刘鹏举, 石宇鹏,张宏志,等. 基于预训练扩散模型的图像实例语义与视觉和谐化[J]. 智能计算机与应用,2025,15(3):7-17. DOI:10.20169/j.issn.2095-2163.24111102

基于预训练扩散模型的图像实例语义与视觉和谐化

刘鹏举[1,2], 石宇鹏[3], 张宏志[3], 姜 峰[2], 左旺孟[3]

(1 哈尔滨工业大学 医学与健康学院, 哈尔滨 150001; 2 哈尔滨工业大学郑州研究院 医学健康研究院, 郑州 450000;
3 哈尔滨工业大学 计算学部, 哈尔滨 150001)

摘 要: 近年来,图像实例和谐化作为图像生成领域中的重要分支得到了迅速发展。然而,如何确保前景实例与背景图像中的各个元素在语义上具备合理的逻辑关系,并使组合后的图像内容和谐一致,仍是当前研究面临的难点。此外,受限于高成本和设备要求,收集大规模的和谐化训练数据存在诸多困难。为解决这些问题,本文提出一种基于大规模预训练扩散模型的图像和谐化方法。该方法基于预训练的 Stable Diffusion 2.0 模型,采用自然语言引导图像填充任务,使模型能够在自然语言描述和待填充区域图像的条件下生成符合语义需求的和谐图像。本方法将实例图像的高频信息与低频信息分别作为控制条件,对预训练模型进行微调,以确保生成结果尽可能保留实例图像的关键内容,最终生成和谐的组合图像。实验结果表明,本方法在生成实例阴影、调节光照等方面均表现出优异的效果,有效提升了图像语义与视觉的和谐化质量。

关键词: 图像实例和谐化; 预训练扩散模型; 自然语言引导; 高频信息与低频信息

中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0007-11

Image instances semantics and visual harmonization using pretrained diffusion model

LIU Pengju[1,2], SHI Yupeng[3], ZHANG Hongzhi[3], JIANG Feng[2], ZUO Wangmeng[3]

(1 School of Medicine and Health, Harbin Institute of Technology, Harbin 150001, China;
2 Zhengzhou Research Institute, Harbin Institute of Technology, Zhengzhou 450000, China;
3 Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Recent advancements in image generation have led to significant progress in image instance harmonization. However, maintaining semantic consistency between foreground and background elements and achieving visually plausible combinations remain a challenging task. Additionally, the scarcity of large-scale harmonization datasets limits the development of effective methods. To address these challenges, the paper proposes a novel image harmonization approach based on a large-scale pretrained diffusion model. Leveraging the powerful capabilities of Stable Diffusion 2.0, the paper formulates image harmonization as a text-guided image inpainting task. By providing natural language descriptions and specifying target regions, the proposed model can generate harmonized images that seamlessly blend with the background. To further enhance the quality of the generated images, the paper incorporates high-frequency and low-frequency information from the foreground instance as control conditions, ensuring that the essential features of the instance are preserved. Experimental results demonstrate that the proposed approach significantly improves image harmonization quality, especially in terms of generating realistic shadows and adjusting lighting effects.

Key words: image instance harmonization; pretrained diffusion model; natural language descriptions; high-frequency and low-frequency information

0 引 言

图像实例和谐化是计算机视觉领域一项基础性任务,旨在无缝融合前景与背景图像[1]。其广泛应用于各种图像处理应用中,但由于光照、色彩和纹理等方面的差异,前景与背景的自然过渡一直是研究

基金项目: 国家自然科学基金面上项目(6237011159); 中国博士后基金项目(2024M754208)。

作者简介: 刘鹏举(1989—),男,博士,助理研究员/工程师,主要研究方向:计算机底层视觉,医学图像处理。

通信作者: 张宏志(1976—),男,博士,副教授,博士生导师,主要研究方向:人工智能,计算机视觉,融合智能与机器学习。Email:zhanghz@hit.edu.cn。

收稿日期: 2024-11-11

难点。合成图像中视觉元素的不匹配常导致画面缺乏和谐统一感。因此,图像实例和谐化技术至关重要,它通过优化组合图像,提升视觉美感,实现自然流畅的视觉效果。传统的图像编辑方法依赖人工设计特征和规则,例如依靠专业人士使用图像处理软件手动调整光影、色彩和明暗细节。这种方法不仅需要高度专业技能,也限制了非专业用户的效率,且在复杂场景下的效果有限[2]。为解决这些问题,研究者们积极探索自动化算法,以期简化操作流程,为非专业人士提供更便捷的图像和谐化途径。

深度学习(Deep Learning)的迅猛发展为图像和谐化带来了革命性的变革[3]。相比传统方法依赖于手工设计的特征和规则,深度学习模型能够自动学习图像的深层次特征,并通过端到端的训练实现对图像的精细化编辑。基于编码器-解码器架构的深度学习模型已成为图像和谐化研究的主流[4]。这些模型通过提取前景和背景图像的多尺度特征,并结合注意力机制,实现对图像信息的自适应融合。此外,生成对抗网络(Generative Adversarial Network, GAN)的引入进一步提升了生成图像的质量,使得合成图像能够更好地融入背景[5]。通过对抗学习,生成器不断学习生成更逼真的图像,而判别器则努力区分生成图像与真实图像。这种对抗过程促使生成器不断提高生成图像的质量,最终实现前景与背景的无缝融合。

近年来,扩散模型(Diffusion Model)[6]在图像生成领域取得了引人注目的进展[7]。大型预训练的扩散模型能够生成细节精致、质量卓越的高分辨率图像,且在众多任务上超越了传统的生成对抗网络。扩散模型在图像实例和谐化领域的应用为这一难题提供了新的解决视角。本文提出了一种基于大规模预训练扩散模型的图像实例语义与视觉和谐化方案,旨在应对背景图像与前景实例不一致时的图像调整挑战。具体而言,本研究深入探索了基于扩散模型的生成方法,通过自然语言引导的图像填充任务,使模型能够在给定自然语言描述和待填充区域图像的条件下,生成符合要求的图像。随后,通过条件微调模型,进一步优化图像填充区域,确保生成图像与前景实例的和谐融合。更具体地,对预训练的基于自然语言的图像填充模型扩展,使得其能够输入更多条件。为了保证模型的对填充区域的生成能力,本研究对前景实例分别提取高频信息以及低频信息,将高低频信息作为 2 种控制条件,使得模型可以生成更具真实性的最终图像。

实验结果显示,基于预训练扩散模型的图像语义与视觉和谐化方法能够生成视觉质量极高的图像,且填充结果展现出卓越的和谐性,显著提升合成图像的整体视觉效果。本文提出的基于预训练扩散模型的图像和谐化方法,不仅在图像合成的精准度上表现优异,还能有效克服传统方法在处理复杂图像合成任务时的局限,展现出更强的自适应性和灵活性,为图像处理技术的发展指明了新的方向。本文的主要贡献总结如下:

(1)本文首次将大规模预训练扩散模型应用于图像和谐化任务,通过充分利用模型在海量数据上学习到的丰富先验知识,显著提升了生成图像的质量和多样性。

(2)通过引入自然语言描述和图像的高频细节与低频语义信息作为控制条件,本文提出的方法实现了对生成图像的精细控制。生成的图像不仅具有丰富的细节纹理,而且能够与背景实现平滑过渡,从而满足多样化的生成需求。

(3)在合成数据集上取得 SOTA 性能。在 Open Image 数据集上进行了大量实验,结果表明本文的方法在生成图像的质量和多样性方面均达到了最先进水平。

1 相关工作

1.1 图像生成模型

图像生成技术旨在设计合理的概率模型,以准确拟合从随机变量到真实图像的映射关系。根据不同的建模方法,基于深度学习的图像生成模型呈现出多种形式,主要包括变分自编码器(Variational Autoencoder, VAE)[8]、生成式对抗网络(Generative Adversarial Network, GAN)[9]和去噪扩散模型(Denoising Diffusion Probabilistic Model, DDPM)[6,10]等。近年来,基于 GAN 和扩散模型的图像生成技术得到了广泛关注和深入研究。由 Goodfellow 等学者[9]提出的 GAN,通过生成器和判别器之间的对抗性训练,逐步逼近真实数据分布,从而生成高质量的图像。研究者们不断优化 GAN 中的生成器和判别器结构,使其在不同图像生成任务中表现出色[10-11]。例如,StyleGAN[12-13]引入自适应特征变换方法来注入图像的风格信息,从而显著提升生成图像的质量,并为后续的图像编辑任务奠定了重要基础[14-18]。此外,BigGAN[19]通过引入类别条件的批量归一化机制,使生成器能够生成多种类别的自然图像,进一步扩展了其在图像生成任务中

的适用范围。

随着计算资源的提升和大规模预训练模型的不断发展，去噪扩散模型（DDPM）逐渐成为学术界的研究热点。DDPM 通过前向加噪和反向去噪过程实现图像生成：在前向过程中，模型逐步向样本数据添加噪声；在反向过程中，模型逐步去除噪声以恢复真实数据分布。基于 DDPM 的模型在多模态生成任务中取得了显著进展[20-22]。其中，Stable Diffusion[22]利用感知压缩方法将图像表示压缩到低维空间中，并在该空间中通过 DDPM 进行建模，极大地提高了生成速度和计算效率，同时减小了模型规模，使其在性能和资源消耗方面具有独特优势。

1.2　图像实例和谐化

在图像处理任务中，将前景实例直接粘贴到背景图像上后，往往会产生明显的画面不协调问题。图像实例和谐化的目的是对这一组合结果进行处理，使前景与背景的视觉效果更加和谐统一。传统图像和谐化方法主要关注图像信号的统计特征，对图像的颜色、对比度等方面进行基本矫正[23-25]。这些算法通常较为简单高效，因此被广泛集成在各类图像处理软件中。然而，传统方法主要用于校正图像内部颜色分布不均等简单场景，难以应对前景和背景在内容、色彩等方面差异较大的复杂情况。图像实例和谐化的目标是更进一步解决前景与背景的边缘不一致、阴影缺失、亮度差异等细节问题。

随着深度学习的快速发展，基于卷积神经网络的图像和谐化技术以其更强的鲁棒性和泛化能力受到广泛关注。近年来，GANs 在图像翻译任务中取得了突破性进展，推动了其在图像和谐化问题中的应用探索。例如，Bargainnet 通过在生成器中引入注意力机制，改进了 U-Net 结构，并设计了基于 Domain Translation 的训练方式，从而提升了前景实例与背景图像在风格和内容上的一致性[26]。类似地，S^2AM 模型通过注意力机制捕获前景与背景之间的交互信息，并将这些信息融入 U-Net 的跳跃连接过程，以增强生成图像的整体协调性。在此基础上，文献[27]进一步改进了 S^2AM，采用预训练模型来提取前景和背景的交互信息，显著提升了处理效果。FRIH[28]通过分阶段处理和自适应聚类，实现了更细粒度的区域感知和谐化，从而生成更逼真的合成图像。虽然上述模型在真实性、效率和一致性等方面得到了快速发展，因语义信息、及真实训练数据缺乏等问题，还需要进一步发展。

1.3　语义图像合成

在给定空间语义标定图（简称语义图）的条件下，语义图像生成的目标是生成逼真的图像，并确保生成图像的空间语义类别与语义图精确对应。作为一种条件图像生成模型，语义图像生成允许用户通过语义图来控制图像内容，从而大大提升了图像生成的灵活性。这项技术具有广泛的应用前景，近年来因其独特的实用价值而受到高度关注，特别是基于 GANs 的语义图像生成技术取得了显著进展。

在基于 GANs 的语义图像生成中，研究者们不断改进生成器和判别器，以提升生成图像的细节和真实感。例如，SPADE[29]提出了自适应空域特征变换方法，通过在生成器的特征批量归一化中自适应学习仿射变换参数，使生成器更好地利用语义信息。CC-FPSE[30]则通过语义图预测生成器中卷积核的参数，以调制卷积过程中的特征响应，从而增强生成图像的语义一致性。同样地，SC-GAN[31]利用语义图预测生成语义向量，自适应学习生成器中相关的卷积参数，进一步改善了语义对齐效果。LGGAN[32]则更进一步，利用局部上下文信息，提出额外分支来精细化生成图像中的细节，从而增强了图像的局部一致性和自然感。另一方面，CollogeGAN[33]通过为特定语义类别训练多个 StyleGAN，以提高图像中特定类别的细节质量，精细化生成的图像效果。总之，基于 GANs 的语义图像生成技术通过一系列创新性的结构设计，使生成器能够更精确地响应语义信息，实现了对图像内容的有效控制，为语义图像生成的实际应用提供了坚实的技术基础。这些方法的不断优化，为高质量、定制化图像生成提供了新的思路和可能性。

2　方法

在本节内容中，首先对基于扩散模型的图像填充进行详细介绍。接下来，将详细阐述本研究动机和详细的方法。最后，将对本文所提出的模型训练及损失函数进行深入介绍。

2.1　基于扩散模型的图像填充

扩散模型（Diffusion Model，DM）作为一种概率生成模型，通过在噪声向量之上实施一系列细致的概率扩散步骤，从而逐步合成出高质量的图像。在每一个扩散阶段，该模型首先将现有的噪声向量与一个新的随机噪声向量相结合，随后通过一个可逆转换网络进行处理。随着扩散步骤的递进，模型能够逐渐细化和完善图像的质量，每一阶段都在不断

地丰富着图像的细节与结构复杂度。扩散模型的学习机制基于一个时长为 T 的马尔可夫去噪流程,该流程可以被划分为 2 个主要阶段:前向加噪过程与反向去噪过程。在前向加噪阶段,假定有一个初始样本 x_0,从给定的真实数据分布 $q(x)$ 中抽取而来。模型将在 x_0 的基础上执行总共 T 步的噪声注入操作。在此过程中,每一步所加入的噪声水平由预定义的一系列方差值 $\{\beta_t \in (0,1)\}_{t=1}^{\mathrm{T}}$ 来控制,这些方差决定了每一次迭代中噪声强度的变化。具体而言,加噪过程遵循特定的算法设计,旨在模拟从清晰到模糊的数据退化路径,为后续反向去噪过程奠定基础。其加噪过程为:

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

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

其中,x_t 表示时间步为 t 时的加噪结果,$\boldsymbol{I}$ 表示标准的高斯噪声。当步数 T 足够大时,x_T 将无限趋近于高斯噪声。在反向去噪过程中,模型 $\epsilon_\theta(x_t, t)$ 预测 x_t 的去噪结果 x_{t-1}。实际应用中,$\epsilon(\cdot)$ 通常使用 U-Net 结构的自编码器实现。训练过程中,目标损失函数可简化为:

$$\mathcal{L}_{\mathrm{ctrl}} = E_{x,\epsilon\sim N(0,1),t}[\| \epsilon - \epsilon_\theta(x_t, t) \|_2^2] \quad (2)$$

其中,t 表示 $\{0, 1, \cdots, T\}$ 中的随机采样。在扩散模型(DM)中,高斯噪声采样经过一系列的去噪过程,逐步逼近真实数据分布。这一过程可视为一个逆扩散过程,通过不断降低噪声水平,最终生成高质量的图像。去噪过程的关键在于一个可学习的去噪函数 $\epsilon(\cdot)$,该函数能够根据当前噪声样本和时间步,预测出上一时刻的噪声样本。

为了缓解高维图像数据带来的训练复杂度和不稳定性问题,Stable Diffusion (SD)[22] 提出了一种基于两阶段的扩散生成模型。Stable Diffusion 首先利用一个自编码器对图像进行感知压缩,将原始的高维图像空间映射到一个低维的潜在空间。具体而言,编码器 EI 将原始图像 $x_0 \in \mathbb{R}^{w_x \times h_x \times 3}$ 编码为潜在表示 $Z_0 \in \mathbb{R}^{w_z \times h_z \times 4}$,其中 $w_x > w_z, h_x > h_z$。通过该方式,Stable Diffusion 有效地去除了图像数据中的冗余信息,降低了模型的训练难度。随后,Stable Diffusion 在这个低维的潜在空间上进行扩散过程,从而生成高质量的图像。由于潜在空间的维度显著降低,模型的训练变得更加稳定。

基于 Stable Diffusion 的图像填充模型[16] 采用了以语义文本为引导的 U-Net 结构去噪网络 $\epsilon(\cdot)$ 来实现图像修补。该模型以待修补图像和相应的语义文本描述作为联合输入,通过优化约束目标函数来生成修补后的图像。具体而言,模型将文本编码为语义向量,并将其与图像的潜在表示相结合,作为去噪网络的输入。类似于式(2),模型的优化目标也写为:

$$\mathcal{L}_{\mathrm{ctrl}} = E_{x,\epsilon\sim N(0,1),t}[\| \epsilon - \epsilon_\theta(x_t, t) \|_2^2] \quad (3)$$

在此过程中,Stable Diffusion 采纳了 CLIP 中的中文文本编码器,对输入的文本进行精准编码,生成文本条件 C_l。同时,利用预训练的自编码器技术,对内容缺失的图像进行有效处理,从而提取出图像条件 C_i。随后,通过利用交叉注意力机制将文本条件 C_l 巧妙地融入到 U-Net 网络中,从而得到:

$$\boldsymbol{Q} = \boldsymbol{W}_Q \Psi(Z_t); \quad \boldsymbol{K} = \boldsymbol{W}_K \Phi(C_l); \quad \boldsymbol{V} = \boldsymbol{W}_V \Phi(Z_l);$$

$$\tau(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\boldsymbol{V}\right) \quad (4)$$

其中,$\Psi(\cdot)$ 与 $\Phi(\cdot)$ 分别表示去噪网络 ϵ_θ 在隐空间中对噪声图像编码 Z_t 以及文本条件 C_l 进行深度特征提取的操作;$\tau(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ 负责融合这些特征;σ 为 Softmax 激活函数,用于增强结果的判别性。设定内容缺失图像的二值掩码为 $M \in \mathbb{R}^{w_x \times h_x}$,其中,掩码值 M 为 1 的区域标志着图像内容的缺失。图像条件 C_i 是通过预训练的自编码器对内容缺失图像进行编码所得,随后,将 C_l 与噪声图像的隐空间编码 Z_t 在通道维度上进行拼接,共同构成去噪网络的输入:

$$Z_{\mathrm{in}} = c(Z_t, C_i, M') \quad (5)$$

其中,Z_{in} 为去噪网络输入;函数 c 为按维度拼接操作;$M' \in \mathbb{R}^{w_z \times h_z}$ 由 M 经最近邻插值得到。

在测试阶段,对于给定的内容缺失图像 I_{in} 及对应的文本描述 T,研究首先利用编码器 E_I 对图像 I_{in} 进行编码,以获得图像条件编码 C_i。同时,借助 CLIP 文本编码器,将文本描述 T 转化为文本编码 C_T。在此基础上,从高斯分布中采样得到一个初始噪声图 $Z_T \in \mathbb{R}^{w_z \times h_z \times 4}$。随后,去噪网络 ϵ_θ 以图像条件编码 C_i 和文本编码 C_l 为条件,对初始噪声图 Z_T 进行逐步去噪处理,直至得到高质量的隐空间表示 Z_0。最后,通过解码器 D_I 对 Z_0 进行解码,从而生成目标填充图像。

2.2 基于扩散模型的图像填充

虽然 Stable Diffusion 的图像填充效果良好,但其生成内容由文本引导,难以与图像实例保持一致。本文提出一种基于 ControlNet[34] 微调 Stable

Diffusion 模型的方案，以最大程度地保留图像实例内容，从而为图像实例和谐化提供一种新的解决方案。本文设计的 Stable Diffusion 网络微调结构如图 1 所示。

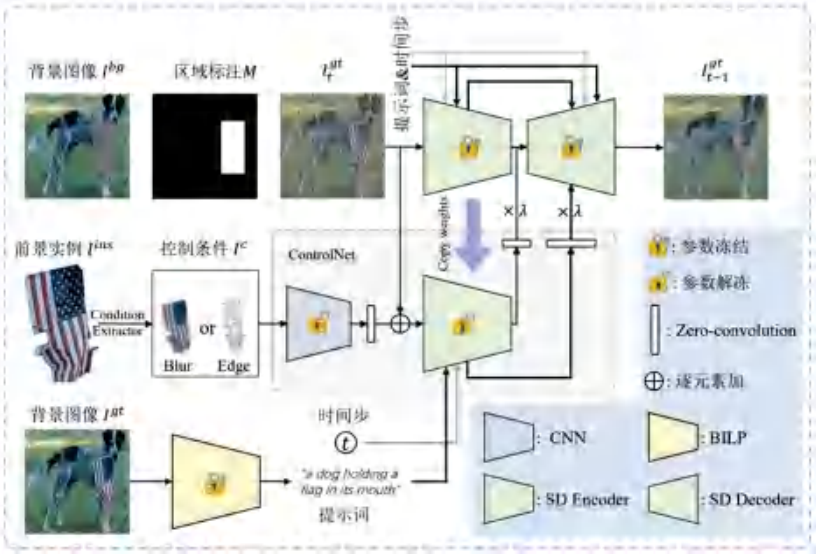


图 1　模型训练架构图

Fig. 1　Model training architecture

通过借鉴 ControlNet[34] 的思路，为 Stable Diffusion 引入了额外的编码器，以增强图像生成的控制能力。具体而言，给定前景实例 I^{ins} 和背景图像 I^{bg}，本方法从前景实例 I^{ins} 中提取条件信息，包括高频和低频信息，将其作为控制条件 I^c。随后，I^c 通过多层卷积网络进一步提取信息，作为 Stable Diffusion 中额外编码器的输入。编码器输出的信息经过缩放后，以跳跃连接的方式作为残差加入到 Stable Diffusion 解码器的每一层特征中，从而提升模型对条件信息的响应。由于数据标注形式的限制，本方法难以获取大量带有文本描述的图像，因此使用 BLIP[35] 生成图像的文本描述，以弥补这一不足。

为提高训练的稳定性，本方法采用 Zero-Convolution 操作，即在将 I^c 提取的特征输入额外编码器之前，先经过参数初始化为 0 的卷积操作，与背景图像 I^{bg} 在时间步 t 的加噪结果相加，然后输入到额外编码器中。同时，额外编码器的输出也经过 Zero-Convolution 后，作为残差加入到 Stable Diffusion 解码器中。整个训练过程中，仅更新卷积神经网络和额外编码器、即 ControlNet 中的参数，从而确保模型在训练初始阶段能够生成较为合理的图像，提供更优的参数初始值，使训练过程更加稳定。整个过程可用如下公式表述：

$$\boldsymbol{F}_c = Z\{F_{\text{ctrl}}[Z[C(\sigma(I^c))] + I_t^{\text{gt}}]\}$$

$$\bar{\boldsymbol{F}}^i_{\text{dec}} = \boldsymbol{F}^i_{\text{dec}} + \lambda \boldsymbol{F}^i_c \tag{6}$$

其中，$\boldsymbol{F}_c$ 表示 ControlNet 的输出特征；Z 表示 Zero-Convolution；F_{ctrl} 表示额外引入的特征编码器；C 表示图 1 中卷积神经网络；σ 表示前景实例图像条件提取算子；$\boldsymbol{F}^i_{\text{dec}}$ 表示 Stable Diffusion 解码器特征；λ 表示控制条件强度；$\bar{\boldsymbol{F}}^i_{dec}$ 表示融合前景图像信息的 Stable Diffusion 解码器特征。

本方法通过使用高斯模糊将前景实例的图像信号分解为高频和低频信息，分别用于保持前景实例的结构细节和颜色特征，并提出了一种多条件融合策略来整合这些特征。具体而言，高频特征有助于保持前景实例的轮廓和纹理信息，而低频特征用于保留其整体色彩和光影效果。通过多条件融合策略，本方法在尽可能保留前景实例原始信息的前提下，实现了前景和背景的和谐融合，生成视觉一致性更高的合成图像。

(1)结构控制：在条件图像合成任务中，边缘信息常作为图像内容的结构控制条件[36-38]。为确保和谐化结果中的实例结构与待融合图像实例的结构尽可能保持一致，本文选用 Canny 算子提取实例图像的边缘信息来引导和谐化图像的生成。不同控制条件和谐化结果如图 2 所示。图 2(a)为未经和谐化的图像(前景实例直接粘贴的效果)，与图 2(b)、图 2(c)(右下角为边缘特征)对比，可以看出利用边缘信息能够有效保留前景实例的结构特征。

(2) 颜色控制：虽然高频特征可以有效保留前景实例的结构信息，但其颜色特征未能得到充分保留，从而导致图像色彩差异较大。为解决该问题，本

方法进一步引入图像的低频特征作为控制条件,以维持前景实例的色彩一致性。本方法采用高斯模糊方式提取前景实例的颜色特征,如图 2(d)右下角所示,该颜色特征反映了前景实例各空间位置的色彩信息。通过使用低频特征来控制图像合成,有效保留了前景实例的颜色特性(见图 2(d))。

(a) 像素粘贴 (b) 无条件 (c) 高频特征条件 (d) 低频特征条件 (e) 联合特征条件

图 2 不同控制条件和谐化结果图

Fig. 2 Harmonization results under different control conditions

(3)高低频特征联合引导的图像实例和谐化:本文提出 一种融合高频和低频条件信息的方法,假设 $\boldsymbol{F}_c^l$ 和 $\boldsymbol{F}_c^h$ 分别表示 ControlNet 通过结构控制和颜色控制获得的中间特征,则两者联合引导下的中间特征可以定义为:

$$\ddot{\boldsymbol{F}}_{\text{dec}} = \boldsymbol{F}_{\text{dec}} + \lambda_l \boldsymbol{F}_c^l + \lambda_h \boldsymbol{F}_c^h \tag{7}$$

其中,λ_h 与 λ_l 分别表示结构控制和颜色控制的强度。模型测试示意如图 3 所示。实验结果表明,通过两者的组合进行前景实例控制,能够有效地保留输入图像实例的视觉特征,同时为用户提供了一种多模态的生成方案。

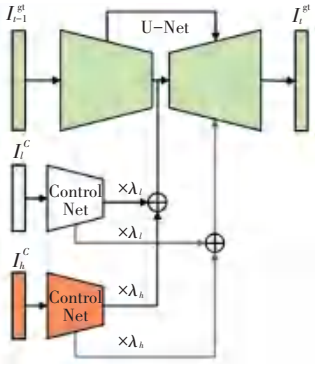


图 3 模型测试示意图

Fig. 3 Schematic diagram of model testing

2.3 模型训练及损失函数

在优化过程中,本方法冻结了 Stable Diffusion 的所有参数,仅对 ControlNet 中的参数进行训练。训练过程依循类似于式(8)的约束方法,先后优化颜色式控制和结构控制模型:

$$\mathcal{L}_{\text{ctrl}} = E_{x,\epsilon\sim N(0,1),t}\left[\| \epsilon - \epsilon_\theta(Z_{\text{in}}, t, \Psi(y), \lambda \boldsymbol{F}_c \|_2^2\right] \tag{8}$$

具体而言,给定背景图像 I^{bg} 及其中的实例区域标注 M(其中 1 表示实例区域),本方法首先获取缺失区域图像 I^{ms} 及其包含的前景实例 I^{ins},研究推得:

$$I^{ms} = (1 - m) \odot I^{\text{bg}} \tag{9}$$

$$I^{\text{ins}} = m \odot I^{\text{bg}} \tag{10}$$

随后,分别将 I^{bg} 和 I^{ms} 编码至隐空间中,得到编码结果 Z_0 和 Z_{ms}。接着,从序列 $\{0,1,\cdots,T\}$ 中随机采样一个时间步 t,并对 Z_0 执行 t 步加噪操作,以得到 Z_t。训练过程中使用式(8)中的损失函数进行约束。

在此,

$$Z_{\text{in}} = c(Z_t, Z_{ms}, M') \tag{11}$$

其中,M' 是通过 M 最近邻插值到与 Z_t 相同尺寸的掩模,三者通过通道维度拼接得到融合特征;y 表示背景图像 I^{bg} 经 BLIP 模型生成的文本描述;$\boldsymbol{F}_c$ 为通过 ControlNet 处理后提取的前景实例的边缘特征(或高斯模糊)结果。由于本方法利用 BLIP 模型生成自然语言描述,训练时仅需使用自然图像样本及其实例级标注,简化了对数据的依赖。

3 实验验证

文中将详细介绍本研究所采用的训练设置,包括数据集的选择、训练过程的实验设置以及评价指标的定义。接下来,通过定量和定性的实验结果,将展示所提方法的有效性。此外,本研究还进行了消融实验,以系统验证模型各个组件在整体性能中的作用和贡献。这些实验结果不仅为方法的有效性提供了实证支持,也为后续研究提供了重要的参考依据。

3.1 训练和测试设置

(1)数据集:本文实验基于 Open Images 数据集[39]进行。该数据集包含 900 万多张图片和超过 1 亿个标注框,涵盖了自然场景、建筑、街景、艺术品等多个领域,共涉及 6 000 余种类别。其标注信息结合了人工标注与机器辅助标注,提供了丰富的实例标注和分类标注,具备较高的数据多样性和标注精度。由于数据集规模较大,本研究方法选取了其

中 12.5%的子集用于模型训练。在测试阶段，实验从 Open Images 数据集中随机抽取了 2 000 张测试样本，确保每张样本图片仅包含一个实例标注，且不对实例类别加以限制，以全面评估模型的泛化能力和不同类别实例的处理效果。

（2）实验设置：本方法使用 8 张 NVIDIA A6000 显卡进行训练，每次迭代的样本数量设为 64。训练采用 Adam 优化器，初始学习率为 1×10^{-6}，权重衰减率设为 0.01，并在训练初期进行 10 000 步的学习率 warmup[40]。所有训练图像的尺寸统一为 512×512。本方法基于 Stable Diffusion 预训练模型的 2.0 版本，且在训练过程中条件控制强度均设置为 1。以边缘信息为控制条件的实验共耗时 24 h，以颜色信息为控制条件的实验耗时 96 h。

（3）评价指标：图像实例和谐化的目标是确保和谐化结果与输入内容的一致性，同时提高生成图像的真实性。为衡量内容一致性，本方法使用均方误差（Mean Squared Error，*MSE*）和峰值信噪比（Peak Signal-to-Noise Ratio，*PSNR*）[41] 作为指标。其中，*MSE* 和 *PSNR* 用于计算和谐化图像与对应真实图像在像素层面的差异。为了评估和谐化结果的真实性，本方法使用 Fréchet Inception Distance（*FID*）[42]。*FID* 通过 Inception-v3 模型提取真实数据集和生成图像集的特征向量，并计算两者特征分布的 Fréchet 距离，其计算公式如下：

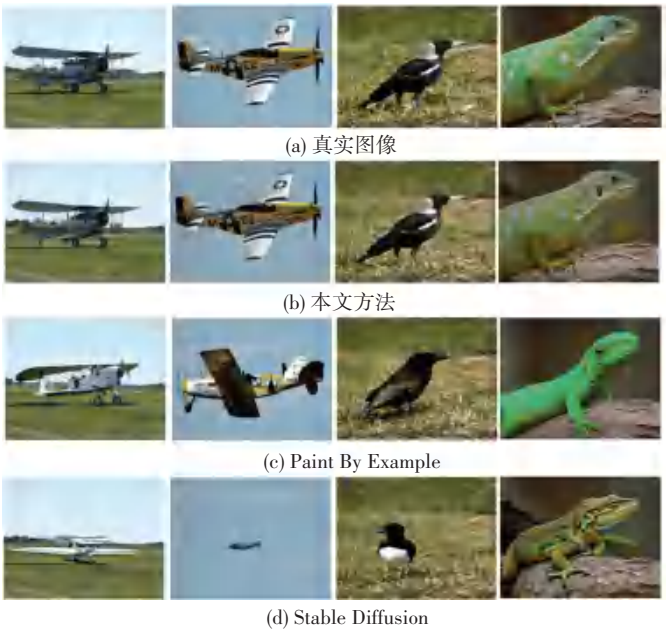
$$FID = \|\mu_x - \mu_y\|^2 + Tr(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y - 2(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_y)^{\frac{1}{2}}) \quad (12)$$

其中，μ_x 和 μ_y 分别表示真实数据集和生成数据集的特征向量集合的均值；$\boldsymbol{\Sigma}_x$ 和 $\boldsymbol{\Sigma}_y$ 分别表示两者的协方差矩阵；*Tr* 表示矩阵的迹运算；$\|\cdot\|$ 表示向量的二范数。*FID* 值越小，表示生成数据集的特征分布与真实数据集更加接近，表明生成模型的图像质量越高。

3.2　定量和定性实验结果

3.2.1　定性实验结果

图 4 展示了图像填充的对比结果。测试样本选自 Open Image 数据集中仅包含一个前景实例的图像。具体而言，本节首先从数据集中选取一张图像，将其作为背景图输入模型，但前景部分进行了擦除；然后，以该图像的前景图作为组合实例输入模型。在本方法中，高频和低频条件强度均设置为 1（$\lambda_l = \lambda_h = 1$）。对于 Paint By Example[43]，测试时以前景图像作为其参考输入；对于 Stable Diffusion，使用 BLIP 获取前景图像的文本描述并输入模型以生成对应图像。对比图 4 中结果图可见，本方法在保持原始图像实例信息方面表现最佳。受限于自然语言描述的容量，Stable Diffusion 仅保留了前景实例的部分特征（如类别和大致颜色），而 Paint By Example 虽提取了图像特征，但由于其特征表示的维度较小（$d = 1\,024$），无法很好地呈现实例图像的细节信息。得益于本方法对前景实例高频和低频特征的有效利用，生成结果能最大程度地保留前景实例的细节和信息完整性。图 5 展示了图像和谐化的结果，从图 5 中可以看出，本方法在调整光照、颜色以及为前景实例增添合理阴影方面效果显著。



(a) 真实图像

(b) 本文方法

(c) Paint By Example

(d) Stable Diffusion

图 4　各个方法可视化对比结果

Fig. 4　Visual results of different models

图 5 图像和谐化结果

Fig. 5 Results of image harmonization

3.2.2 定量实验结果

本节还对本方法与 2 个基准模型 - Stable Diffusion 和 Paint By Example 进行了定性比较。实验中,本方法的条件控制强度均设为 1。对于 Stable Diffusion,本方法使用 BLIP 为背景图像生成描述文本并作为提示词以获取和谐化结果;对于 Paint By Example,以前景实例的区域标注作为其掩码输入。

不同方法在 Open Dataset 的 *PSNR/MSE/FID* 定量指标见表 1。Stable Diffusion 通过提示词提取实例图像的基本信息,仅能对前景实例的类别和颜色等属性做出简单概括,但因提示词的局限性,难以完整表达前景实例的详细内容。而 Paint By Example 采用 CLIP 的图像编码器,将前景实例的内容信息编码至图像特征空间。然而,由于 Paint By Example 在图像信息表示向量上的维度限制(1 024),图像中的大部分细节信息在过程中被丢失。相比之下,本方法通过对图像高频和低频特征的有效利用,在和谐化图像中最大程度地保留了前景实例的丰富信息。

表 1 不同方法在 Open Dataset 的 *PSNR/MSE/FID* 定量指标

Table 1 Quantitative results of *PSNR/MSE/FID* on testing sets

方法	*FID* ↓	*PSNR* /dB ↑	*MSE* ↓
Stable Diffusion	18.93	30.58	61.09
Paint By Example	17.52	30.16	65.99
本文方法	**11.79**	**31.26**	**53.08**

表 2 进一步展示了控制强度(λ_h 与 λ_l)对图像生成质量的影响。为方便测试,每组实验的控制强度 λ_h 与 λ_l 设置为相同值(即 $\lambda_h = \lambda_l$,简记为 λ_c),此简化不会影响实验的有效性。由表 2 可见,尽管训练时控制强度设置为恒定值 1,测试阶段的最佳效果并不一定在相同强度下实现。这一发现提示,通过合理调节 λ_c,生成图像质量可进一步优化。

表 2 条件控制强度 λ_c 对图像生成质量的影响

Table 2 Impact of conditional control strength λ_c on image generation quality

λ_c	*FID* ↓	*PSNR* /dB ↑	*MSE* ↓	λ_c	*FID* ↓	*PSNR* /dB ↑	*MSE* ↓
0.1	18.10	30.59	61.01	0.6	**11.31**	31.28	54.23
0.2	17.29	30.28	63.13	0.7	11.88	**31.29**	**53.07**
0.3	15.55	30.83	59.89	0.8	11.71	31.23	53.17
0.4	12.95	30.93	57.47	0.9	11.80	31.23	53.10
0.5	12.35	31.19	54.15	1.0	11.79	31.26	53.08

3.3　消融实验

针对高频和低频控制强度的设置进行了消融实验，具体探讨了高频控制强度 λ_h 和低频控制强度 λ_l 对图像和谐化效果的影响，如图 6 所示。实验结果表明，随着高频控制强度 λ_h 的增加，合成结果中的空间结构逐渐与真实图像的前景实例结构一致。然而，当缺少低频信息的控制时，和谐化图像中的实例颜色难以与目标图像相匹配。类似地，当低频条件缺失于图像合成过程中时，合成图像在结构上难以与输入实例保持一致。由此可见，同时应用高频和低频条件信息可显著提升模型的生成效果和一致性。

图 6　高低频条件控制强度对生成效果的影响

Fig. 6　Effects of high and low frequency conditional control strength

除了利用高低频特征分解对图像合成结果进行联合控制外，本方法还对不同输入信号形式进行了消融实验。具体而言，研究对比了直接使用原始图像信息作为输入条件的效果，结果如图 7 所示。由于在获取组合图像时可能对前景实例进行插值操作，直接使用原始图像作为生成条件往往无法实现对质量较差的前景实例的细节增强。因此，本方法基于高低频分解的图像和谐化方案在生成效果上具备更高的鲁棒性，能够有效地处理各类前景实例的细节和色彩问题。

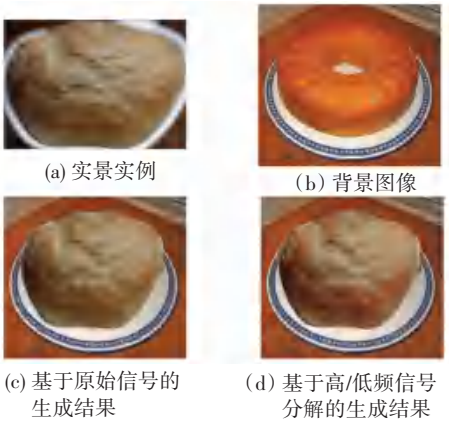


图 7　基于不同输入条件的生成结果

Fig. 7　Visual results of different input conditions

4　结束语

在大规模扩散预训练模型的推动下，基于 Stable Diffusion 的图像填充模型展示了极高的生成真实性。然而，该模型的填充内容主要依赖于文本描述，难以捕捉前景图像实例的细节特征，因为前景实例的许多细微特征难以通过自然语言进行准确表达。针对该问题，本文提出从前景实例中提取边缘信息作为高频特征，并使用高斯模糊生成低频特征，以此分别保留前景实例的结构和颜色信息。采用 ControlNet 框架，分别以高频和低频特征为控制条件，训练出多层控制的图像填充模型。在推理阶段，通过特征加权和的方式有效融合高频和低频信息，以确保生成图像能完整地呈现前景实例的特征信息。为验证模型的有效性，设计了 2 组基准实验，从定性和定量角度对生成效果进行评价，结果显示所提出的方法在图像实例和谐化问题上具有显著优势。此外，本文还开展了消融实验，进一步探讨了条件控制权重和输入条件形式对模型性能的影响。实验结果表明，本文设计的方法合理有效，在和谐化质量上表现优异。

参考文献

[1] SEONI S, SHAHINI A, MEIBURGER K M, et al. All you need is data preparation: A systematic review of image harmonization techniques in Multi - center/device studies for medical support systems[J]. Computer Methods and Programs in Biomedicine, 2024,250: 108200.

[2] SUNKAVALLI K, JOHNSON M K, MATUSIK W, et al. Multi-scale image harmonization[J]. ACM Transactions on Graphics (TOG), 2010, 29(4): 1-10.

[3] TSAI Y H, SHEN Xiaohui, LIN Zhe, et al. Deep image harmonization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3789-3797.

[4] GUO Zonghui, GUO Dongsheng, ZHENG Haiyong, et al. Image harmonization with transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 14870-14879.

[5] CONG Wenyan, ZHANG Jianfu, NIU Li, et al. Dovenet: Deep image harmonization via domain verification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8394-8403.

[6] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[7] LU Lingxiao, LI Jiangtong, CAO Junyan, et al. Painterly image harmonization using diffusion model[J]. arXiv preprint arXiv, 2308.02228,2023.

[8] REZENDE D J, MOHAMED S. Stochastic backpropagation and approximate inference in deep generative models[J]. arXiv preprint arXiv,1401.4082, 2014.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Montreal, Canada: NIPS Foundation, 2014: 2672- 2680.

[10] RAMESH A, GOYAL Y, LIANG Z, et al. Diffusion models beat GANs on image synthesis[J]. International Conference on Machine Learning, 2021,11: 8447-8457.

[11] 张彬, 周粤川, 刘杨, 等. 基于多级残差映射器的文本驱动人脸图像生成和编辑[J]. 软件学报, 2023, 34(5): 2101-2115.

[12] 谈馨悦, 何小海, 王正勇. 基于 Transformer 交叉注意力的文本生成图像技术[J]. 计算机科学, 2022, 49(2): 107-115.

[13] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE,2019: 4401-4410.

[14] KARRAS T, LAINE S, HELLSTEN J, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8110-8119.

[15] ABDAL R, QIN Yipeng, WONKA P. Image2stylegan: How to embed images into the stylegan latent space? [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019:4431-4440.

[16] TOV O, ALALUF Y, NITZAN Y, et al. Designing an encoder for stylegan image manipulation[J]. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-14.

[17] RICHARDSON E, ALALUF Y, PATASHNIK O, et al. Encoding in style: A stylegan encoder for image-to-image translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2287-2296.

[18] 龚颖, 许文韬, 赵策, 等. 生成对抗网络在图像修复中的应用综述[J]. 计算机科学与探索, 2024, 18(3):553-573.

[19] BROCK A, DONAHUE J, SIMONYAN K. Large scale GaN training for high fidelity natural image synthesis[J]. arXiv preprint arXiv, 1809.11096,2018.

[20] RAMESH A, LIU M, FISCHER P, et al. Dall·e 2: Exploring cross-modal embed dings for multi-modal generation[J]. arXiv preprint arXiv,2110.11487, 2021.

[21] NICHOL A, DHARIWAL P, RAMESH A, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[J]. arXiv preprint arXiv, 2112.10741, 2021.

[22] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10684-10695.

[23] PITIE F, KOKARAM A C, DAHYOT R. N-dimensional probability density function transfer and its application to color transfer[C]//IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2005: 1434-1439.

[24] REINHARD E, ADHIKHMIN M, GOOCH B, et al. Color transfer between images[J]. IEEE Computer Graphics and Applications, 2001, 21(5): 34-41.

[25] COHEN-OR D, SORKINE O, GAL R, et al. Color harmonization[M]. ACM Transactions on Graphics, 2006: 624-630.

[26] CONG Wenyan, NIU Li, ZHANG Jianfu, et al. Bargainnet: Background-guided domain translation for image harmonization[C]//Proceedings of 2021 IEEE International Conference on Multi-Media and Expo. Piscataway, NJ: IEEE, 2021: 1-6.

[27] SOFIIUK K, POPENOVA P, KONUSHIN A. Foreground-aware semantic representations for image harmonization[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 1620-1629.

[28] PENG Jinlong, LUO Zekun, LIU Liang, et al. Frih: Fine-grained region-aware image harmonization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4478-4486.

[29] PARK T, LIU Mingyu, WANG Tingchun, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2337-2346.

[30] LIU Xihui, YIN Guojun, SHAO Jing, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis[J]. Advances in Neural Information Processing Systems, 2019, 32: 570-580.

[31] WANG Yi, QI Lu, CHEN Yingcong, et al. Image synthesis via semantic composition[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 13749-13758.

[32] TANG Hao, XU Dan, YAN Yan, et al. Local class-specific and global image-level generative adversarial networks for semantic-

guided scene generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 7870-7879.

[33] LI Yuheng, LI Yijun, LU Jingwan, et al. Collaging class-specific gans for semantic image synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 14418-14427.

[34] ZHANG Lvmin, AGRAWALA M. Adding conditional control to text-to-image diffusion models[J]. arXiv preprint arXiv, 2302.05543, 2023.

[35] LI Junnan, LI Dongxu, XIONG Caiming, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation [C]// International Conference on Machine Learning. Vienna, Austria : PMLR, 2022: 12888-12900.

[36] CHEN Wengling, HAYS J. SketchyGAN: Towards diverse and realistic sketch to image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9416-9425.

[37] LEE J, KIM E, LEE Y, et al. Reference-based sketch image colorization using augmented self-reference and dense semantic correspondence[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020:5800-5809.

[38] MOU Chong, WANG Xintao, XIE Liangbin, et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models[J]. arXiv preprint arXiv, 2302.08453, 2023.

[39] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[J]. International Journal of Computer Vision, 2020, 128(4): 1037-1052.

[40] GOYAL P, DOLLÁR P, GIRSHICK R, et al. Accurate, large minibatch SGD: Training imagenet in 1 hour[J]. arXiv preprint arXiv, 1706.02677, 2017.

[41] HUYNH-THU Q, GHANBARI M. The scope of validity of PSNR in image/video quality assessment[J]. Electronics Letters, 2008, 44(13): 800-801.

[42] HEUSEL M, RAMSAUER H, UNTERTHINER T, 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, 30:6627-6638.

[43] YANG Binxin, GU Shuyang, ZHANG Bo, et al. Paint by example: Exemplar-based image editing with diffusion models [J]. arXiv preprint arXiv, 2211.13227, 2022.