InterLCM：低质量图像作为潜在一致性模型中间态的高效盲人脸修复方法

李森茂 [1,2*] **王凯** [2†] **Joost van de Weijer**[2] **Fahad Shahbaz Khan**[3,4]
郭春乐 [1,6] **杨诗琪** [5] **王亚星** [1,6] **杨健** [1] **程明明** [1,6]
[1]VCIP, CS, 南开大学 [2] 计算机视觉中心, 巴塞罗那自治大学
[3] 穆罕默德·本·扎耶德人工智能大学 [4] 林雪平大学 [5]SB Intuitions, 软银
[6] 南开国际深圳研究院（深圳福田）, 南开大学
{senmaonk,shiqi.yang147.jp}@gmail.com, {kwang,joost}@cvc.uab.es
fahad.khan@liu.se, {guochunle,yaxing,csjyang,cmm}@nankai.edu.cn

Abstract

目前已通过在图像修复数据集上恢复低质量图像来微调扩散模型（DMs），将扩散先验应用于盲人脸修复（BFR）。然而，直接应用 DMs 存在一些显著局限：（i）扩散先验的语义一致性较差（如身份特征、结构与色彩等），增加了 BFR 模型的优化难度；（ii）依赖数百次去噪迭代，阻碍了与感知损失的有效协同，其中感知损失对真实修复至关重要。我们发现潜在一致性模型（LCM）在 ODE 轨迹上学习到了噪声到数据的一致性映射，因此在身份特征、结构信息和色彩保持方面展现出更强的语义一致性。为此，我们提出 *InterLCM* ，利用潜在一致性模型的语义一致性与高效性解决上述问题。通过将低质量图像视为潜在一致性模型的中间态，*InterLCM* 能够从更早的潜在一致性模型步骤开始，实现保真度与质量的平衡。潜在一致性模型还允许在训练中融合感知损失来提升修复质量，特别是在真实世界场景中会生成更好的效果。为降低结构与语义不确定性，*InterLCM* 引入一个视觉模块和一个空间编码器，视觉模块用来提取视觉特征，空间编码器用来捕获空间细节，能够增强图像修复保真度。大量实验表明，*InterLCM* 能够在更快的推理速度下，在合成与真实数据集上的效果均超越现有方法。项目主页：https://sen-mao.github.io/InterLCM-Page/

1 引言

盲人脸修复（BFR）旨在从具有复杂未知退化的低质量（LQ）输入中恢复高质量（HQ）图像，这些退化包括：下采样 (Chen et al., 2018; Bulat et al., 2018)、模糊 (Zhang et al., 2017; 2020; Shen et al., 2018)、噪声 (Dogan et al., 2019) 以及压缩损失 (Dong et al.,

*Work done during a research stay at Computer Vision Center, Universitat Autònoma de Barcelona.
†The corresponding author.

图 1: (左图)4 步潜在一致性模型和 SD Turbo 模型的中间状态。潜在一致性模型所使用的网络直接映射到真实图像空间，SD Turbo 则采用渐进式去噪方法处理噪声图像。(右图）给定提示文本 “一张头像，画着一位戴着帽子和眼镜的男性”，我们分别用潜在一致性模型和 SD Turbo 模型生成 1000 张图像，随后采用 DreamSim 相似度、结构相似性指数（SSIM）和色彩直方图距离（HDist）量化评估生成结果在身份特征、空间结构和色彩保持三个维度的语义一致性。

2015) 等。近年来，盲人脸修复领域取得了显著进展。现有方法主要集中于学习低质量到高质量图像的直接映射，通常结合各种先验知识以提升修复性能。早期研究主要利用几何先验，如面部关键点 (Chen et al., 2018)、解析图 (Chen et al., 2021; Shen et al., 2018) 和热力图 (Yu et al., 2018)，为脸部修复提供显式指导信息。基于指导先验的方法 (Gu et al., 2022; Zhou et al., 2022) 则引入额外高质量图像来增强低质量图像修复效果。最近，生成先验 (Wang et al., 2021a; Yang et al., 2021) 被广泛应用于盲人脸修复，以获得更真实的纹理细节。

随着基于数十亿级数据训练 (Schuhmann et al., 2022) 的先进扩散模型 (Ramesh et al., 2022) 展现出卓越的生成能力，扩散先验方法 (Wang et al., 2023; Miao et al., 2024; Lu et al., 2024) 已被探索用于解决 BFR 问题。虽然已经取得了合理结果，但现有基于扩散的方法 (Wang et al., 2021a; Yue & Loy, 2024) 存在几个关键局限：（i）扩散先验在语义一致性、身份一致性、结构稳定性和色彩保持等方面表现欠佳，这增加了 BFR 模型 (Zhou et al., 2022) 的优化难度。举例来说，我们对比了传统扩散模型 SD Turbo (Sauer et al., 2023)[1]与潜在一致性模型（LCM） (Luo et al., 2023a) 在各步骤生成图像时的语义一致性，如 Fig. 1所示，结果表明传统扩散模型相比一致性模型具有更弱的语义先验信息。(ii）依赖标准扩散模型的基于扩散的方法在取样方面存在挑战，因为它们需要多次迭代来产生真实图像输出，难以有效结合针对最终输出的感知损失计算。尽管现有方法 (Chung et al., 2023; Laroche et al., 2024) 尝试在中间步骤获取真实图像来计算感知损失，但这些图像与最终输出存在显著外观差异（详见 Appendix E.6)。

为解决上述问题，我们首次将潜在一致性模型（LCM）引入盲人脸修复任务。具体而言，潜在一致性模型学习将 ODE 轨迹 (Song et al., 2023) 上的任意点映射至其原点以实现生成建模，该特性显著区别于传统扩散模型，后者通过迭代采样过程逐步消除随

[1]我们将 SD Turbo 视为扩散模型的典型代表，因其完整继承了扩散模型特性；同时潜在一致性模型通过一致性正则化进行蒸馏。

机初始向量中的噪声。基于潜在一致性模型的特性，我们提出 *InterLCM* 方法，将低质量（LQ）图像视为 *LCM* 模型的中间状态输入，然后执行 4 步潜在一致性模型中的剩余少量去噪步骤（如 3 步）即可获得高质量输出。通过这种方式，*InterLCM* 保持了源自潜在一致性模型的更好语义一致性。同时，得益于这一特性，我们可以整合感知损失 (Johnson et al., 2016) 和对抗损失 (Goodfellow et al., 2014) 这两种修复模型训练中常用的损失函数，从而获得高质量、高保真的人脸修复结果。

然而，直接将潜在一致性模型应用于盲人脸修复会给生成的结构和语义带来随机性，这种随机性源自随机采样路径（详见 Sec. 3.2和 Fig. 5）。我们因此在 *InterLCM* 中引入了两个额外组件。首先，采用 CLIP 图像编码器和视觉编码器作为视觉模块，帮助从人脸中提取语义信息，为潜在一致性模型提供人脸特异性先验。其次，为了防止内容（如结构）发生变化，我们加入了空间编码器以利用潜在一致性模型的强大语义一致性。具体来说，我们遵循 ControlNet 架构设计，复制 UNet 编码器部分作为空间编码器。需要注意的是，空间编码器在训练方案上与 ControlNet 有所不同：ControlNet 通常使用扩散损失进行训练，而我们的空间编码器通过从真实世界图像（经过去噪步骤）到初始低质量图像的梯度反向传播进行训练。在此过程中，视觉编码器和空间编码器都会随着梯度更新。

在实验中，我们在 CelebA、LFW、WebPhoto 等合成和真实世界数据集上进行了大量实验，将 *InterLCM* 与现有方法进行比较。我们的方法在获得更好定性和定量性能的同时，同时还实现了更快的推理速度。总的来说，我们的工作贡献如下：

- 我们提出了 *InterLCM* ，一个简单但有效的 BFR 框架，能够利用了潜在一致性模型（LCM）的先验。通过将低质量图像视为潜在一致性模型的中间状态，我们能够有效地在面部修复中保持更好的语义一致性。
- 利用潜在一致性模型将每个状态映射到原始图像级别的特性，我们的方法 *InterLCM* 具有额外优势：只需少量采样步骤即可获得更快的速度，并且可以将我们的框架与面部修复中常用的感知损失和对抗损失相结合。
- 通过在合成和真实图像数据集上的大量实验，我们证明了 *InterLCM* 在恢复高质量图像方面的有效性和真实性，特别是在具有不可预测退化的真实世界场景中。

2 相关工作

2.1 盲人脸修复。

在真实世界场景中，人脸图像可能遭受多种类型的退化，如噪声、模糊、下采样、JPEG 压缩伪影等。盲人脸修复（BFR）旨在从遭受未知退化的低质量图像中恢复高质量人脸图像。现有 BFR 方法主要聚焦于探索更好的人脸先验，包括几何先验、参考先验和生成先验。近年来备受关注的扩散先验属于更广泛的生成先验范畴。基于几何先验的方法主要利用人脸图像中的高度结构化信息。这些结构信息：例如面部关键点 (Chen et al., 2018)、人脸解析图 (Shen et al., 2018; Chen et al., 2021) 和 3D 形状 (Hu et al., 2020;

Zhu et al., 2022; Lu et al., 2024)，可作为指导信息辅助修复。然而，由于从退化输入估计的几何先验可能不可靠，往往导致后续 BFR 任务性能欠佳。部分现有方法 (Dogan et al., 2019; Li et al., 2018) 采用与退化图像同身份的高质量参考图像来指导修复，称为参考先验 BFR 方法。这类方法的主要局限在于对高质量参考图像的依赖，而这些图像在某些场景中难以获取。最新研究直接利用生成模型中封装的丰富先验进行 BFR，称为生成先验方法。

GAN 先验方法 通过应用 GAN 反演技术 (Xia et al., 2022)，早期生成先验研究 (Gu et al., 2020; Menon et al., 2020) 通过迭代优化预训练 GAN 的潜在 code 来获得理想的高清目标图像。为规避耗时的优化过程，部分研究 (Yang et al., 2021; Chan et al., 2021) 直接将预训练 StyleGAN (Gal et al., 2021) 的解码器嵌入 BFR 网络，显著提升了修复的表现。VQ-GAN (Crowson et al., 2022) 在图像生成领域中的成功也启发多项 BFR 研究来设计不同策略 (Wang et al., 2022; Zhou et al., 2022)，以改进退化输入与潜在高质量图像间码本元素的匹配效果。

扩散先验方法 最近，扩散模型已被证明相比 GAN (Dhariwal & Nichol, 2021) 具有更好的稳定性，图像生成也更加多样化，这一优势也引起了盲人脸修复领域的关注。IDM (Zhao et al., 2023) 在 SR3 (Saharia et al., 2022) 基础上引入外部预清洗流程以提升 BFR 性能。为加速推理，LDM (Rombach et al., 2022) 提出在潜空间训练扩散模型。为规避耗时耗力的重新训练过程，多项研究 (Lin et al., 2023; Wang et al., 2023) 探索了将预训练扩散模型作为生成先验用于修复任务。具体而言，DiffBIR (Lin et al., 2023) 和 SUPIR (Yu et al., 2024) 利用预训练 Stable Diffusion (Rombach et al., 2022) 作为生成先验，相比现有方法能提供更丰富的先验知识。DR2 (Wang et al., 2023) 和 CCDF (Chung et al., 2022) 先将输入图像扩散至噪声状态（此时各类退化的强度低于添加的高斯噪声），随后在去噪过程中捕获语义信息。这类基于噪声状态 (Wang et al., 2023; Chung et al., 2022) 或扩散桥 (Liu et al., 2023) 的修复方法可显著加速推理。这些方法的共同核心是通过引入明确定义或人工假设的退化模型作为额外约束，来修改预训练扩散模型的反向采样过程。尽管在理想场景下表现良好，但由于盲人脸修复任务的退化模型未知且复杂，这些方法难以直接适用。然而，这些扩散先验方法仍面临着由于需要多步迭代导致推理耗时的问题。其次，它们大多主要继承自潜在扩散模型的重建损失进行训练，难以有效整合图像修复任务中常用的感知损失，从而导致生成结果的感知质量欠佳。

2.2 文生图模型

扩散模型 (Shonenkov et al., 2023; Ho et al., 2022; Chen et al., 2023) 已成为文生图领域的新一代最先进的模型。这类模型通常采用预训练语言编码器，例如 CLIP (Radford et al., 2021) 和 T5 (Raffel et al., 2020) 对文本提示进行编码，随后通过交叉注意力机制将编码结果输入扩散模型。在基础架构方面，广泛采用 UNet (Ronneberger et al., 2015) 和 DiT (Peebles & Xie, 2023)。本文主要采用文生图模型中的一个代表性强大模型：Stable Diffusion (Rombach et al., 2022) 来构建我们的方法。

文生图模型蒸馏技术。 扩散模型的生成速度受限于其缓慢的采样过程。近年来，基于蒸馏的技术 (Hinton et al., 2014) 被广泛应用于加速扩散模型。从预训练教师模型蒸馏的学生模型 (Luo et al., 2023a; Sauer et al., 2023) 通常具有更快的推理速度。早期研究 (Salimans & Ho, 2022; Meng et al., 2023) 采用渐进式蒸馏逐步减少学生扩散模型的采样步数。此外，预训练教师模型的采样时间也制约着训练效率。为解决这一限制，多项工作 (Gu et al., 2023; Nguyen & Tran, 2023) 提出了不同的引导技术：例如 Boot (Gu et al., 2023) 基于连续两个采样步骤进行引导训练，实现了无需真实图像的蒸馏。SDXL-Turbo (Sauer et al., 2023) 则引入判别器并与分数蒸馏损失相结合。

文生图模型的附加图像控制 虽然文本描述能引导扩散模型生成图像，但难以实现对生成结果的细粒度控制。细粒度控制信号具有多种模态，包括布局图、分割图、深度图等。鉴于文生图模型强大的生成能力，已有多种方法 (Li et al., 2024a; Zavadski et al., 2023; Lin et al., 2024) 致力于为文生图模型添加图像控制功能。作为典型代表，ControlNet (Zhang et al., 2023) 提出使用文生图扩散模型中 UNet 编码器的可训练副本对附加条件信号进行编码，通过零卷积将编码结果注入扩散模型 UNet 主干。这种简洁但是高效的设计在空间控制方面展现出泛化稳定的性能，因而被广泛应用于各类下游任务。类似地，T2I-Adapter (Mou et al., 2024) 训练了一个额外的控制编码器，将中间表示添加到 Stable Diffusion 预训练编码器的特征图中。

然而，现有附加图像条件的文生图模型仍从高斯噪声开始生成图像。如何探索其在图像修复任务中的应用潜力尚未得到充分研究。本文创新性地实现了从退化低质量图像出发生成高质量图像的修复方法，并将其与加速版文生图模型结合。

3 方法

盲人脸修复（BFR）旨在从未知复杂退化的低质量（LQ）图像中恢复高质量（HQ）图像，同时保持与低质量图像的语义一致性。在本节中，我们首先在 Sec. 3.1中介绍潜在扩散模型和潜在一致性模型的基础知识，随后在 Sec. 3.2中详述我们提出的 *InterLCM* 方法。遵循潜在一致性模型的加噪过程，我们首先在 *InterLCM* 中研究低质量图像应被视为潜在一致性模型的何种中间状态。接着引入视觉模块和空间编码器来保持重建高质量图像中的语义信息和结构完整性。*InterLCM* 的详细说明参见 Appendix B中的 Fig. 3与 Algorithm 1。

3.1 预备知识

潜在扩散模型。 为了让扩散模型（DM）在有限的计算资源保持高生成质量地训练，一系列潜在扩散模型（LDMs）(Rombach et al., 2022) 把一张图片 x 通过编码器 $\mathcal{E}$ 编码至潜在表征 z_0，且使用解码器 $\mathcal{D}$ 重建这张图片。潜在扩散模型的目的是通过扩散损失训练一个噪声预测网络 ϵ_θ：

$$\mathcal{L} = \mathbb{E}_{z_0,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(z_t, c, t)\|_2^2 \tag{1}$$

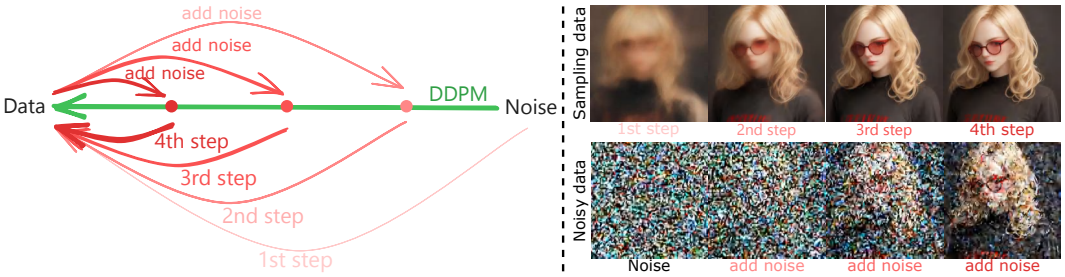

图 2: （左图）4 步潜在一致性模型在各采样步骤的映射过程：噪声 $\xrightarrow{\text{第一步}}$ 采样数据 $\xrightarrow{\text{添加噪声}}$ 含噪数据 $\xrightarrow{\text{第二步}}$ 采样数据 $\xrightarrow{\text{添加噪声}}$ 含噪数据 $\xrightarrow{\text{第三步}}$ 采样数据 $\xrightarrow{\text{添加噪声}}$ 含噪数据 $\xrightarrow{\text{第四步}}$ 采样数据。第一步，从随机噪声预测原始图像，后续每一步都在前一步生成的原始图像上添加噪声。（右图）展示各步骤预测的原始图像（第一行），第一步至第三步的随机噪声与含噪数据（第二行）。举例来说，给定提示词 “戴红色眼镜穿黑色衬衫的金发女性”，可以注意到各步骤生成图像在身份特征、结构信息和色彩保持方面均保持语义一致性。

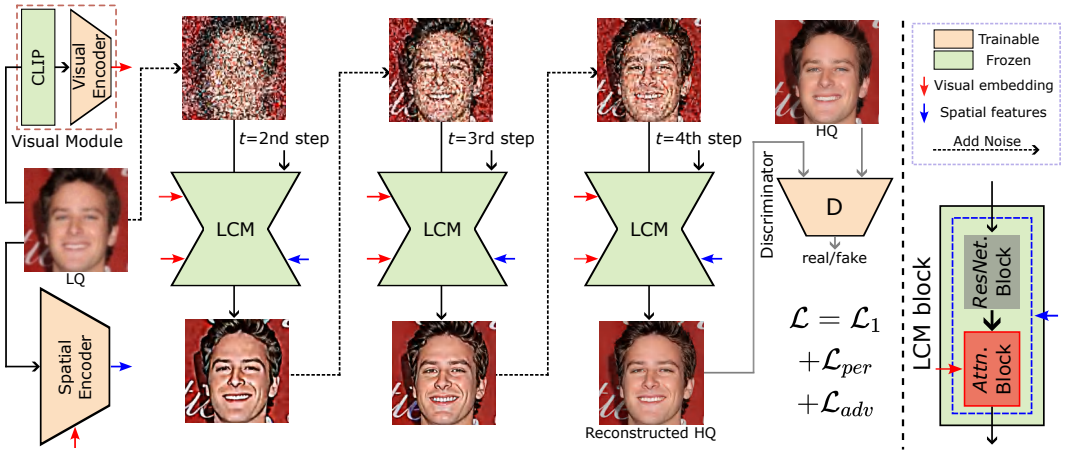

图 3: 我们提出的 *InterLCM* 框架概览。视觉模块处理低质量图像生成视觉嵌入，空间编码器提供结构信息。我们将低质量图像视为潜在一致性模型的中间状态，通过结合视觉嵌入和空间特征的标准潜在一致性模型条件化处理，可将低质量输入重建为高质量图像。

在扩散推理阶段，潜在扩散模型利用文本条件 c，通过预训练的去噪网络 $\boldsymbol{\epsilon_\theta}(z_t, c, t)$ 预测噪声，并遵循 DDPM 调度器 (Ho et al., 2020) 获得潜在表征 z_{t-1}（参见 Fig. 2（左图），绿色箭头所示）。通过该迭代过程最终获得潜在表示 z_0。

潜在一致性模型。 一致性模型（CMs）(Song et al., 2023) 采用一致性映射，直接将 ODE 轨迹上的任意点映射回其原点，相比潜在扩散模型更能促进语义一致性生成。潜在一致性模型 $\boldsymbol{f}_\theta(z_{\tau_n}, c, \tau_n)$ 能够通过一致性蒸馏损失 (Song et al., 2023) 从预训练的潜在扩散模型（如 Stable Diffusion (Rombach et al., 2022)）中蒸馏，来实现少步推理，其中 c 为给定文本条件。潜在一致性模型直接预测增强 PF-ODE 轨迹 (Luo et al., 2023a) 的起点 z_0，实现单步采样。潜在一致性模型通过交替执行去噪和加噪步骤（参见 Fig. 2

（左图）多条红色箭头），在保持语义一致的同时提升样本质量。具体来说，在第 n 次迭代中：首先对前次预测样本 $z_0 = \boldsymbol{f}_\theta(z_{\tau_{n+1}}, c, \tau_{n+1})$ 应用加噪前向过程，得到 z_{τ_n}。这里 τ_n 表示递减时间步序列，其中 $n \in \{1, \cdots, N-1\}$，$\tau_1 > \tau_2 > \cdots > \tau_{N-1}$，且 N（$N = 4$）为潜在一致性模型的总步数。随后再次预测下一个 $z_0 = \boldsymbol{f}_\theta(z_{\tau_n}, c, \tau_n)$。

3.2 *InterLCM*：低质量图像作为潜在一致性模型的中间状态

我们提出的 *InterLCM* 基于潜在一致性模型。如 Fig. 3所示，在已包含复杂未知退化的低质量图像 x_l 上添加随机噪声。视觉模块以低质量图像作为输入并返回视觉嵌入，其将替代标准潜在一致性模型中的文本嵌入以提供人脸特异性语义信息。为保持低质量图像的结构信息，我们采用空间编码器为潜在一致性模型提供结构特征。通过结合视觉嵌入和空间特征的标准潜在一致性模型，低质量输入便可重建为高质量输出。在本小节中，根据潜在一致性模型加噪过程，我们首先研究确定低质量图像应插入的潜在一致性模型中间状态，随后详细说明视觉模块与空间编码器的设计。

第二步中间状态。 为了利用潜在一致性模型 (Luo et al., 2023a) 中固有的内容一致性，我们保留预训练模型并沿用其采样流程。如图 Fig. 2（右图，第一行）所示，4 步潜在一致性模型采样流程能够生成语义一致的图像。在第一步中，潜在一致性模型直接从随机噪声中预测生成图像。在后续每一步中，潜在一致性模型首先对前一阶段的图像添加噪声，据此预测更精细的输出。基于 4 步潜在一致性模型中每一步的三次噪声添加过程，我们首先将低质量图像添加到潜在一致性模型的每个中间状态。如图 Fig. 4所示，我们实证发现低质量图像的分布更接近于第一次添加噪声后所生成的图像（第二步添加噪声）而非其他中间状态（详细信息请参见 Appendix C.1）。因此，我们将低质量图像作为潜在一致性模型第一次噪声添加后的中间状态。随后，从第二步开始应用潜在一致性模型。

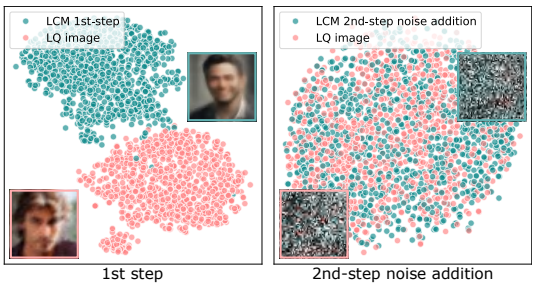

图 4: t-SNE 特征分布可视化展示了潜在一致性模型第一步采样结果与低质量图像的相似性（FID=103.70），与经过潜在一致性模型的第二步噪声扩散后它们的含噪中间状态的相似性（FID=2.83）。

视觉编码器。 理想情况下，模型应同时重建图像质量并与低质量图像保持语义对齐。然而，噪声扩散过程会引入随机性，无论使用空文本提示或文本提示，都会改变低质量图像的原始语义。如图 Fig. 5所示，当给定低质量图像和空文本提示（$\varnothing$ =“”）时，生成图像中的发色会变为白色（如

图 5: 改变了低质量图像的原始语义的标准潜在一致性模型（例如头发）。

图 Fig. 5（第二列））。即使使用从高质量图像获得的文本提示（如“金发微笑的女性”[2]），生成图像中的直发仍会变为卷发（如图 Fig. 5（第三列））。

为提供给潜在一致性模型人脸特异的先验知识来生成语义一致的内容，我们提出使用视觉模块（ Fig. 3）来提供。该视觉模块为预训练的潜在一致性模型提供人脸特异性语义信息，类似于标准文本条件图像生成中的文本提示 (Luo et al., 2023a)。我们采用视觉嵌入，首先从低质量图像 x_l 中提取总的 CLIP 视觉特征 (Radford et al., 2021)，其通过视觉编码器（VE）蒸馏得到人脸特异性语义信息，定义为 $c_v = VE(CLIP(x_l))$。这种方法使 c_v 与潜在一致性模型通常用于文本条件采样的文本嵌入对齐。此外，使用视觉嵌入避免了需要应用复杂文本提示来详细准确描述低质量图像的问题 (Liao et al., 2024; Li et al., 2024b)。

空间编码器。尽管人脸特异性视觉嵌入 c_v 对于捕捉全局语义属性至关重要，但其在保持全局结构方面存在不足（见 Fig. 8中的 ①）。为解决这一问题，我们引入空间编码器（SE）来有效提取并增强空间结构保持能力（Fig. 3）。我们使用 Stable Diffusion 中预训练的 UNet 编码器来捕获包括结构信息的低质量图像中的全部内容。当与视觉嵌入结合时，空间编码器随后提取空间特征，记为 $\mathsf{f}_v = SE(x_l, c_v)$。其中，*ResNet* 和 *Attn* 模块分别代表潜在一致性模型中的标准 ResNet 和交叉注意力 Transformer 模块。*ResNet* 模块的输出作为 *Attn* 模块中的 Query 特征，而视觉嵌入 c_v 则作为其中的 Key 和 Value 特征。

然后空间特征与 *Attn* 模块输出相结合，经过三次潜在一致性模型采样迭代，最终生成重建的高质量图像 $x_{rec} = \mathcal{D}$

训练目标。为训练视觉编码器与空间编码器，我们采用三种图像级损失函数：重建损失 $\mathcal{L}_1$、感知损失 (Johnson et al., 2016; Zhang et al., 2018) $\mathcal{L}_{per}$ 以及对抗损失 (Goodfellow et al., 2014; Esser et al., 2021) $\mathcal{L}_{adv}$：

$$\mathcal{L}_1 = \|x_h - x_{rec}\|_1; \quad \mathcal{L}_{per} = \|\Phi(x_h) - \Phi(x_{rec})\|_2^2; \quad \mathcal{L}_{adv} = [\log D(x_h) + \log(1 - D(x_{rec}))],$$

其中 x_h 表示高质量图像，Φ 代表 VGG19 (Simonyan & Zisserman, 2014) 的特征提取器。模型的完整目标函数为：

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{per} + \lambda \mathcal{L}_{adv}, \tag{2}$$

其中 λ 为权衡参数，在后续实验中默认设置为 0.1。

4 实验

4.1 在合成与真实世界数据集上的评估

我们在一个合成数据集和三个真实世界数据集上评估我们的模型，这些数据集都常用于评估盲人脸恢复任务 (Wang et al., 2021a; Zhou et al., 2022; Yue & Loy, 2024; Yang et al., 2024)。

[2]我们使用 BLIP (Li et al., 2022) 描述生成模型为高质量图像生成文本提示。

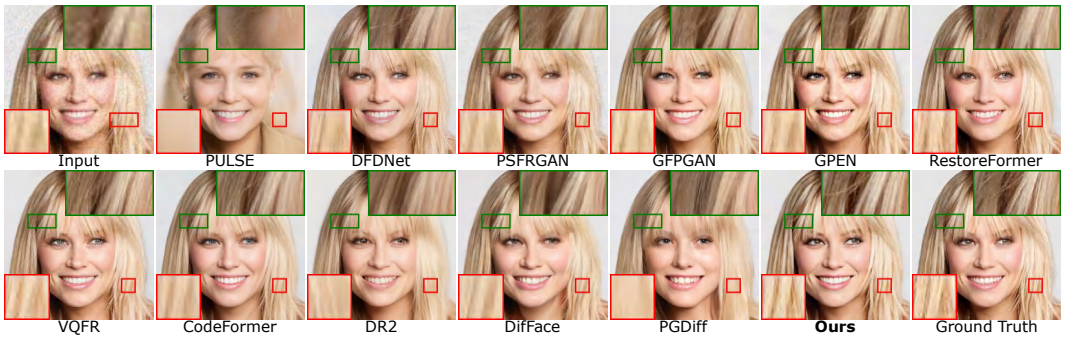

图 6: 对于 BFR 任务在合成数据集 CelebA-Test 上的多基准方法量化对比（放大以便观看，更多结果见 Appendix F）。

我们将我们的方法与现有基准方法进行对比，包括（基于卷积/Transformer 的方法）PULSE (Menon et al., 2020), DFDNet (Li et al., 2020), PSFRGAN (Chen et al., 2021), GFPGAN (Wang et al., 2021a), GPEN (Yang et al., 2021), RestorFormer (Zamir et al., 2022)，VQFR (Gu et al., 2022)，CodeFormer (Zhou et al., 2022)，（基于扩散的方法）DR2 (Wang et al., 2023)，DifFace (Yue & Loy, 2024)，PGDiff (Yang et al., 2024) 和 WaveFace (Miao et al., 2024)。更多细节见 Appendix A。

对于合成数据集（即 CelebA-Test (Karras et al., 2017)）的评估，我们采用五种量化指标：LPIPS (Zhang et al., 2018)、FID (Heusel et al., 2017)、MUSIQ、PSNR 和 SSIM (Wang et al., 2004)，类似于 CodeFormer (Zhou et al., 2022) 使用的指标和 VQFR (Gu et al., 2022) 中使用的 IDS（亦称 Deg）指标。各方法的评估结果汇总于 Tab. 1（第二至第七列）。在图像质量指标 LPIPS 和 MUSIQ **(MUS.)** 方面，本方法 *InterLCM* 较现有方法获得了更优的分数。此外，本方法能忠实保持身份特征和结构完整，如最佳 IDS 和 SSIM 分数所示。Fig. 6进一步表明，本方法显著优于其他方法，而对比方法均未能产生令人满意的修复结果。例如，DFDNet、PSFRGAN、GFPGAN、GPEN、DifFace 和 PGDiff 均引入了明显伪影，而 PULSE 和 DR2 则产生过度平滑而缺失关键面部细节的结果。虽然 RestoreFormer、VQFR 和 CodeFormer 能生成高质量纹理细节（如头发），但仍存在轻微伪影。相比之下，本方法略优于这些方法（参见 Fig. 6中放大区域）。

在真实场景数据集（例如 LFW-Test (Huang et al., 2008)、WebPhoto-Test (Wang et al., 2021a) 和 WIDER-Test (Yang et al., 2016)）的评估中，我们参照 CodeFormer (Zhou et al., 2022) 的设置采用两项量化指标，即 FID 和 MUSIQ。对比实验结果汇总于 Tab. 1（第八至第十三列）。我们观察到本方法在中度和重度退化条件下的 WebPhoto-Test 和 WIDER-Test 数据集上均取得最优性能。同时，它在轻度退化的 LFW-Test 数据集上获得最高的 MUSIQ 评分。如 Fig. 7的定性对比所示，本方法对真实场景退化表现出卓越的鲁棒性，产生最具视觉满意度的结果。即使在严重退化图像中，本方法仍能生成丰富的纹理细节，而对比方法均出现明显伪影。举例来说，如 Fig. 7（第五、六行）显示，在低质量图像存在严重退化时，所有对比方法生成的人脸图像均存在显著伪影，而本方法则能重建出具有丰富头发细节的高质量人脸图像。

表 1: 合成和真实世界数据集上的量化对比。最好的结果**加粗**表示，次好的结果<u>下划线</u>表示。

数据集		合成数据集 Celeba-Test						真实世界数据集 LFW-Test		WebPhoto-Test		WIDER-Test		时间（秒）
	指标 / 方法	LPIPS↓	FID↓	MUSIQ↑	IDS↓	PSNR↑	SSIM↑	FID↓	MUSIQ↑	FID↓	MUSIQ↑	FID↓	MUSIQ↑	
	输入	0.574	145.22	72.81	47.94	22.72	0.706	138.87	26.87	171.63	18.63	201.31	14.22	–
基于卷积/Transformer	PULSE	0.356	68.33	66.46	43.98	22.10	0.592	67.01	65.00	85.69	63.88	70.65	63.01	3.509
	DFDNet	0.332	54.21	72.08	40.44	24.27	0.628	60.28	73.06	92.71	68.50	59.56	62.02	0.438
	PSFRGAN	0.294	54.21	73.32	39.63	24.66	0.661	49.89	73.60	85.42	71.67	85.42	71.50	**0.041**
	GFPGAN	0.230	49.84	73.90	<u>34.56</u>	24.64	0.688	50.36	73.57	87.47	72.08	39.45	72.79	<u>0.059</u>
	GPEN	0.290	63.44	67.52	36.17	**25.48**	<u>0.708</u>	61.04	68.96	99.09	61.10	46.25	62.64	0.109
	RestoreFormer	0.241	50.04	73.85	36.16	24.61	0.660	48.77	73.70	78.85	69.83	50.04	67.83	0.066
	VQFR	0.245	<u>41.84</u>	75.18	35.74	24.06	0.660	51.33	71.74	<u>75.77</u>	72.02	44.09	<u>74.01</u>	0.177
	CodeFormer	<u>0.227</u>	52.94	<u>75.55</u>	37.27	25.15	0.685	52.84	<u>75.48</u>	83.95	<u>74.00</u>	39.22	73.41	0.085
基于扩散	DR2	0.264	54.48	67.99	44.00	25.03	0.617	<u>45.71</u>	71.50	109.24	62.37	48.20	60.28	1.775
	DifFace	0.272	**39.23**	68.87	45.80	24.80	0.684	46.31	69.76	80.86	65.37	37.74	65.02	3.248
	PGDiff	0.300	47.26	71.81	55.90	22.72	0.659	**44.65**	71.74	101.68	67.92	38.38	68.26	14.768
	WaveFace	–	–	–	–	–	–	53.88	73.54	78.01	70.45	<u>37.23</u>	72.89	19.370
	Ours	**0.223**	45.38	**76.58**	**33.64**	<u>25.19</u>	**0.718**	51.32	**76.16**	**75.48**	**75.88**	**35.43**	**76.29**	0.421

表 2: 视觉编码器（VE）和空间编码器（SE）的消融实验，以及中间起始点的对比研究。

	文本嵌入				开始时间步				LFW-Test		WebPhoto-Test		WIDER-Test	
Exp.	VE	Null Text		SE	1st	2nd	3rd	4th	FID↓	MUS.↑	FID↓	MUS.↑	FID↓	MUS.↑
①	✓					✓			69.99	76.11	93.40	75.58	57.66	76.14
②		✓		✓		✓			55.56	76.02	76.06	75.15	37.28	75.68
③			✓	✓		✓			55.07	75.75	77.76	75.30	36.15	75.98
④	✓			✓	✓				54.94	71.50	92.33	72.92	40.72	71.00
⑤	✓			✓		✓			**50.48**	75.06	86.53	73.66	38.71	73.18
⑥	✓			✓			✓		50.59	71.36	77.25	72.01	50.70	70.41
⑦‡	✓			✓		✓			51.32	**76.16**	**75.48**	**75.88**	**35.43**	**76.29**

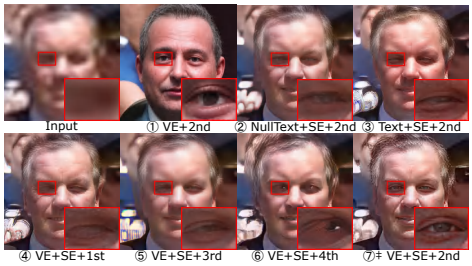

图 8: 各样设计变体方法的消融实验可视化。‡ 展示了我们的结果。

4.2 消融实验

视觉编码器与空间编码器的有效性。本文提出的方法从第二步开始，同时结合了视觉编码器（VE）的视觉嵌入特征和空间编码器（SE）的空间特征。我们首先通过探索不同消融设计并对比它们的性能来评估从第二步开始的视觉嵌入和空间特征的有效性。消融设计包括：① VE+2nd：移除 SE，仅训练 VE；② NullText+SE+2nd：仅训练 SE，VE 替换为空文本；③ Text+SE+2nd：仅训练 SE，VE 替换为文本。性能结果和对比见 Fig. 8（第一行第二至第四列）和 Tab. 2（第一至第三行）。我们注意到① VE+2nd 能够捕捉低质量图像的人脸特异性语义信息并生成高质量细节，但由于视觉嵌入仅提供语义一致性内容，在保持全局结构方面表现不足；② NullText+SE+2nd 和③ Text+SE+2nd

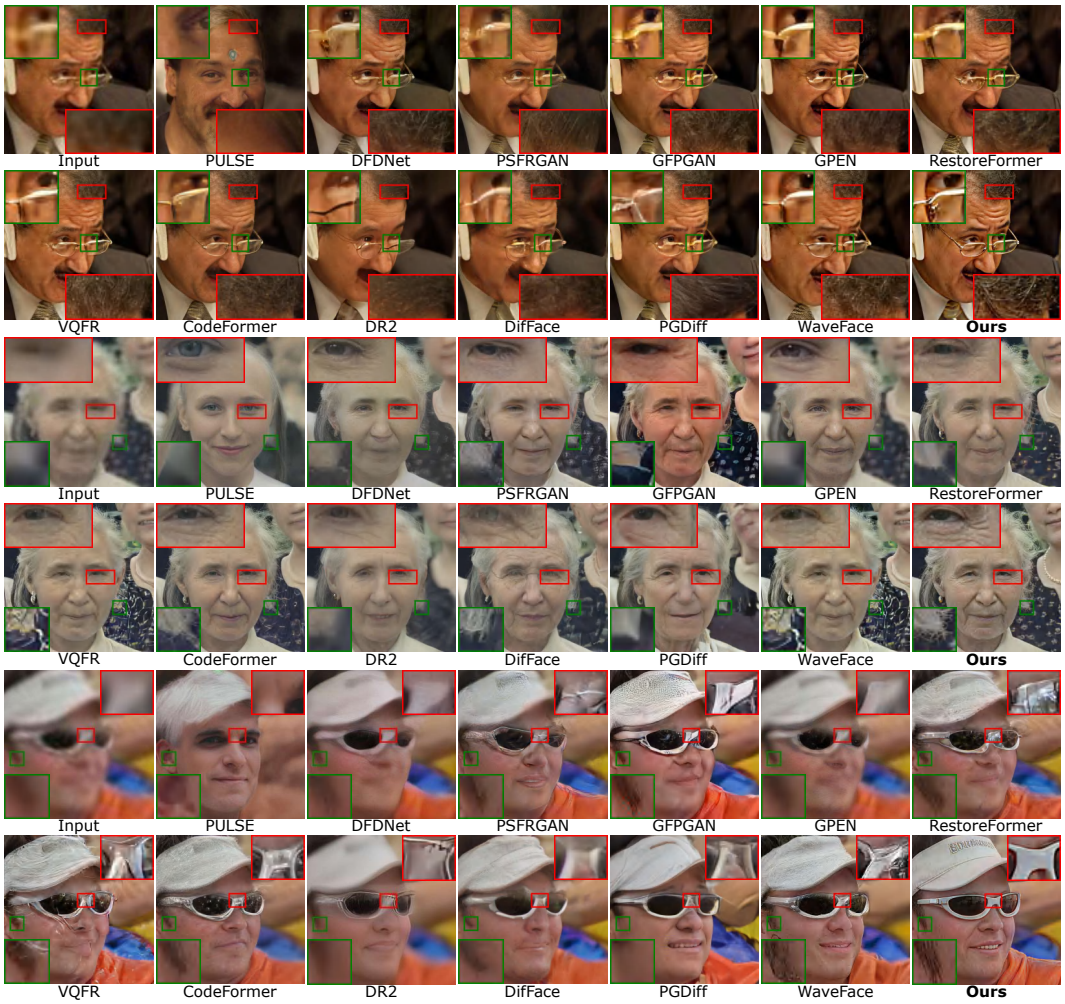

图 7: 基线方法在来自于 LFW-Test，WebPhoto-Test 和 WIDER-Test 的真实世界图像上的量化对比（更多结果参见 Appendix F）。**放大以便观看**。

（例如 “一张人脸的图片”，如 Fig. 8所示）能够通过空间特征有效捕捉低质量图像的全局人脸结构，然而，仍然在细节内容（如眼睛和皱纹）上有所欠缺。

我们也通过实验验证了起始步骤的选择，结果展示在 Fig. 8（第二行）和 Tab. 2（第四至第七行）。可以观察到如果从初始步骤（即纯噪声）开始会生成细节纹理（如皱纹），但会引入随机性（如眼睛），如④ VE+SE+1st 所示。如果从较晚步骤开始会导致输出模糊并保留低质量图像的原始纹理（ Fig. 8第二行第三列），但无法生成精细细节，这是由于去噪迭代次数的限制（ Fig. 8第二行第二列），如⑤ VE+SE+3rd 和⑥ VE+SE+4th 所示。因此，我们最终选择从第二步开始在潜在一致性模型中同时结合视觉嵌入和空间特征，既便于捕捉人脸特异性信息，又能生成精细细节（如 Fig. 8中的⑦和 Tab. 2的最后一行所示）。

推理时间。 Tab. 1（最后一列）展示了不同方法的推理时间。所有方法均在 Quadro RTX 3090 GPU（24GB 显存）上进行评估，输入图像分辨率为 512×512。本方法的采样时间与基于卷积/Transformer 的方法，例如 DFDNet (Li et al., 2020)、GPEN (Yang

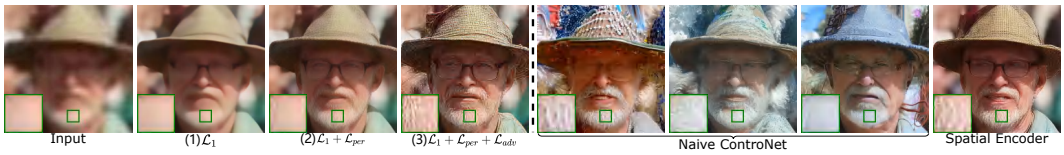

图 9: (左图) 表示感知损失与对抗损失的消融实验可视化结果。(右图) 表示标准 ControlNet 与本文空间编码器的消融实验可视化结果。

表 3: 感知损失与对抗损失的消融实验。

Exp.	$\mathcal{L}_1$	$\mathcal{L}_{per}$	$\mathcal{L}_{adv}$	LFW-Test FID↓	LFW-Test MUSIQ↑	WebPhoto-Test FID↓	WebPhoto-Test MUSIQ↑	WIDER-Test FID↓	WIDER-Test MUSIQ↑
(a)	✓			87.12	43.14	141.86	39.37	93.61	33.71
(b)	✓	✓		57.57	67.99	95.02	66.24	44.83	63.94
(c) Ours	✓	✓	✓	**51.32**	**76.16**	**75.48**	**75.88**	**35.43**	**76.29**

表 4: 标准 ControlNet 与本文空间编码器的消融实验。

Exp.	Loss	LFW-Test FID↓	LFW-Test MUSIQ↑	WebPhoto-Test FID↓	WebPhoto-Test MUSIQ↑	WiDER-Test FID↓	WiDER-Test MUSIQ↑
标准 ControlNet	Eq. (1)	**35.43**	75.03	81.91	73.63	49.58	74.20
空间编码器	Eq. (2)	55.07	**75.75**	**77.76**	**75.30**	**36.15**	**75.98**

et al., 2021) 和 VQFR (Gu et al., 2022) 方法的运行速度相近。同时，本方法的推理时间显著优于其他基于扩散模型的方法，如 PGDiff (Yang et al., 2024) 和 WaveFace (Miao et al., 2024)，它们仍受限于扩散模型固有的迭代采样过程。

感知损失与对抗损失的有效性。 我们认为 *InterLCM* 优越的修复性能主要源于在图像域同时结合了感知损失 (Johnson et al., 2016) 和对抗损失 (Goodfellow et al., 2014)，它们在常被用于重建模型训练中，且通常能产生高质量、高保真的人脸修复结果。为验证这两种损失的有效性，我们在 Fig. 9（左）和 Tab. 3中进行了消融实验。我们注意到当不使用感知损失和对抗损失时，量化指标显著下降（ Tab. 3（第一行）），因为仅使用重建损失难以获得良好的视觉质量（ Fig. 9（第二列））。向图像域加入感知损失和对抗损失后，能够有效恢复真实细节。此外，我们还对 *InterLCM* 中的空间编码器与标准 ControlNet 进行了对比研究（ Fig. 9（右）和 Tab. 4）。两者的主要区别在于训练时使用的损失函数。虽然原始 ControlNet 能生成高质量图像并保持结构，但由于其去噪损失聚焦于语义信息而忽略保真度 (Zhang et al., 2023)，导致保真度下降。

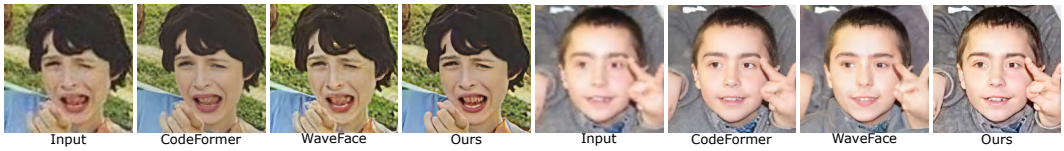

图 10: 输入手部可能复原失败的低质量图像。

5 结论

在本文中，我们提出了一种新颖的盲人脸修复（BFR）框架 *InterLCM* ，该框架利用潜在一致性模型（LCM）来提升语义一致性并从低质量输入中恢复高质量图像。通过将低质量图像视为潜在一致性模型的中间状态，*InterLCM* 与传统基于扩散的方法相比，能以更少的采样步骤实现更精确的修复。此外，我们整合了基于 CLIP 的图像编码器和视觉编码器以捕捉人脸特异性语义信息，以及基于 ControlNet 的空间编码器来保证结构一致性。在合成数据集和真实世界场景数据集上的大量实验表明，*InterLCM* 在图像质量和推理速度方面均优于现有方法，特别是在具有不可预测退化的复杂真实世界场景中。

局限。尽管我们的方法在盲人脸修复任务中优于现有方法，但仍无法避免一些局限：当 *InterLCM* 处理包含手部的图像时，虽然能生成更精细的面部细节，但无法产生真实的手部结构（见 Fig. 10）。这一局限可能源于 FFHQ 训练数据集中此类样本的稀缺性。潜在的改进方案是通过增加包含手部的多样化人脸图像来扩展训练数据。

致谢

本研究工作得到了国家自然科学基金（项目编号：62225604）和青年科学基金（项目编号：62202243）的支持。我们同时感谢西班牙政府通过 MCIN/AEI/10.13039/501100011033 和 FEDER 资助的 PID2022-143257NB-I00 项目支持。我们同样感谢“科创甬江 2035”关键技术攻关计划项目（2024Z120）的资助。本工作计算资源由南开大学超算中心（NKSC）提供。特别感谢南开大学李重仪教授在研究中提出的宝贵建议和深入讨论。

References

Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 185–200, 2018.

Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14245–14254, 2021.

Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11896–11905, 2021.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2492–2501, 2018.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12413–12422, 2022.

Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6059–6069, 2023.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE international conference on computer vision*, pp. 576–584, 2015.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Josh Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. *arXiv preprint arXiv:2306.05544*, 2023.

Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3012–3021, 2020.

Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NIPS Deep Learning Workshop*, 2014.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Xiaobin Hu, Wenqi Ren, John LaMaster, Xiaochun Cao, Xiaoming Li, Zechao Li, Bjoern Menze, and Wei Liu. Face super-resolution guided by 3d facial priors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 763–780. Springer, 2020.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *ICLR: international conference on learning representations*, pp. 1–15. ICLR US., 2015.

Charles Laroche, Andrés Almansa, and Eva Coupete. Fast diffusion em: a diffusion model for blind inverse problems with application to deconvolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5271–5281, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024a.

Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, 2020.

Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a $\mathcal{W}_+$ adapter for personalized image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Zhenyi Liao, Qingsong Xie, Chen Chen, Hannan Lu, and Zhijie Deng. Fine-tuning diffusion models for enhancing face quality in text-to-image generation. *arXiv preprint arXiv:2406.17100*, 2024.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I2sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

Xiaobin Lu, Xiaobin Hu, Jun Luo, Ben Zhu, Yaping Ruan, and Wenqi Ren. 3d priors-guided diffusion for blind face restoration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1829–1838, 2024.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023b.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. *arXiv preprint arXiv:2403.12760*, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024.

Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. *arXiv preprint arXiv:2312.05239*, 2023.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *OpenReview*, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8260–8269, 2018.

Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-if. https://github.com/deep-floyd/IF, 2023.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. In *International Journal of Computer Vision*, 2024.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021b.

Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1704–1713, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17512–17521, 2022.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3121–3138, 2022.

Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances in Neural Information Processing Systems*, 36, 2024.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, 2016.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 672–681, 2021.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.

Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 217–233, 2018.

Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models. *arXiv preprint arXiv:2312.06573*, 2023.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *arXiv preprint*, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yang Zhao, Tingbo Hou, Yu-Chuan Su, Xuhui Jia, Yandong Li, and Matthias Grundmann. Towards authentic face restoration with iterative diffusion models and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7312–7322, October 2023.

Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022.

Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7662–7671, 2022.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*, 2024.