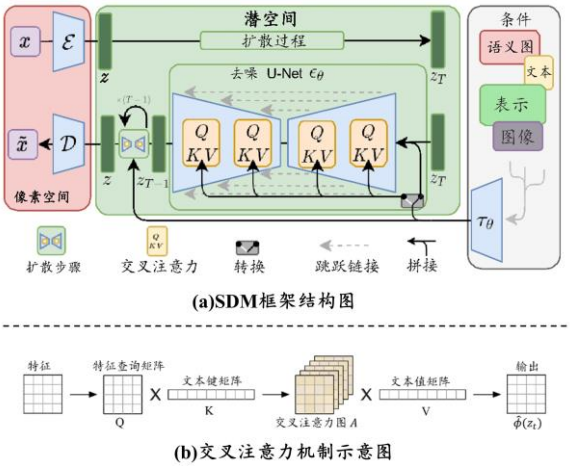


图 3 SDM[1]框架结构图与交叉注意力机制示意图

交叉注意力(Cross Attention, CA)来源于 Transformer，指的是指编码器和解码器之间的注意力层。图 3(b)所示查询矩阵来自编码器，键矩阵和值矩阵来自解码器。在扩散模型中，查询矩阵来自

潜在噪声特征经过下采样后的深层空间特征，键矩阵和值矩阵来自文本嵌入。通过可学习的线性投影函数 ℓ_Q，ℓ_K 和 ℓ_V，将潜在噪声特征 $\phi(z_t)$ 的深层空间特征映射到查询矩阵 $Q=\ell_Q(\phi(z_t))$，将文本嵌入被映射到键矩阵 $K=\ell_K(\phi(P))$ 和值矩阵 $V=\ell_V(\phi(P))$。接着使用以下公式计算交叉注意力图：

$$A=\text{Softmax}(\frac{QK^T}{\sqrt{d}}) \quad (5)$$

其中 d 是键矩阵和查询矩阵的潜在投影维度。最后将上一步得到的交叉注意力图 A 与值矩阵 V 相乘得到交叉注意力图的输出。自注意力机制的过程与其类似，不同的是自注意力的键矩阵和值矩阵都来自潜在噪声特征的深层空间特征。

3 基于扩散模型的图像编辑算法研究现状

回顾扩散模型的发展历程，其快速发展完善离不开富有创造性的社区环境。基于扩散模型的图像编辑是扩散模型的一个子任务，在电影、游戏、绘画和虚拟现实等领域都有重大应用潜力。图像编辑任务涉及在给定一张原始图像及相关的编辑指令、参考图像或交互式操作（如点拖动）的情况下，生成一张既符合编辑要求又保留原图特征的目标图像。这一任务的核心挑战在于如何有效地平衡编辑效果与对原始内容的忠实度。在实施图像编辑时，必须细致地调整和优化处理流程，以确保最终图像既满足用户的具体编辑需求，同时又不损害原图的核心视觉属性。如图 1(a) 中将长椅颜色编辑为红色的关键在于确保编辑内容的充分准确性和稳定性，以避免编辑失效。此外，必须精确地保留长椅的原始特征，并确保背景及其他不需编辑的部分与原图保持高度一致性。这种操作不仅要求对颜色的精确调整，还需要维护图像的整体视觉和结构完整性。本文将图像编辑任务细分为纹理编辑、形状改变、风格变化、物体消除、非刚性编辑以及物体添加等六类。表 1 依据不同方法所需的输入类型及其实现的具体编辑任务，对基于扩散模型的经典图像编辑技术进行了系统的归纳与总结。

在明确了图像编辑的定义之后，本文将根据是否需要对网络进行额外训练，系统地概述各种图像编辑技术。

3.1 需要额外训练的方法

需要额外训练的方法指的是通过大规模数据集重新训练扩散模型，这种方法在数据分布建模上更为全面，从而在一些编辑任务中表现出更稳定的效果。基于算法所能够成功编辑的图像域数量，可以将这些方法划分为特殊域方法和通用场景两大类，如图 4 **需要额外训练的方法的发展脉络**所示。从特殊域到通用场景的演变过程中，这些前期探索充分展示了基于扩散模型的图像编辑方法所蕴含的巨大潜力，并为后续研究提供了宝贵的经验借鉴。

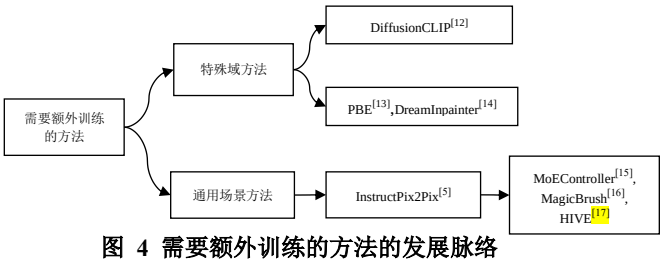


图 4 需要额外训练的方法的发展脉络

3.1.1 特殊域方法

特殊域方法关注于解决某一特定类别的编辑任务，例如将狗转化为猫、将人的表情编辑为笑脸等。这类方法与传统的计算机视觉图像翻译任务相似。在 SDM 中，CLIP 预训练模型被用作文本编码器，以确保编辑后的图像在 CLIP 空间中与目标编辑文本的向量表示具有较高相似性。受 GAN Inversion 方法的启发，DiffusionCLIP[12]提出了一个新的损失函数，该函数旨在确保文本编辑方向与图像编辑方向的一致性:

$$\mathrm{L}_{\text{direction}}(x_{\text{gen}},y_{\text{tar}};x_{\text{ref}},y_{\text{ref}}):=1-\frac{\langle \Delta I,\Delta T\rangle}{\|\Delta I\|\|\Delta T\|} \quad (6)$$

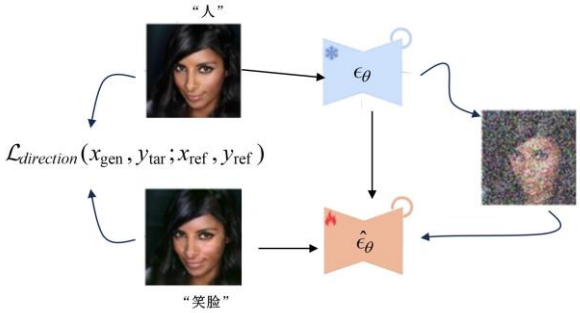
其中 x_{gen} 为生成的图像，y_{tar} 为目标文本描述，x_{ref} 为参考图像，y_{ref} 为参考文本描述。其中 $\Delta T=E_T(y_{\text{tar}})-E_T(y_{\text{ref}}),\Delta I=E_I(x_{\text{gen}})-E_I(x_{\text{ref}})$，$E_T$ 是 CLIP 模型的文本编码器，将文本描述转换为 CLIP 空间中的向量表示。E_I 是 CLIP 模型的图像编码器，将图像转换为 CLIP 空间中的向量表示。ΔT 即为目标文本描述和参考文本描述在 CLIP 空间中的向量差异。ΔI 即为生成图像和参考图像在 CLIP 空间中的向量差异。

为了实现这一点，该方法首先利用 DDIM 反演获取潜在的噪声特征，然后利用上述损失函数重新训练扩散模型。训练流程与框架设计如图 5 所示。尽管 DiffusionCLIP 在人脸相关任务上取得了显著

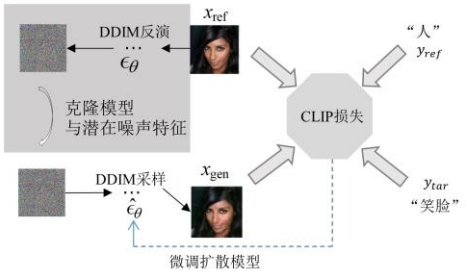
表 1 基于扩散模型的图像编辑经典方法概览

分类	模型	输入	纹理编辑	形状改变	风格变化	物体消除	非刚性编辑	物体添加
需要额外训练	DiffusionCLIP[12]	文本，类别	√		√			
	PBE[13]	参考图像						√
	DreamInpainter[14]	参考图像，掩码，文本				√	√	√
	InstructPix2Pix[5]	文本	√	√	√		√	√
	MoEController[15]	文本	√		√			√
	MagicBrush[16]	文本	√	√	√	√		√
	HIVE[17]	文本	√		√	√		√
需要微调	Null-text Inversion[18]	文本	√		√		√	
	DPL[19]	文本						
	DDS[20]	文本	√		√		√	√
	Imagic[6]	文本	√				√	√
	SINE[21]	文本	√		√			√
	StyleDiffusion[22]	参考图像，文本			√			√
	DragDiffusion[23]	点拖动操作		√			√	
	MasaCtrl[24]	文本，布局草图					√	
	DragNoise[25]	点拖动操作		√			√	
无需微调	Prompt-to-Prompt[4]	文本	√		√			√
	Plug-and-Play[26]	文本			√			
	TIC-inversion[27]	文本，掩码					√	
	BARET[28]	文本	√	√	√		√	
	TI-Guided-Edit[29]	文本，参考图像	√	√			√	
	TF-ICON[30]	文本，参考图像，掩码						√
	MAG[31]	文本，掩码	√	√	√	√	√	√
	CDS[32]	文本	√					√
	KV Inversion[33]	文本					√	
	PnP Inversion[34]	文本	√	√	√		√	
	Edit Friendly DDPM[35]	文本	√	√	√		√	
	Disentanglement Diffusion[36]	文本	√		√			√
	Null-Text Guidance[37]	参考图像，文本			√			
	Negative Inversion[38]	文本	√		√	√		√
	ProxEdit[39]	文本	√		√	√		√
	BD[40]	文本，掩码	√			√		√
	BLD[41]	文本，掩码	√					√
	PFB-Diff[42]	文本，掩码	√					√
	DiffEdit[43]	文本	√		√	√		√
	Watch Your Step[44]	文本	√		√			√
	MFL[45]	文本	√			√		
	Instruct-Edit[46]	文本			√			√
	Inpaint Anything[47]	文本，掩码						

成效，但每种编辑任务都需要单独训练，这增加了计算和时间成本。



(a)DiffusionCLIP训练流程简化图



(b)DiffusionCLIP框架图

图 5 DiffusionCLIP[12]模型示意图

然而，DiffusionCLIP 主要局限于基于文本的编辑任务，而实际应用中，基于参考图像的编辑、消除物体编辑等需求同样重要。为此，Paint by Example[13] (PBE)进一步扩展了图像编辑的范围，它不仅限于基于文本的编辑。PBE 利用参考图像边界框内的内容作为参考，以自监督的方式进行训练，同时将边界框外的内容作为原图像。为了避免简单的图像复制粘贴，并增强模型对上下文的理解，PBE 基于边界框创建了一个任意形状的掩码，并使用 CLIP 图像编码器将参考图像的信息压缩为扩散模型的条件。

为了保留更多的低级细节，DreamInpainter[14]则关注于 U-net 的下采样网络特征提取。在训练过程中，它向整个图像添加噪声，要求扩散模型在详细的文本描述指导下恢复清晰的图像。这些创新方法不仅丰富了图像编辑的多样性，也为后续研究提供了更广阔的探索空间。

3.1.2 通用场景方法

通用场景方法通过单次训练能够实现通用场景下的多类编辑任务，如更改动物种类、更改衬衫颜色或调整碗的形状等。相较于特殊域方法，通用场景方法显著减少了计算资源和时间的消耗，实现了更高效的任务处理。其中，InstructPix2Pix[5] (IP2P)是一个代表性方法，它结合了先进的 Generative Pre-trained Transformer 3[48] (GPT3)自然语言大模型，构

图 6 InstructPix2Pix[5]框架结构图与结果对比图

建了一个大型图像编辑示例数据集。在此基础上，训练出了一个条件扩散模型，只需输入图像和编辑指令，即可生成相应的编辑后图像。这种方法避免了重复训练，实现了在统一框架内完成多样化编辑任务的目标。

IP2P 的特点在于其编辑指令的简洁性和高效性。用户只需输入如“让它变成黑色”这样的指令，即可实现编辑，无需输入与编辑对象无关的文本信息，降低了使用门槛。如**错误!未找到引用源。**(a)所示，其训练过程分为两个阶段：生成数据集和训练扩散模型。Generative Pre-trained Transformer[49] (GPT)生成的编辑文本对与 SDM 结合 Prompt-to-Prompt[4] (P2P)生成的图像对相匹配，形成训练数据。在 U-net 网络中进行完全监督的训练，确保了模型对输入图像和编辑指令的准确响应。然而这样的指令编辑也带来了一系列问题，如**错误!未找到引用源。**(b)所示，编辑指令的轻微区别会给编辑结果带来显著影响。

在 IP2P 的基础上，后续方法进行了进一步改进。MoEController[15] 引入了 Mixture-of-Expert Controllers (MOE)架构，包含三个专家，分别负责细粒度局部翻译、全局风格迁移和复杂局部编辑任务，从而提升了编辑效果。MagicBrush[16]则通过聘请 Amazon Mechanical Turk(AMT, https://www.mturk.com/)的工人手动执行编辑，构建全新数据集，并基于此数据集重新训练 IP2P，增强了模型的实用性。Harnessing human feedback for Instructional Visual Editing[17] (HIVE)则为了与人类偏好相一致，引入了强化学习，在提出的新数据集上训练了奖励模型，确保了编辑结果与人类审美的一致性。这些思路在评价指标中也得到了体现，为图像编辑领域的发展提供新方向。

3.2 无需训练的方法

需要额外训练网络的方法往往伴随着巨大的时间和计算资源消耗。然而，扩散模型本身具备丰富的信息表示能力，包括交叉注意力、自注意力和潜在噪声特征等。一些方法巧妙地利用了这些特性进行模型微调或其他操作，从而在编辑效果和计算成本之间达到更理想的平衡。根据是否需要对扩散模型进行微调，本节将这些方法分为两类：需要微调的方法和无需微调的方法。

3.2.1 需要微调模型的方法

需要微调的方法通常涉及利用 SDM 模型的丰富内部表达能力，或者通过引入外部网络，构建特定的损失函数来对扩散模型的 U-net 部分进行微调。值得注意的是，也有一些方法专注于优化噪声、潜在噪声特征或文本嵌入。在本文中，我们将那些需要对扩散模型的 U-net 部分进行微调的方法定义为需要微调的方法，而将其他方法统称为无需微调的方法。如图 7 所示**图 7 需要微调的方法的发展脉络**，根据是否引入外部的网络参与模型的微调，本节将需要微调模型的方法分为基于扩散模型的微调和基于外部网络的微调。

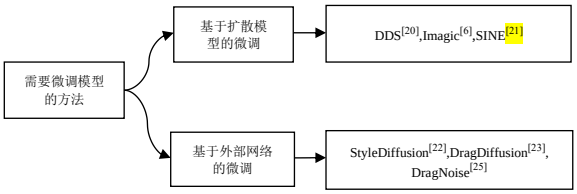


图 7 需要微调的方法的发展脉络

基于扩散模型的微调 基于扩散模型的微调尝试根据 U-NET 网络中的特定输出，如自注意力和交叉注意力等构建损失函数，从而微调去噪 U-net 网络以调整潜在噪声特征。Delta denoising score[20] (DDS) 利用分数蒸馏采样 Score Distillation Sampling[50] (SDS)机制进行图像编辑，利用两个图像-文本对:一个原图像和原文本，另一个目标对象和目标文本。DDS 计算这两个图像-文本对之间的差异，通过这种比较得出损失，通过这个损失优化潜在噪声特征。Imagic[6]使用两种策略共同完成图像编辑任务：文本嵌入优化和扩散模型微调。首先最小化重建图像和原始图像之间的距离来优化目标文本嵌入。同时，对扩散模型进行微调，以便更好地进行上一步得到的优化文本嵌入引导图像重建。最后将原始文本嵌入和优化后的文本嵌入进行插值，利用该插值引导微调后的扩散模型生成编辑图像。Single image editing with text-to-image diffusion models[21] (SINE)对文本编码器和扩散模型进行微调。它提出了一种基于补丁的微调策略，可以有效地帮助模型生成任意分辨率的图像，这使得基于扩散模型的图像编辑在实际生活中的应用潜力更进一步加强。

基于外部网络的微调 基于外部网络的微调通过引入新的网络构建损失函数从而微调扩散模型或微调网络参数，从而引导编辑。这部分方法借鉴了先前研究的经验和成果，取得了良好的编辑效果。StyleDiffusion[22]使用映射网络将输入图像的特征映射到与文本提示嵌入空间对齐的嵌入空间，结合 SDM 本身的交叉注意力机制有效地生成提示嵌入并根据反演和重建过程中对应的潜在噪声特征和注意力构建损失函数优化这个映射网络的参数，进而通过简单的文本描述即可实现良好的重建和有效、无溢出的编辑操作。

在基于外部网络的微调方法中有一类十分有趣的方法:基于点拖动操作的方法。这类方法的代表性工作是 DragGAN[51]。它使用点拖动操作代替文本在非刚性编辑任务中作为一种条件信息输入。它不仅使得精确定量的图像编辑成为可能，同时大大降低了图像编辑的使用门槛。

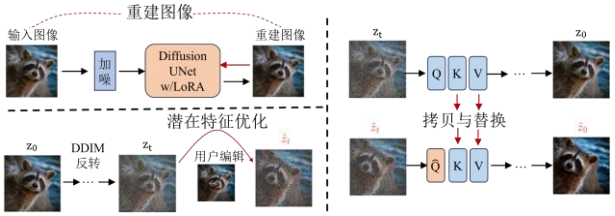


图 8 DragDiffusion[23]模型框架

DragDiffusion[23]受到 DragGAN 的启发，如图 8 所示，首先使用 Low-rank adaptation of large language models[52] (LORA)微调扩散模型保留身体特征，接着基于用户提供的点拖动操作信息提出了运动监督和点偏移损失优化潜在噪声特征，最后借鉴 MasaCtrl[24]中的相互自注意力进行身份特征的进一步保留，提升编辑质量。DragDiffusion 在一些非刚性编辑任务中展现出了良好的性能，并且由于点拖动操作降低了普通用户的使用门槛，有很强大的应用潜力。但是 DragDiffusion 直接更新整个潜在噪声特征容易造成非忠实的编辑和梯度消失，导致编辑失败。针对这一问题，DragNoise[25]将优化的对象从潜在噪声特征转向 U-net 的特定层从而控制单步去

噪 U-net 的噪声，减少了优化时间的同时增强了编辑的稳定性和效果。

3.2.2 无需微调模型的方法

需要微调的方法虽然在一些任务上可以达到较好的效果，但是由于其对计算资源的占用和微调模型可能产生的过拟合，无需微调的方法应运而生。无需微调模型的方法指的是在每一张图像的推理过程中对模型部分进行优化的方法，区别与需要微调模型的方法，该类方法无需微调扩散模型并且耗时更短，同时该类方法包含对潜在噪音特征、注意力等中间特征进行优化而不涉及对 U-net 进行直接微调的方法。如图 9 所示，本节将围绕基于注意力注入的方法、基于反演的方法、基于掩码的方法对该类无需微调模型的图像编辑方法展开介绍。

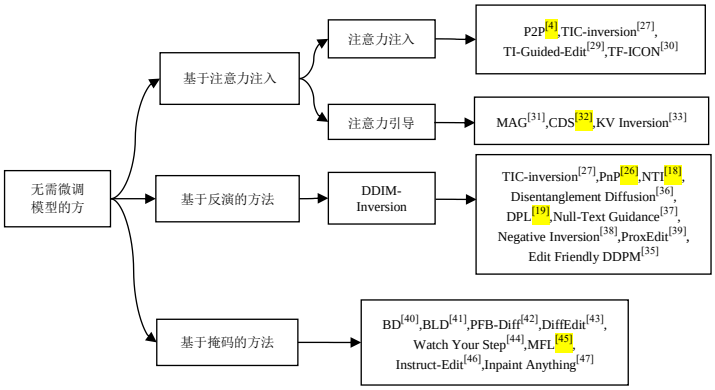


图 9 无需微调模型方法的发展脉络

基于注意力的方法 基于注意力的方法重点分析了去噪 U-net 中注意力机制对输出图像的重要作用，并通过对注意力层进行各种操作来实现图像编辑功能。这些方法利用注意力机制的细节调整，从而精确地控制图像生成过程中的特定区域和细节。具体而言，这些方法可以根据操作注意力的方式不同分为两大类：注意力注入和注意力引导。接下来，本文将详细介绍这两类基于注意力的方法，探讨它们如何通过调整或引导注意力层来优化图像编辑的效果。

P2P 开创性地提出了一种只需要编辑文本的图像编辑方法。具体来说，它使用重建分支和编辑分支共同完成图像编辑任务，根据编辑文本相较原文本的不同操作（换词、细化和调整权重）将重建分支的交叉注意力图和自注意力图替换、插入编辑分支的对应的注意力图或将编辑分支的交叉注意力图乘以不同的权重。P2P 启迪了许多基于注意力注入的方法。不同于 P2P 替换的是注意力图, Plug-and-Play [26] (PnP)通过直接将重建分支的空间特征和自注意力的查询矩阵和键矩阵插入到编辑分支实现更细粒度的编辑效果，而 MasaCtrl 提出了相互自注意力，即替换自注意力的键矩阵和值矩阵，同时，掩码引导的相互自注意力策略通过解决前景和背景混淆进一步增强了一致性,这使得 MasaCtrl 的非刚性编辑能力相较于 P2P 得到加强，如动作改变、数量改变、视角变换等，拓宽了图像编辑的潜能。实践证明 MasaCtrl 提出的相互自注意力策略可以作为一种增强编辑忠实性的手段，广泛应用于其他编辑方法中。相似地，TIC-Inversion[27]直接使用 DDIM 反演分支的自注意力的键矩阵和值矩阵而不是来自重建分支，从而更好直接利用原图的信息，在非刚性编辑中更好地平衡了保真度和编辑性。而 Balanced Attention Based Real Image Editing Driven by Target-Text Inversion[28] (BARET)提出了注意力平衡模块，重新组合重建分支、DDIM 采样分支和过度分支的交叉注意力和自注意力，更好地平衡了编辑效果和内容保留。

前文所述的方法直观且有效，但对于一些复杂的任务：如提供参考图像，使得原图像的某部分的纹理风格和参考图像相似等任务则束手无策。TI-Guided-Edit[29]提出了一种既可以完成上述使用参考图像进行编辑任务，又可以完成常见刚性编辑和非刚性编辑任务的框架，该框架通过组合原图像和参考图像的自注意力的查询、键和值并进一步借鉴工作[53]中的注意力对比方法来强化编辑过程中的结构匹配同时消除伪影。TF-ICON[30]则专注于跨域图像的合成，通过重组自注意力使得参考图像更自然协调地融入原图像。它引入额外提示词确保图像自注意力的精确反演，以此来提升编辑效果。

总体而言，注意力注入方法通过更加深入地挖掘扩散模型固有的信息，增强了模型的信息处理能力。然而，这种方法的复杂性相应提高，且可能会对模型的泛化能力带来一定的限制。

不同于注意力注入类直接改变图像注意力的方法，注意力引导方法通常会根据注意力构造一个损失函数，从而更新噪声或潜在噪声特征。MAG[31]聚焦于交叉注意力，根据需要编辑的词元和提供的掩码构造了两个损失函数，分别从词元维度和空间维度提高了需要编辑词元的交叉注意力图的比重，并且可以迭代更新潜在噪声特征，从而解决了复杂场景编辑错误定位或效果微弱的问题。Contrastive Denoising Score[32] (CDS)在 DDS 的基础上结合了

Contrastive learning for unpaired image-to-image translation[54] (CUT)的工作，使用 U-net 中的丰富自注意力表示来构建损失函数，增强了 DDS 的编辑性能。KV Inversion[33]则构造重建损失函数学习交叉注意力中的键和值，称为内容保存自注意力，它在实现真实图像动作编辑的有效性和忠实性之间实现了更好的平衡。

此类方法的侧重点在于如何根据注意力等特征构造有效合理的损失函数从而引导采样，依托于损失函数的多样性，此类方法通常十分灵活，应用广泛，但是由于需要计算梯度并更新，计算需求相对较大，同时超参数的存在会导致该类方法编辑性能的不稳定性。

基于反演的方法 基于反演的方法将视角从去噪 U-net 转向初始的反演过程上：即如何根据一张图像获取其对应不同时间步上的潜在图像特征。如图 所示，许多编辑方法使用反演将一张真实图像映射到潜在噪声空间中，接着使用重建和编辑双分支完成编辑任务，这是常用的编辑框架。而反演完成了其中的第一步，因此反演的质量直接影响了编辑结果。由于其加速采样的策略和确定采样的过程使得 DDIM 反演成为很多工作的基准，它根据

$$\frac{z_{t-1}}{\sqrt{\alpha_{t-1}}}=\frac{z_t}{\sqrt{\alpha_t}}+\left(\sqrt{\frac{1}{\alpha_{t-1}}-1}-\sqrt{\frac{1}{\alpha_t}-1}\right)\cdot\dot{o}_t \quad (7)$$

对初始潜在图像特征进行加噪。

而根据这个公式延伸出 DDIM 采样的公式为

$$\frac{z_t^*}{\sqrt{\alpha_t}}=\frac{z_{t-1}^*}{\sqrt{\alpha_{t-1}}}+\left(\sqrt{\frac{1}{\alpha_t}-1}-\sqrt{\frac{1}{\alpha_{t-1}}-1}\right)\cdot\dot{o}_{t-1}^* \quad (8)$$

将相同时刻的上两式作差可得

$$\frac{z_{t-1}-z_{t-1}^*}{\sqrt{\alpha_{t-1}}}=\left(\sqrt{\frac{1}{\alpha_{t-1}}-1}-\sqrt{\frac{1}{\alpha_t}-1}\right)\cdot(\dot{o}_t-\dot{o}_{t-1}^*) \quad (9)$$

其中 z_t 表示时间步数 t 的潜在图像特征，z_t^* 表示经过不同处理或不同路径计算得到的 z_t。α_t 表示时间步数 t 的缩放系数，用于控制噪声水平。$\dot{o}_t$ 表示时间步 t 的噪声项，$\dot{o}_t^*$ 表示通过不同方法计算得到的 $\dot{o}_t$。可以观察到反演和重建过程中潜在噪声特征的差别来源于噪声项的不同。因此本文将基于反演的方法分为对齐噪声项和对齐潜在噪声特征项，并根据该分类依次介绍各方法。

对齐噪声项的方法观察到潜在噪声特征产生差别的根本原因，通过多种手段使得重建过程中的噪声和反演过程对应的噪声逼近。TIC-Inversion 将 DDIM 反演过程中自注意力的键和值替换到编辑过程中，通过注意力机制使得重建的噪声靠近反演的噪声。

对应噪声项的不同导致了潜在噪声特征的不同，直接使得重建过程和反演过程中的潜在噪声特征相一致。其中具有代表性的方法是 PnP Inversion[34]它在 DDIM 反演的基础上提出了一种新颖的架构，将常用的重建-编辑双分支解耦，直接利用反演分支的潜在噪声特征，噪声项的差几乎为零，将重建分支加回到原始 DDIM 反演分支，而保持编辑分支不被影响，从而增强重要信息保留并进一步提升编辑效果。而 Edit Friendly DDPM[35]在重建中利用了反演过程的一系列潜在噪声特征，相较于一般的 DDPM 反演不仅保证了重建的良好性能，而且允许更多的编辑可能。

尝试利用 DDIM 反演噪声信息的同时，也有方法关注到文本嵌入对于噪声的影响。其中空文本微调作为一种代表性的方法，在图像编辑领域取得了显著成效，尤其是空文本反演 Null-text Inversion[18] (NTI)的出现，更是推动了反演方法的发展，使其达到了当前的最优水平。NTI 针对 DDIM 反演过程中经常出现的重建失败问题，通过在单次 DDIM 反演的基础上，为空文本嵌入构建一个损失函数，旨在减少采样与 DDIM 反演之间的距离。这种方法在保

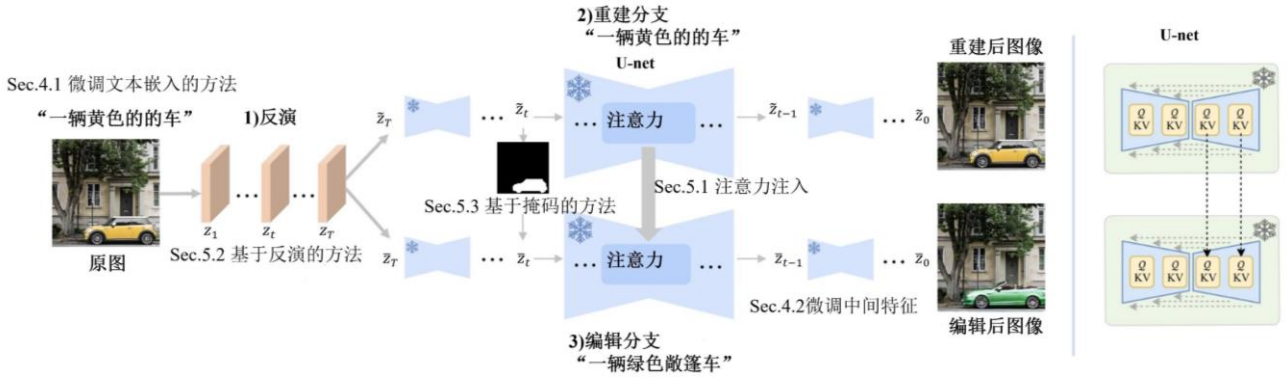


图 10 图像编辑流程图

留用户输入的编辑文本嵌入的同时，有效缩短了采样与反演之间的距离，实现了出色的重建效果。然而，这种优化过程通常需要较长的时间。

此外，NTI 在实际应用中有时会面临名词交叉注意力错误定位的问题。为了解决这一问题，Dynamic Prompt Learning (DPL) [19]方法应运而生。DPL 在 DDIM 反演后，结合掩码技术，提出了三种损失函数，以动态地微调空文本嵌入。这种方法不仅缓解了交叉注意力的泄露问题，还进一步提升了交叉注意力的质量和图像重建的准确性。综合应用这些方法可以更好地利用基于扩散模型的图像编辑技术，实现更精确、更高效的图像编辑操作。除了对文本嵌入直接进行优化，Disentanglement Diffusion [36]则提出了一种不同的策略。观察到 SDM 本身具备解纠缠能力，并据此设计了一种简单轻量的解缠算法。该算法通过优化两个文本嵌入的混合权重，实现样式匹配和内容保持，而无需微调扩散模型，仅需优化大约 50 个参数。BARET 则将 NTI 的优化对象从空文本转移到条件文本上以实现仅由文本引导的快速反演从而进一步完成包括非刚性编辑在内的多种编辑任务。

然而上述方法均涉及到对文本嵌入或其权重的优化，对于时间效率和计算资源拥有一定要求。Null-Text Guidance[37]证明了扰动无分类器引导的空文本可以实现图像的动画化并进一步提出回滚扰动和图像扰动策略。Negative Inversion[38]同样体现了无分类器引导的思想，用原文本替换编辑分支的空文本，借由无分类器引导的公式从反向增强编辑效果，在重建效果和 NTI 的重建效果相近的同时大大加快了反演速度。在此基础上，ProxEdit[39]为了解决负文本反演可能的伪影的问题，同时也针对具有较大无分类器引导权重的 DDIM 反演缺乏理想的重建效果，提出近端指导，通过正则化项和反转引导增强负文本反演的效果，并结合相互自注意力控制以拓展编辑潜能。

无论是对齐噪声项的方法还是对齐潜在噪声特征的方法，基于反演的方法是基于扩散模型的图像编辑算法的基础，直接关系到重建和编辑的结果。通常基于反演的方法可以和其他编辑方法一起使用从而提升编辑效果。

基于掩码的方法 在许多图像编辑场景中，通常只需修改原图像的特定部分，而保持其他区域不变。基于掩码的方法通过在潜在空间使用掩码来重新组合编辑后的图像和原始图像，有效地实现这一目标。这种方法能够精确地保留原图中不需要编辑的重要信息，同时提供更细致的编辑控制。由于其能够支持细粒度的编辑操作，基于掩码的方法在各种实际应用场景中显示出广泛的应用潜力和实用价值。

基于掩码的混合编辑方法通常来说将掩码视为需要编辑和需要保留区域的分界，从反演或者重建过程中提取需要保留区域的信息，从编辑分支中提取需要编辑的信息。其中 Blended Diffusion[40] (BD)就是其中非常经典的一种方法，它通过使用预训练的 CLIP 模型来引导编辑分支中图像的生成，通过掩码将重建分支的噪声特征和编辑分支的噪声特征混合。特别地，它还将噪声经过若干线性变化求得的 CLIP 梯度相加代替直接求 CLIP 梯度，使用该拓展增强优化了编辑效果。在此基础上，Blended latent Diffusion[41] (BLD)通过使用文本到图像的潜在扩散模型，将混合操作的维度从噪声特征转向更低纬度的潜在噪声特征，大大提升编辑速度的同时也避免了每个时间步对 CLIP 梯度的求解。上述两种方法殊途同归，都是在每一步去噪之后的结果上进行混合操作，虽然已经可以实现许多编辑任务如：物体消除、替换、添加等，但是仍有许多更细粒度的任务完成效果欠佳，如改变人的发型等。上述方法掩码内的编辑物体缺少约束，会出现物体形状和掩码不匹配、无缝性差的问题从而进一步导致编辑后图像和文本一致性差、编辑图像质量低下。而 PFB-Diff[42]在 U-net 的采样过程中对更细粒度的特征进行渐进多层次的混合操作，并在交叉注意力的编辑词上通过掩码进行混合，能够实现更细粒度编辑的同时地控制了编辑效果不会溢出。

上述方法将掩码作为用户提供的输入之一，在提升了编辑精确度的同时也带来一个问题，即如何获取能够高效引导编辑的掩码。有一些方法将掩码获取模块划分到编辑框架内部。DiffEdit[43]提出了一种方法，以空文本（或原文本）和目标文本作为条件进行去噪，计算潜在噪声特征上的差异并进行二值化以得到掩码，接着以 DDIM 反演结果作为双分支编辑的起点，在编辑分支中将目标文本作为交叉注意力的查询矩阵，最后利用先前计算的掩码进行混合操作。虽然这种方法在前后景的拓展性上有所欠缺，但是它利用掩码简化语义编辑，保留了重要信息。类似地，Watch Your Step[44]在 IP2P 的基础上，将空文本和编辑指令分别作为IP2P的输入以此获得

噪声差异并进一步获得相关图和掩码。编辑过程中和在IP2P作为编辑分支的基础上在噪声上进行混合操作。DiffEdit 和 Watch Your Step 通过对不同文本嵌入或编辑指令的噪声进行差异处理得到掩码，而 Masactrl 则将获取掩码信息的源头转向去噪过程中的交叉注意力。它从交叉注意力中结合编辑的词元自动提取掩码，用掩码来区分前后景并用编辑前后的词元的对应掩码在交叉注意力的层面上进行混合操作。除了在扩散模型内部使用噪声图或中间特征自动生成掩码，也有方法利用任务迁移的思想进行掩码自动生成。MFL[45]作为一种为文本引导图像编辑量身定制的无掩码方案，就借鉴利用了《Open World Entity Segmentation》[55]中的操作根据原文本提取图像掩码，接着使用该掩码和基于文本条件的扩散模型生成一系列图像，通过多模态质量评估选出编辑效果最好的图像。而 Instruct-Edit[46]则通过 BLIP2[56]和 GPT 根据编辑指令获得的编辑前后的文本，将进一步结合 Grounded Segment Anything[57]获得定位更准确的掩码从而提升了编辑效果。而 Inpaint Anything[47]利用 Segment AnythingModel[58] (SAM)进行掩码生成,利用 LaMa[59]等进行物体移除,最后利用 SDM 进行物体的填充或替换。

基于掩码的方法目前使用掩码的思路较为单一，如何更简单获取更精细的掩码并充分利用掩码信息仍然是基于掩码的编辑方法的未来研究方向。

4 数据集和评价指标

4.1 数据集

数据集的选择是评价模型性能的基础，根据数据集元素和目的的区别，数据集可划分为以下两类：（1）通用图像数据集 此类数据集仅包含图像，不包含编辑相关的信息，根据域间差异可进一步划分为以下两类：场景类和真实对象类，场景类的域间差异小，代表的有 Stanford Car[60]，数据集规模也相对较小，而真实场景域间差异较大，数据集规模也相对较大，代表的有 COCO[61]和 ImageNet[62]。

此类数据集由于缺乏编辑所需要的额外信息如文本描述、编辑指令等，并不是单独为图像编辑任务服务，更多被用于定量比较。（2）针对图像编辑的数据集 此类数据集图像大多取自（1）类，并在（1）类的基础上针对图像编辑任务增加了信息，代表性的数据集是 MagicBrush 和 PIE-Bench[34]。MagicBrush 是第一个大规模的、手动注释的数据集，用于指令引导的真实图像编辑，涵盖多种场景，包

图 10 PIE-Bench[34]数据集示意图

含超过 10,000 个手动注释的三元组（原图像、指令、目标图像），支持训练大规模文本引导图像编辑模型。如图 11 所示，PIE-Bench 覆盖 10 类编辑任务共 700 张图像，为每一张图像提供了原文本、编辑文本、编辑指令和高质量掩码。此类数据集的规模较小，更多被用于定量评估模型。

4.2 指标

图像编辑的评价指标从图像质量、编辑效果以及图像在内容上的保持程度这四个方面来衡量模型的性能。以下介绍 4 种常用指标。

FID Fréchet Inception Distance[63] (FID)以在 ImageNet 数据集上训练的 Inception-V3[64]模型作为特征提取器，计算真实图片和生成图片的特征向量的距离。计算公式如公式（10）所示。

$$FID(g,r) = \| \mu_g - \mu_r \|_2^2 + Tr\left(\sum_g + \sum_r - 2(\sum_g \sum_r)^{\frac{1}{2}} \right) \quad (10)$$

其中，g 表示生成图像，r 表示真实图像，μ 和 $\sum$ 分别表示均值和协方差。当生成图像和真实图像特征的均值和协方差相近时，生成图像的分布接近真实图像的分布，即FID越小，生成的图像质量越好。

CLIP Score CLIP 是一种多模态视觉和语言模型。它可用于图像文本相似性和零样本图像分类。文本和视觉特征都被投影到具有相同维度的 CLIP 潜在空间。将投影图像和文本特征之间的点积称为 CLIP 余弦相似度，其值一百倍定义为 CLIP Score。CLIP 余弦相似度和 CLIP Score 通常被用于衡量编辑图像和编辑文本的匹配程度，分数越高匹配程度越高。除了使用 CLIP 来刻画图像和文本的关系，也可以使用它来刻画两张图像之间的差异程度,它通过将两张图像投射到潜在 CLIP 空间并计算余弦相

似度来评估编辑前后图像的一致性。

SSIM Structural Similarity[65] (SSIM)，即结构相似性，用于评估两幅图像相似度的指标，常用于衡量图像失真前与失真后的相似性，也用于衡量模型生成图像的真实性。

图像编辑是一个较为复杂的任务，不仅需要考虑到编辑的效果，还需要考虑到结构的保持、伪影的严重程度、编辑的方向、编辑任务的种类等指标的协调，因此单一的评价指标往往无法客观公正地评价模型。有一些研究针对图像编辑，期望对各种图像编辑方法进行更客观更综合的定量评价。

以下介绍这里具有代表的 2 种指标：EditVal[66] 和 Human Preference Score[67] (HPS)。

EditVal EditVal 并不是单一的一种指标，而是一种自动评估框架。它包含一系列图像数据集并从 13 种可能的编辑类型中提取的每个图像的一组可编辑属性，使用预先训练的视觉语言模型来评估每个方法每种编辑类型的生成图像的保真度。

HPS 上述评价指标有时会和人类的偏好相冲突，而依赖于 midjourney (https://www.midjourney.com/) 等平台提供的基于扩散模型的图像生成社区，研究者可以从此收集海量的数据并分发问卷，通过模拟人类的偏好对生成的图像与文本的匹配程度进行评分。从 Stable Foundation Discord (https://discord.com/) 频道收集的数据集揭示了人类在图像生成中的选择偏好。实验表明，目前生成模型的评估指标与人类的选择不太相关。因此，它用新数据集训练一个人类偏好分类器，并基于该分类器得出人类偏好评分 Human Preference Score (HPS)。

5 模型比较

为了比较基于扩散模型的图像编辑模型，本小节在 PIE-Bench，Ted-Bench[6]和 MagicBrush 数据集上进行定量和定性比较。

5.1 定性比较

考虑到不同模型的侧重点，本文根据编辑任务的种类不同将编辑任务分为纹理编辑、形状变化、风格变化、物体消除、非刚性编辑、物体添加和点拖动操作编辑。每一种方法使用相同且固定的种子。所有实验均在单张显存为 40GB 的 NVIDIA A100 显卡上进行。

定性地比较了不同模型在各类任务上的表现。得益于庞大的训练数据，IP2P 在多种编辑任务上都有良好的编辑效果，然而出现效果溢出、特征丢失问题。针对非刚性编辑提出的 MasaCtrl 在形状变化和非刚性编辑任务上都取得了比较好的效果并保留了编辑物体的原始身份特征。基于注意力注入的方法如 P2P，PnP 在纹理和风格上都有较好的编辑效果，但 PnP 有时面临编辑效果不明显的问题，而 P2P 也存在效果溢出和身份特征丢失的问题。LEDITS++[68]虽然可以很好完成纹理编辑且没有编辑效果的溢出，但是在风格编辑任务上遭遇了特征的完全丢失。Imagic 在多个任务上都没有表现出令人满意的效果。

基于点编辑的方法面临编辑效果和输入不一致的问题，对于输入的点是复制物体还是移动物体的方法没有明确的区分。

5.2 定量比较

本节从 PIE-Bench 的十个任务中每个随机抽取十张图像作为测试集。定量比较实验结果如下。本节选择 FID,SSIM,CLIP,HPS 分别衡量编辑图像的生成质量、结构保留、编辑效果和整体效果，其中较低的 FID 和较高的 SSIM,CLIP 和 HPS 值标志着更优越的编辑性能。定量比较结果如表 2 所示。

表 2 各模型定量比较评价结果

模型 \ 评价指标	FID	SSIM	CLIP	HPS
I-P2P	**74.23**	0.61	25.86	**22.28**
SINE	<u>164.30</u>	0.51	25.32	20.93
Imagic	152.75	<u>0.45</u>	24.54	<u>20.41</u>
P2P	82.11	**0.64**	**26.43**	22.17
PnP	99.23	0.47	26.27	21.73
MasaCtrl	118.02	0.51	<u>24.19</u>	21.19

在图像生成质量上 IP2P 取得了最好的效果，而微调的方法图像生成质量普遍较差，反映出模型可能出现过拟合的问题。

在结构保留上 P2P 表现较好，Imagic 在结构保留上也存在问题。除了模型本身的问题外，该测试集中包含部分动作变化，这会降低模型在 SSIM 指标上的得分。

在编辑效果上，所有方法都取得了相对不错的效果，MasaCtrl 取得了较低的得分，因为 MasaCtrl

编辑种类：纹理编辑(把车变成黄金的)

Source P2P[4] PnP[26] IP2P[5] SINE[21] Imagic[6] LEDITS++[68]

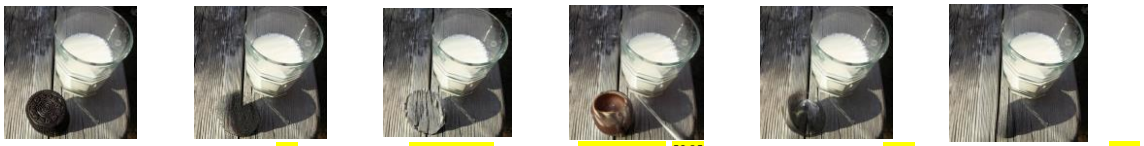
编辑种类：形状变化（把猫变成狐狸）

Source P2P[4] CDS[32] IP2P[5] MasaCtrl[24] Imagic[6] BLD[41]

编辑种类：风格变化（把它变成梵高的画）

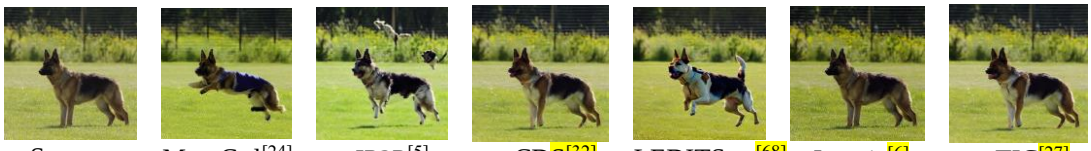
Source P2P[4] IP2P[5] Imagic[6] LEDITS++[68] ProxEdit[39] PnP[26]

编辑种类：物体消除（消除饼干）



Source LD[1] BLD[41] ProxEdit[39] LEDITS++[68] Inpaint Anything[47]

编辑种类：非刚性编辑（让狗跳起来）



Source MasaCtrl[24] IP2P[5] CDS[32] LEDITS++[68] Imagic[6] TIC[27]

编辑种类：物体添加（给雕像添加一副墨镜）

Source MagicBrush[16] IP2P[5] LEDITS++[68] Imagic[6]

编辑种类：基于点拖动的编辑

Source DragDiffusion[23] DragonDiffusion[69] DragNoise[25]

图 11 经典模型分类定性比较结果

是针对非刚性编辑提出的模型，在纹理、风格等任务上表现较差，而测试数据集中包含较多的纹理、风格等编辑任务。

此外，本节还使用 EditVal 进行模型的定量比较，结果如图 13 所示。定量和定性结果表明不同的方法各有侧重。

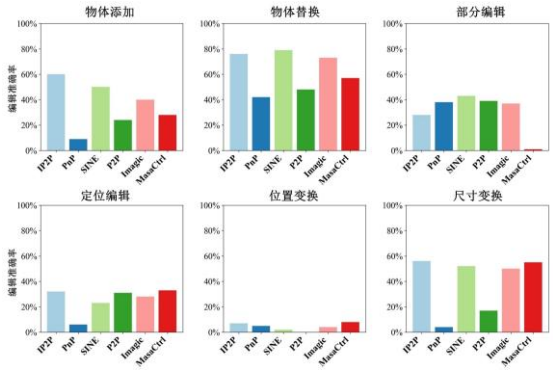


图 12 各模型根据 EditVal[66]进行定量评价评估比较结果

6 总结和展望

基于扩散模型的图像编辑技术已成为图像编辑领域的重要组成部分，其在生成图像质量和编辑效果方面均取得了显著进展。本文首先对图像编辑进行了简要介绍，并系统梳理了基于扩散模型的图像编辑研究历程。同时，本文整理了常见的数据集和评价指标，并对经典模型进行了定量和定性的对比分析。

扩散模型凭借其高质量的条件生成能力、逐步去噪的噪声模型以及逆向过程，展现了出色的可编辑性和可控性，为图像编辑提供了独特的优势。未来，该领域的研究可围绕以下方向展开：

（1）开发用户友好的通用编辑工具至关重要。当前图像编辑领域面临的一大挑战是如何降低编辑工具的使用门槛。尽管基于文本和基于图像、布局的部分编辑方法已经取得了一定进展，但这些方法往往依赖于高质量的输入信息。而随着大语言模型的发展，也有研究者尝试在编辑过程中加入大语言模型的协助以更好的平衡编辑效果和用户友好性。相比之下，基于点拖动操作的方法通过直观的点拖动操作方式替代了模糊的文本描述，显著提升了用户友好性[23, 25, 51]。然而，这种方法目前主要侧重于非刚性编辑。因此，开发兼具易用性和通用性的编辑工具仍具有巨大的研究价值和应用前景。

（2）简化模型结构同样值得探索。现有模型往往结构复杂，训练和微调时间长，且受限于巨大的计算开销，通常只能处理分辨率较低的潜空间表示。如何在保持模型性能的同时简化其结构，是一个亟待解决的问题。这有望推动模型在更高分辨率的图像上实现更高效的编辑操作。

（3）实现更富想象力的编辑也是未来研究的重要方向。当前编辑内容主要集中在以纹理颜色为主的刚性编辑和以形状视角动作为主的非刚性编辑上。然而，实际应用中许多编辑任务需要更丰富的想象力和创造力，如将闭嘴的老虎编辑为张嘴的老虎等。这类编辑不仅涉及动作和形状的变化，还需生成原图中不存在的信息。已有研究工作开始关注这类更具想象力和实际应用价值的编辑操作[70]，预示着未来该领域将涌现出更多富有创意的编辑实现。

参考文献：

[1] Rombach R, Blattmann A, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10684-10695.

[2] Ho J, Jain A, et al. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[3] Song J, Meng C, et al. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.

[4] Hertz A, Mokady R, et al. Prompt-to-prompt image editing with cross attention control[J]. arXiv preprint arXiv:2208.01626, 2022.

[5] Brooks T, Holynski A, et al. Instructpix2pix: learning to follow image editing instructions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18392-18402.

[6] Kawar B, Zada S, et al. Imagic: text-based real image editing with diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6007-6017.

[7] Huang Y, Huang J, et al. Diffusion Model-Based Image Editing: A Survey[J]. arXiv preprint arXiv:2402.17525, 2024.

[8] Goodfellow I, Pouget-Abadie J, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[9] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.

[10] Schuhmann C, Beaumont R, et al. Laion-5b: An open large-scale dataset for training next generation image-text models[J]. Advances in Neural Information Processing Systems, 2022, 35: 25278-25294.

[11] Radford A, Kim JW, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning, 2021: 8748-8763.

[12] Kim G, Kwon T, et al. Diffusionclip: text-guided diffusion models for robust image manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 2426-2435.

[13] Yang B, Gu S, et al. Paint by example: exemplar-based image editing with diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18381-18391.

[14] Xie S, Zhao Y, et al. DreamInpainter: Text-Guided Subject-Driven Image Inpainting with Diffusion Models[J]. arXiv preprint arXiv:2312.03771, 2023.

[15] Li S, Chen C, et al. MoEController: Instruction-based Arbitrary Image Manipulation with Mixture-of-Expert Controllers[J]. arXiv preprint arXiv:2309.04372, 2023.

[16] Zhang K, Mo L, et al. Magicbrush: A manually annotated dataset for instruction-guided image editing[J]. Advances in Neural Information Processing Systems, 2024, 36.

[17] Zhang S, Yang X, et al. Hive: harnessing human feedback for instructional visual editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9026-9036.

[18] Mokady R, Hertz A, et al. Null-text inversion for editing real images using guided diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6038-6047.

[19] Yang, Fei and Yang, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing[J]. Advances in Neural Information Processing Systems,2024,36.

[20] Hertz A, Aberman K, et al. Delta denoising score[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 2328-2337.

[21] Zhang Z, Han L, et al. Sine: single image editing with text-to-image diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6027-6037.

[22] Wang Z, Zhao L, et al. Stylediffusion: controllable disentangled style transfer via diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 7677-7689.

[23] Shi Y, Xue C, et al. Dragdiffusion: harnessing diffusion models for interactive point-based image editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 8839-8849.

[24] Cao M, Wang X, et al. Masactrl: tuning-free mutual self-attention control for consistent image synthesis and editing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 22560-22570.

[25] Liu H, Xu C, et al. Drag your noise: interactive point-based editing via diffusion semantic propagation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 6743-6752.

[26] Tumanyan N, Geyer M, et al. Plug-and-play diffusion features for text-driven image-to-image translation[C]//Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, 2023: 1921-1930.
[27] Duan X, Cui S, et al. Tuning-free inversion-enhanced control for consistent image editing[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 1644-1652.
[28] Qiao Y, Wang F, et al. BARET: balanced attention based real image editing driven by target-text inversion[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4560-4568.
[29] Wang J, Liu P, et al. Unified Diffusion-Based Rigid and Non-Rigid Editing with Text and Image Guidance[J]. arXiv preprint arXiv:2401.02126, 2024.
[30] Lu S, Liu Y, et al. Tf-icon: diffusion-based training-free cross-domain image composition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 2294-2305.
[31] Mao Q, Chen L, et al. MAG-Edit: Localized Image Editing in Complex Scenarios via Mask-Based Attention-Adjusted Guidance[J]. arXiv preprint arXiv:2312.11396, 2023.
[32] Nam H, Kwon G, et al. Contrastive denoising score for text-guided latent diffusion image editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9192-9201.
[33] Huang J, Liu Y, et al. Kv inversion: kv embeddings learning for text-conditioned real image action editing[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2023: 172-184.
[34] Elarabawy A, Kamath H, et al. Direct inversion: Optimization-free text-driven real image editing with diffusion models[J]. arXiv preprint arXiv:2211.07825, 2022.
[35] Huberman-Spiegelglas I, Kulikov V, et al. An edit friendly ddpm noise space: inversion and manipulations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 12469-12478.
[36] Wu Q, Liu Y, et al. Uncovering the disentanglement capability in text-to-image diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 1900-1910.
[37] Zhao J, Zheng H, et al. Null-text guidance in diffusion models is secretly a cartoon-style creator[C]//Proceedings of the 31st ACM International Conference on Multimedia, 2023: 5143-5152.
[38] Miyake D, Iohara A, et al. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models[J]. arXiv preprint arXiv:2305.16807, 2023.
[39] Han L, Wen S, et al. Proxedit: improving tuning-free real image editing with proximal guidance[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024: 4291-4301.
[40] Avrahami O, Lischinski D, et al. Blended diffusion for text-driven editing of natural images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18208-18218.
[41] Avrahami O, Fried O, et al. Blended latent diffusion[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-11.
[42] Huang W, Tu S, et al. Pfb-diff: Progressive feature blending diffusion for text-driven image editing[J]. arXiv preprint arXiv:2306.16894, 2023.
[43] Couairon G, Verbeek J, et al. Diffedit: Diffusion-based semantic image editing with mask guidance[J]. arXiv preprint arXiv:2210.11427, 2022.
[44] Mirzaei A, Aumentado-Armstrong T, et al. Watch your steps: Local image and scene editing by text instructions[J]. arXiv preprint arXiv:2308.08947, 2023.
[45] Liu Z, Zhang F, et al. Text-guided mask-free local image retouching[C]//2023 IEEE International Conference on Multimedia and Expo (ICME), 2023: 2783-2788.
[46] Wang Q, Zhang B, et al. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions[J]. arXiv preprint arXiv:2305.18047, 2023.
[47] Yu T, Feng R, et al. Inpaint anything: Segment anything meets image inpainting[J]. arXiv preprint arXiv:2304.06790, 2023.
[48] Brown T, Mann B, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
[49] Radford A, Narasimhan K, et al. Improving language understanding by generative pre-training[J]. OpenAI, 2018. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf
[50] Alldieck T, Kolotouros N, et al. Score Distillation Sampling with Learned Manifold Corrective[J]. arXiv preprint arXiv:2401.05293, 2024.
[51] Pan X, Tewari A, et al. Drag your gan: interactive point-based manipulation on the generative image manifold[C]//ACM SIGGRAPH 2023 Conference Proceedings, 2023: 1-11.
[52] Hu, Edward J, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.
[53] Alaluf Y, Garibi D, et al. Cross-image attention for zero-shot appearance transfer[J]. arXiv preprint arXiv, 2023, 2311.03335.

[54] Park T, Efros AA, et al. Contrastive learning for unpaired image-to-image translation[C]//Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX 16, 2020: 319-345.
[55] Qi L, Kuen J, et al. Open world entity segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
[56] Li J, Li D, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning, 2023: 19730-19742.
[57] Liu S, Zeng Z, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection[J]. arXiv preprint arXiv:2303.05499, 2023.
[58] Kirillov A, Mintun E, et al. Segment anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4015-4026.
[59] Suvorov R, Logacheva E, et al. Resolution-robust large mask inpainting with fourier convolutions[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2149-2159.
[60] Dehghan A, Masood SZ, et al. View independent vehicle make, model and color recognition using convolutional neural network[J]. arXiv preprint arXiv:1702.01721, 2017.
[61] Lin TY, Maire M, et al. Microsoft coco: common objects in context[C]//Computer Vision--ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, 2014: 740-755.
[62] Deng J, Dong W, et al. Imagenet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.
[63] Heusel M, Ramsauer H, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.
[64] Szegedy C, Vanhoucke V, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818-2826.
[65] Wang Z, Bovik AC , et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
[66] Basu S, Saberi M, et al. Editval: Benchmarking diffusion based text-guided image editing methods[J]. arXiv preprint arXiv:2310.02426, 2023.
[67] Wu X, Sun K, et al. Human preference score: better aligning text-to-image models with human preference[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 2096-2105.
[68] Brack M, Friedrich F, et al. Ledits++: limitless image editing using text-to-image models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 8861-8870.
[69] Mou C, Wang X, et al. Dragondiffusion: Enabling drag-style manipulation on diffusion models[J]. arXiv preprint arXiv:2307.02421, 2023.
[70] Jung Y, Lee S, et al. Latent Inversion with Timestep-aware Sampling for Training-free Non-rigid Editing[J]. arXiv preprint arXiv:2402.08601, 2024.