arXiv:2603.01328v1 [cs.CV] 1 Mar 2026

你只需要一个阶段：从单张盲人脸图像合成新视角

Taoyue Wang[1], **Xiang Zhang**[1],**Xiaotian Li**[1],**Huiyuan Yang**[2],**Lijun Yin**[1]

[1]纽约州立大学宾汉姆顿分校计算学院 [2]密苏里科技大学计算机科学系 twang61@binghamton.edu, zxiang4@binghamton.edu, xli210@binghamton.edu, hyang@mst.edu, lijun@cs.binghamton.edu

摘要

我们提出一种新颖的单阶段方法 **NVB-Face**，可直接从单张**盲人脸**图像生成一致的**新视**图像。现有针对物体或人脸的新视角合成方法通常需要以高分辨率 RGB 图像作为输入。处理退化图像时，传统流程遵循两阶段流程：首先将图像恢复至高分辨率，然后基于恢复结果合成新视角。然而，这种方法高度依赖恢复图像的质量，常导致最终输出不准确且不一致。为解决这一局限，我们直接从盲人脸图像中提取单视角特征，并引入一种特征操纵器，将这些特征转换为三维感知的多视角潜在表示。借助扩散模型强大的生成能力，我们的框架能够合成高质量且一致的新视角人脸图像。实验结果表明，无论在一致性还是保真度方面，我们的方法都显著优于传统的两阶段方法。

1 引言

从单张人脸图像重建三维结构或生成新视图图像，长期以来一直是一项基础研究课题，其应用涵盖数字人建模与三维动画。人类凭借对面部结构和多视图一致性的丰富先验知识，能够毫不费力地想象给定人脸的不同视点。然而，在计算机视觉中，在精确保持身份、表情、背景及其他属性的同时忠实合成新视角，仍然是一个极具挑战性且不适定的问题。

目前，新视图人脸生成的主流方法主要分为三大类：基于回归的模型，如三维可变形模型（3DMM）（Paysan 等，2009；Wu 等，2019；Deng 等，2019b；Dib 等，2021；Zhuang 等，2022），基于神经辐射场（NeRF）的方法（Chan 等，2021；Gu 等，2021；Zhang 等，2022；Chan 等，2022；Yin 等，2023；Yuan 等，2023；An 等，2023；Trevithick 等，2023；Bhattarai、Nießner 和 Sevastopolsky，2024），以及利用 Stable Diffusion 和 ControlNet 的生成方法（Rombach 等，2022；Zhang、Rao,

版权所有 ©2026，人工智能促进协会（www.aaai.org）。保留所有权利。

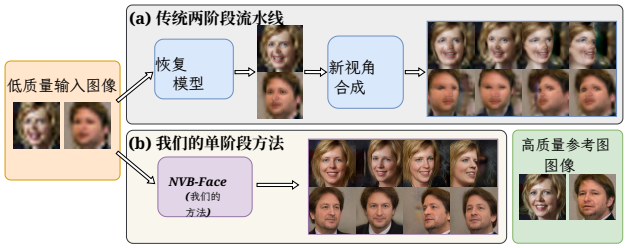


图 1：我们将我们的方法与典型的两阶段流程进行比较，例如 CodeFormer（Zhou 等，2022） + PanoHead-PTI（An 等，2023）。这些方法先复原退化图像，再合成新视角。显然，当复原阶段未能恢复准确细节时，这些误差会在新视角合成过程中进一步放大，导致结果与原始身份和外观显著偏离。相比之下，我们的方法直接从低质量输入图像出发，在单个阶段中生成新视角。此端到端设计抑制了误差累积，因此生成的新视角图像更加可靠、忠实。

以及 Agrawala，2023；Gu 等，2024、2025）。尽管这些方法各不相同，但它们有一个关键局限——**都要求输入为高质量单视角人脸图像**。遗憾的是，获取高质量人脸图像并非总是可行，因为大多数图像不可避免地会受到不同程度的退化。处理低分辨率、模糊、含噪或压缩退化人脸图像时，传统方法采用两阶段流水线：先修复输入图像，再进行新视角合成。然而，这种两阶段推理过程带来了一个根本问题：新视角生成的性能完全取决于修复后图像的质量。由于两个阶段彼此独立，缺乏显式连接，上游图像修复步骤中的任何错误都会在下游新视角合成过程中被放大，显著降低最终输出质量（如图 1 所示）。此外，这种流水线本身效率低下。只有在获得并评估修复结果之后，才能进行新视角合成，而为确保质量，往往还需要额外筛选或过滤。这种依赖关系

造成了大规模部署的主要瓶颈，使其在现实应用中实用性较低。

为解决这一问题，我们提出一种基于 Stable Diffusion（Rombach 等，2022）的端到端框架，能够在单次推理阶段中，从单张盲人脸图像直接生成高质量的新视角图像。我们的整体流程首先从低质量输入人脸图像中提取特征。随后，这些特征被转换为与目标相机视角相对应的新视角特征表示。最后，我们利用 Stable Diffusion 强大的生成能力，从变换后的低质量特征重建高分辨率新视角图像。具体而言，为实现潜在空间中的视点变换，我们引入基于 Transformer 的三维特征构建模块，该模块从单视角输入构建潜在三维人脸表示网格。借助相机参数，我们能够在潜在空间中将该三维表示直接投影为新视角特征。需要强调的是，与以往依赖 ControlNet（Zhang、Rao 和 Agrawala，2023）的基于扩散的新视角合成方法（Papantoniou 等人，2024；Gu 等，2024、2025）不同——这些方法先利用相机参数生成图像模板作为条件输入——我们的方法直接使用相机参数，对生成的三维特征网格执行视点条件映射。这不仅提供了更直观、更灵活的建模策略，而且由于采用了显式的三维特征表示，显著提升了多视角一致性。此外，我们的流程采用单阶段端到端推理过程，避免了多阶段方法中常见的误差累积，从而获得更高的计算效率和更强的泛化能力。

我们的贡献总结如下。

1. 我们提出了一种新颖的免调优框架，可在指定视点下直接从单张盲人脸输入生成高质量的人脸新视角图像。据我们所知，我们是首个探索这种端到端方式的工作。2. 我们提出了一种新颖的人脸特征三维潜在空间表示，可实现一致且准确的多视点特征投影。这种结构化表示有助于在新视角合成过程中保持跨视图一致性。3. 大量定性和定量实验表明，我们的单阶段新视角合成方法达到了最先进的性能，验证了与传统两阶段流水线相比，我们的框架能够实现更优的感知质量。

2 相关工作

我们提出的方法融合了两项具有挑战性的任务：**盲人脸修复** 和**新视角合成**。本节将简要概述这两个领域的相关工作，并探讨为何将这两项任务直接整合到单个推理阶段，以从退化图像生成高质量新视角，本质上具有很大难度。

盲图像修复。 将退化图像修复为高分辨率、细节丰富的版本，本质上是一个病态问题。在过去十年中，研究人员利用深度模型强大的生成能力，尤其是生成对抗网络（Yang 等，2021；Dib 等，2021；Wang 等，2021a、2022；Zhou 等，2022；Wang 等，2023a）和扩散模型（Wang 等，2023b；Yue、Wang 和 Loy，2023；Wang 等，2024b；Lin 等，2024；Yue 和 Loy，2024；Sun 等，2024；Yue、Wang 和 Loy，2023；Wu 等，2025；Wang 等，2024a），来重建退化图像中缺失的细节。基于 GAN 的方法通常采用 U-Net 风格的生成器，将退化图像映射为高质量输出，同时使用判别器来确保真实感。然而，生成对抗网络存在两大主要局限。首先，它们出了名地难以训练，因为维持生成器与判别器之间的平衡一直是长期存在的挑战。其次，其性能高度依赖大规模训练数据，往往需要数百万样本才能良好泛化。为解决这些问题，近期研究探索了用于图像修复的基于扩散的方法。早期用于盲图像修复的基于扩散的方法（Wang 等，2023b；Lin 等，2024；Yue 和 Loy，2024；Sun 等，2024）通常以退化图像为条件，使用 ControlNet（Zhang、Rao 和 Agrawala，2023）或类似的图像编码器（Wang 等，2024b）等模块提取特征，并将其注入冻结的预训练 Stable Diffusion 模型。随后，Stable Diffusion 强大的生成能力重建出高质量图像。鉴于扩散模型的显著优势，我们也采用基于扩散的骨干网络。然而，我们并未沿用现有基于扩散的图像修复范式——这些范式要求先修复单视角高分辨率图像，再单独执行多视角合成步骤——而是提出一种专门设计的新流程，以与新视角人脸合成任务无缝集成。

新视角人脸合成。 传统的基于回归的模型，例如三维可变形模型（3DMM）（Paysan 等，2009年；Wu 等，2019年；Deng 等，2019b；Dib 等，2021年；Zhuang 等，2022年），通过估计人脸的参数化几何先验，并渲染三维人脸重建结果来合成新视角。然而，由于参数空间的表达能力有限，基于 3DMM 的方法难以捕捉丰富的身份变化、表情及其他面部属性。因此，研究人员提出了基于 NeRF 的方法（Chan 等，2021年；Gu 等，2021年；Zhang 等，2022年；Trevithick 等，2023年），以及效率更高的演进方案——三平面表示（Chan 等，2022年；Bhattarai、Nießner 和 Sevastopolsky，2024年；An 等，2023年；Yin 等，2023年；Yuan 等，2023年）。其中，EG3D（Chan 等，2022年）是首个利用三维感知生成对抗网络生成潜在三平面表示，并渲染高质量、多视角一致人脸图像的方法。在此基础上，一系列研究（Yin 等，2023年；Yuan 等，2023年；Trevithick 等，2023年；Bhattarai、Nießner 和 Sevastopolsky，2024年）探索了 GAN反演技术，以适配 EG3D 的框架。然而，所有这些方法都依赖于先从输入图像中提取相机视点信息，以确保与 EG3D 的三平面表示保持一致。如果输入图像存在退化，准确提取相机参数就会极具挑战性，这使得

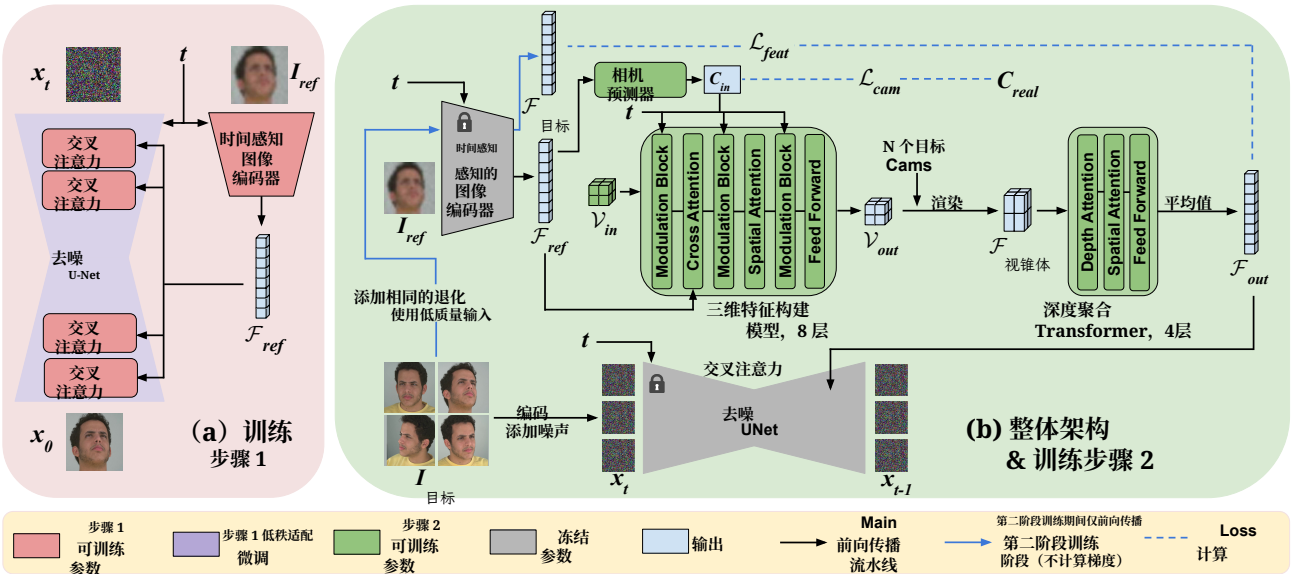


图 2：所提出的 **NVB-Face** 架构概览。**(a)** 我们的第一个训练步骤仅专注于图像修复。**(b)** 在第二个训练步骤中，我们仅更新新引入模块的参数（以深绿色高亮显示），并保持网络其余部分冻结。训练完成后，这一两步流程构成了我们完整的推理流程。

必须先修复图像，才能进行新视角合成——这一过程不可避免地会引入误差累积，并使结果依赖于修复质量。最近，Gu 等人（2024年）提出了首个用于新视角人脸生成的基于扩散的方法。然而，我们的实验揭示了该方法的一项重大局限：虽然它对高质量输入图像效果良好，但面对退化输入时表现明显不佳，导致严重的信息丢失。

为克服这些挑战，我们提出了一种单阶段框架，将盲人脸修复与新视角合成无缝整合为统一流程。与以往方法不同，我们的方法能够直接接收任意质量的人脸图像，并在指定的相机视点生成高分辨率且一致的新视角输出。

3 方法

给定一幅具有任意退化程度的输入图像 $\mathcal{I}_{ref}$，我们的目标是重建 $\mathcal{I}_{ref}$的原始细节，同时生成一幅高分辨率图像，使其保持相同的身份、表情和属性，但来自指定的新相机视角 $\mathcal{C}^i$。为此，我们提出NVB-Face，这是一个分两步训练的框架。首先，图像编码器从 $\mathcal{I}_{ref}$中提取潜在特征 $\mathcal{F}_{ref}$，随后将其输入Stable Diffusion（SD）模型（第3.1节）。借助其强大的生成能力，SD模型在保持身份一致性的同时恢复精细细节，并重建退化图像的高质量版本。接下来，我们引入三维感知视点变换模型，该模型接收单视角特征 $\mathcal{F}_{ref}$，并根据第 i个相机视角 $\mathcal{C}^i$将其变换为新特征 $\mathcal{F}^i_{out}$。随后，变换后的特征被送入我们预训练的SD模型，以合成高分辨率新视图图像（第3.2节）。最后，我们设计了一种定制损失函数，从而有效地将盲人脸修复与新视角合成衔接为统一的训练目标。我们进一步从理论和实证两方面论证了该损失形式如何确保在单阶段推理流程中对两项任务进行无缝优化（第3.3节）。

3.1 图像修复

如图 2a 所示，给定一张低质量（LQ）图像，我们的 NVB-Face 的修复组件首先使用图像编码器 Enc提取参考图像 $\mathcal{I}_{ref}$的特征 $\mathcal{F}_{ref}$，然后通过交叉注意力将这些特征输入 SD模型，生成一张高质量图像。需要注意的是，我们没有对最后一层的输出特征应用平均池化，而是在将其输入交叉注意力模块之前，保留了原始特征 $\mathcal{F}_{ref} \in \mathbb{R}^{H\times W\times C}$ 的维度。由于我们的任务是图像修复，保留特征的完整潜在空间分辨率能够确保保留输入图像中更多的精细细节，从而忠实地重建高分辨率输出，并准确保留原始内容。

遵循 Liu 等人 2023年和 Wang 等，2024b 的方法，我们将时间信息融入图像编码器，以提升生成质量和训练稳定性。我们将 SD模型中使用的相同时间步嵌入融入图像编码器的每个模块，使图像编码器具备时间感知能力，确保提取的特征与下游 SD模型的每个时间步同步。

$$\mathcal{F}_{ref} = Enc(\mathcal{I}_{ref},\, t) \tag{1}$$

由于我们用自定义图像编码器替换了 SD模型中的原始 CLIP（Radford 等人，2021）文本编码器，现有的交叉注意力模块（原本为

文本条件而设计）不再有效。为此，我们联合微调图像编码器和交叉注意力参数。此外，为了让模型更适合我们的特定任务，同时保留 SD模型固有的生成能力，我们使用低秩适配（Hu 等人，2022；Luo 等人，2023）微调 SD模型的所有组件，但交叉注意力模块除外。

训练目标与原始SD模型保持一致，即模型在每个时间步被训练以预测噪声。

$$\mathcal{L}_{SD} = \mathbb{E}_{x_0, \mathcal{F}_{ref}, t, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(x_t, t, \mathcal{F}_{ref}) \|^2 \right], \quad (2)$$

其中 x_0 是由VAE编码器编码的潜在图像，并经过 t 个时间步扩散为服从高斯分布的 $x_t \sim \mathcal{N}(0,1)$。而 ϵ_θ 是一个采用U-Net结构的噪声预测器。

除了使用常规的高质量人脸数据集外，我们还在第一个训练步骤中利用多视角人脸数据集。这确保了无论第二步生成的特征属于何种主体ID或视点，SD模型都能持续重建具有准确身份和细节保留的高分辨率图像。我们实验中所使用数据集的细节将在第4.1节中讨论。

3.2 新视图合成

3D特征构建模型。为了生成 $\mathcal{I}_{ref}$的新视角图像，先前的方法（Gu 等人，2024；Papantoniou等人，2024）试图从目标视点合成模板图像，并将其输入ControlNet（Zhang、Rao和Agrawala，2023年），以利用相机参数对模型进行条件控制。然而，存在两个主要问题。首先，在我们的情况下，SD模型已在第一个训练步骤中专门针对图像复原任务进行了微调，这可能会部分削弱其在其他任务（如新视角生成）上的生成能力。其次，ControlNet本身无法保证多视角一致性，因此需要在SD模型中集成一个可训练的视角一致性模块（Gu 等人，2024），从而增加了整体模型复杂度。

受（Peebles与Xie，2023年）以及（Hong 等人，2023）的启发，我们提出了一种基于Transformer的**3D特征构建模型**，记为 $Trans$，它将从图像编码器提取的单视角特征 $\mathcal{F}_{ref}$ 转换为融合多视角信息的3D特征体 $\mathcal{V}_{out}$ ，从而在3D空间中强制实现多视角一致性。

$$\begin{aligned} \mathcal{V}_{out} &= Trans(\mathcal{V}_{in}, \mathcal{F}_{ref}, \mathcal{C}_{in}, t), \\ \mathcal{C}_{in} &= CameraPredictor(\mathcal{F}_{ref}) \end{aligned} \quad (3)$$

其中 $\mathcal{C}_{in}$ 表示输入图像的相机参数，而 t 表示时间步嵌入，该嵌入与图像编码器和SD模型同步。该设计受到（Hong 等人，2023）的启发，其主张同时利用图像特征及相应的相机参数来构建3D表示。主要目标是将面部上下文（例如身份、表情）与姿态解耦，从而获得更稳定且一致的三维特征表示。然而，在推理过程中，往往难以获得准确的视点信息（来自退化的输入图像）。为解决这一问题，我们引入了一个**相机预测器** 模块，该模块直接根据输入图像特征估计相机参数。随后，预测得到的视点用于监督三维特征表示的生成，从而确保即使在推理时没有显式的相机输入，该表示仍保持视点一致性。

为了生成3D特征体 $\mathcal{V}_{out}$，我们定义空间位置嵌入 $\mathcal{V}_{in}$ 作为查询，随后将其输入交叉注意力，以整合来自 $\mathcal{F}_{ref}$的信息，如图2b所示。该机制注入了当前视点之外的信息，包括身份、表情、属性以及其他上下文细节。为了使3D表示以输入的相机视角为条件，我们引入了基于自适应层归一化的时间感知相机调制块，记为 $ModLN$（图2b中的调制块）

$$\begin{aligned} \gamma,\ \beta &= MLP(\mathcal{C}_{in}), \\ ModLN(f') &= LayerNorm(f) \cdot (1+\gamma) + \beta + t, \end{aligned} \quad (4)$$

它应用于 Transformer 中每个注意力模块与前馈层之间。

二维特征采样与聚合。遵循 Liu 等人（2023年）和 Yang 等人（2024年）的方法，随后通过沿每个观察方向进行射线采样和特征插值，将三维网格特征变换为与指定相机视角 $\mathcal{C}^i$ 对应的视锥体积特征

$$\mathcal{F}^i_{frustum} = Warp(\mathcal{V}_{out},\ \mathcal{C}^i),\ \mathcal{F}^i_{frustum} \in \mathbb{R}^{D \times H \times W \times C}. \quad (5)$$

由于我们的方法完全在潜在空间中运行，因此我们不采用 NeRF （ Mildenhall 等人，2021 年；Zhang 等，2022 ）中使用的体渲染，后者基于显式物理建模融合深度信息。相反，我们设计了一个**深度聚合 Transformer** $Aggr$，其模块包含跨深度注意力和空间注意力（图 2b 最右侧），以增强每个新视图生成的二维特征的表达能力。最后，我们沿视锥体积的深度维度进行平均池化，生成与输入特征 $\mathcal{F}_{ref}$ 空间维度相匹配的新视角特征 $\mathcal{F}^i_{out}$。

$$\mathcal{F}^i_{out} = AvgPool(Aggr(\mathcal{F}^i_{frustum})),\ \mathcal{F}^i_{out} \in \mathbb{R}^{H \times W \times C} \quad (6)$$

需要注意的是，在第二个训练步骤中，仅更新 3D特征构建模型、深度聚合 Transformer 和相机预测器的参数。图像编码器和 SD模型的参数保持冻结，因为它们已在第一个训练步骤中针对图像复原任务进行了微调。修改这些参数会对图像修复质量造成负面影响。总而言之，我们的方法在训练期间将目标任务解耦为两个独立目标，而在推理期间则将其融合为统一的流程。这种模块化设计增强了模型的可扩展性和鲁棒性，同时也降低了训练复杂度。

3.3 损失函数

在第二个训练步骤中，我们使用与公式 2 相同的扩散损失，同时额外引入特征损失 $\mathcal{L}_{feat}$，以使生成的新视角特征与对应的真实值对齐。真实值特征的获取方式是：首先使用与 $\mathcal{I}_{ref}$ 相同的退化程度，对新视角真实图像 $\mathcal{I}_{tgt}$ 进行退化。随后，将这些退化图像输入第一步中使用的预训练图像编码器，以提取新视角真实特征 $\mathcal{F}_{tgt}$。由于 SD模型已经学会如何根据退化特征修复高分辨率图像，应用相同的噪声模式可确保生成特征与真实值特征具有相同的分布。这样一来，模型便能专注于新视角特征生成，而非图像修复，从而有效地将这两个任务解耦。特征损失的公式如下

$$\mathcal{F}^i_{tgt} = Enc(Degrade(\mathcal{I}^i_{tgt}),\ t),$$

$$\mathcal{L}_{feat} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\mathcal{F}^i_{out} - \mathcal{F}^i_{tgt}\right)^2 + \lambda\left(1 - \frac{\mathcal{F}^i_{out}\cdot\mathcal{F}^i_{tgt}}{\|\mathcal{F}^i_{out}\|\|\mathcal{F}^i_{tgt}\|}\right)\right) \tag{7}$$

其中，N 是相机视图的数量，$\mathcal{L}_{feat}$ 由均方误差损失和平方余弦损失两部分组成，用于共同约束像素级和特征空间级的生成特征。此外，为确保相机预测器 $\mathcal{C}_{in}$ 预测的相机参数与输入图像的真实值一致，我们计算预测相机参数与真实相机参数之间的均方误差损失，并将其称为相机损失，

$$\mathcal{L}_{cam} = \|C_{\text{in}} - C_{\text{real}}\|_2^2 \tag{8}$$

因此，第二个训练步骤中的总损失函数定义为

$$\mathcal{L}_{total} = \mathcal{L}_{SD} + \lambda_1\mathcal{L}_{feat} + \lambda_2\mathcal{L}_{cam}. \tag{9}$$

4 实验

4.1 实验设置

数据集。 与 (Gu 等人，2024) 中的数据集设置类似，我们在第二个训练步骤中使用混合训练数据集进行新视角合成，其中包括照片级真实感多视角数据集 NeRSemble（Kirschstein等人，2023）以及使用 PanoHead（An 等，2023）生成的合成数据。NeRSemble 数据集包含从 16 个固定视点同步拍摄的 220 名受试者的动态表情视频。我们从该数据集中采样了 3000 组多视角帧。为扩大视点范围、增加主体多样性并增强背景变化，我们进一步使用 PanoHead 生成的 3000 组图像扩充训练数据。这种增强弥补了 NeRSemble 多样性较低的问题，因为它是一个可变性有限的实验室控制数据集。对于第一个训练步骤，我们增强了

通过纳入上述数据集，提升模型对多视角人脸的适应能力。此外，我们还加入高分辨率人脸数据集 FFHQ（Karras、Laine 和 Aila，2019），以增强模型的泛化能力。训练期间，所有输入图像均按照 EG3D 所采用的预处理策略进行裁剪和对齐。为生成



图3：在 NeRSem-ble（Kirschstein等人，2023年）数据集上的定性比较。如结果所示，与两阶段方法相比，这种端到端策略实现了更优的感知质量，并能更有效地保留身份和表情信息，同时最大限度地减少关键面部属性的损失。

低质量输入图像 $\mathcal{I}_{ref}$，我们采用 Real-ESRGAN（Wang 等人，2021b）流程进行退化模拟。为进行定量评估，我们从 NeRSemble 数据集中抽取 300 组未见过的多视图图像。为评估模型在真实场景数据上的泛化能力，我们在 LFW-Test（Huang 等人，2008）数据集上进行定性评估，并在 CelebA-Test（Karras 等人，2017年）数据集上对单视图重建进行定量评估。

训练细节。在第一个训练步骤中，图像编码器、SD模型的交叉注意力层以及 LoRA 参数（秩为16）共同优化200K次迭代，学习率为5e-5，批量大小为8。在第二个训练步骤中，我们将公式3中的超参数设为 $\lambda = 2.0$、 $\lambda_1 = 10$，并将公式9中的超参数设为 $\lambda_2 = 0.05$ 。对于每张输入图像，我们从其余视图（总视图数为 $N = 8$）中随机采样7个不同的目标视点，以监督潜在三维表示的生成。此步骤以 $1e-4$的学习率、2的批量大小训练300K次迭代。推理时，我们采用 DDIM（Song、Meng和Ermon，2020年）去噪采样器，步数为50。所有训练和推理实验均在由8张 NVIDIA RTX 6000 Ada GPU 组成的集群上进行。

4.2 定性比较

我们与多种最先进的新视角合成方法进行了对比实验，包括 PanoHead-PTI（An 等，2023）、GOAE（Yuan 等，2023）、Tri-PlaneNet（Bhattarai、Nießner 和 Sevastopolsky，2024）以及 DiffPortrait3D（Gu 等人，2024），并分别在 NeRSemble（Kirschstein等人，2023）和 LFW-Test（Huang 等人，2008）数据集上展示了定性比较结果，如图3和图4所示。在对比的方法中，DiffPortrait3D 基于 SD模型，其余方法则基于 GAN。由于这些现有方法并非为处理退化人脸图像而设计，因此我们首先使用 CodeFormer（Zhou 等，2022）进行盲人脸修复，再将修复结果输入各方法进行新视角生成。值得注意的是，基于 GAN 的方法需要准确的相机参数作为输入，才能进行新视角合成。然而，由于 CodeFormer 修复后的图像质量有所退化，



图4：在 LFW-Test（Huang 等人，2008）数据集上的定性比较。我们的方法在不同程度的输入退化下均能稳定地产生结果。与其他方法相比，即使在严重退化的情况下，我们生成的图像也能保留原始输入中的更多信息，并呈现出更高的视觉真实感。

通常难以从中提取可靠的相机参数。为了确保这些方法能够成功生成结果，我们只能改用原始高质量图像来提取输入相机参数。事实上，在真实世界的盲恢复场景中，这种做法并不可行，因此凸显了两阶段方法的一个关键局限：修复质量不佳，使得为新视角合成准确估计相机参数难以实现。我们的单阶段方法则完全避免了这一问题，因为它不依赖外部参数估计。

如图3所示，无论在身份还是表情保持方面，我们的方法都取得了最接近真实值的结果。相比之下，基于恢复的流程在盲恢复过程中不可避免地会引入误差，例如身份偏移、表情偏移和视觉伪影。当这些不完美的恢复结果被用作新视角生成的输入时，误差往往会被放大，导致信息大量丢失，图像质量明显退化。我们还在 LFW-Test 数据集上评估了我们的方法，以展示其在真实条件下面对不同退化程度时的鲁棒性。我们调整 Real-ESRGAN（Wang 等人，2021年b）的超参数，生成从弱到强的等级 1 和等级 2 退化。如图4所示，当输入为清晰图像时，我们的方法能够取得与先前方法相当的结果。然而，随着退化程度增加，其他方法的性能显著下降，生成的新视角质量也随之变差。虽然我们的方法也可能出现一些身份和表情偏移，但这些变化通常是基于退化输入所作出的合理预测，而且生成结果总体上更加稳定与其他方法相比。

4.3 定量比较

与 Gu 等人，2024 的做法类似，我们在多视角生成和单视角重建任务上评估我们的方法。为进行定量评估，我们采用一组全面的指标，包括学习感知图像块相似度（Zhang 等人，2018）、DISTS（Ding 等人，2020）、结构相似性（Wang，2004年）、身份相似度、弗雷歇起始距离（Heusel 等人，2017）以及姿态误差。身份相似度的计算方法是：使用（Deng 等，2019a）中的方法从生成图像中提取人脸嵌入，然后测量其与真实值嵌入之间的余弦相似度。姿态误差的计算方法是：使用（Deng 等，2019b）中的方法估计面部姿态，再计算该姿态与根据真实值图像估计出的姿态之间的均方误差（MSE）。对于多视角生成任务，我们使用 NeRSemble 数据集的未见测试划分，模型根据单张退化输入图像生成所有其他新视角。对于单视角重建任务，我们使用 CelebA-Test 数据集中的退化图像，目标是从相同姿态重建原始视图。竞争方法采用两阶段流水线：先使用 CodeFormer（Zhou 等，2022）复原退化图像，再进行新视角合成。

如表 1a 所示，我们提出的方法在所有评估指标上均优于所有基线方法。这些结果表明，我们的模型不仅能够忠实保留原始内容，还能从严重退化的输入中稳健地生成新视角。

	PanoHead-PTI	GOAE	TriplaneNet	DiffPortrait3D	Ours
结构相似性↑	0.73/0.62	0.73/0.58	0.75/0.63	0.68/0.62	**0.78/0.63**
学习感知图像块相似度↓	0.48/0.50	0.45/0.53	0.51/0.51	0.52/0.49	**0.17/0.49**
DISTS↓	0.26/0.28	0.25/0.27	0.32/0.29	0.26/0.27	**0.10/0.22**
FID↓	83.19/65.88	90.00/71.12	111.84/63.63	80.16/67.06	**5.67/22.71**
ID↑	0.32/0.22	0.27/0.21	0.30/0.34	0.29/0.21	**0.77/0.46**
姿态↓	0.0446/-	0.0425/-	0.0456/-	0.0341/-	**0.0084/-**

(a)

	w/o $\mathcal{L}_{feat}$	w/ $\mathcal{L}_{feat}$
结构相似性↑	0.75	**0.78**
学习感知图像块相似度↓	0.24	**0.17**
DISTS↓	0.16	**0.10**
FID↓	11.72	**5.67**
ID↑	0.57	**0.77**

(b)

表 1：(a) 在 NeRSemble 数据集（新视角合成）/CelebA-Test（单视图重建）上与其他方法的定量比较。(b) 在 NeRSemble 数据集上进行消融研究，以评估特征损失在我们框架中的有效性。具体而言，我们比较了模型在有无特征损失时的定量表现。



图 5：在 LFW-Test 数据集上进行定性消融研究，比较我们的方法在有无特征损失时的表现。



图 6：训练步骤 2 的特征校正。

4.4 消融研究

不使用特征损失。 我们在 NeRSemble 和 LFW-Test 数据集上进行了消融研究，以评估特征损失对定量和定性表现的影响。在消融设置中，我们移除特征损失，仅优化扩散损失，同时保持其他所有训练流程不变。如表 1b 所示，移除特征损失会导致模型在 NeRSemble 测试集上的表现显著下降。为进一步探究模型表现中受影响最大的方面，我们在包含具有挑战性的野外样本的 LFW-Test 数据集上评估模型。如图 5 所示，移除特征损失会导致生成的输出出现严重的多视角不一致性。我们认为，这是由于训练期间缺乏对潜在空间的显式约束。没有特征损失时，模型只需生成视觉上合理的最终图像，而忽略了潜在空间中新视角特征的对齐与一致性。尽管扩散模型强大的生成能力可以在训练期间表面上弥补特征未对齐的问题，但这种捷径无法泛化。在未见过的数据上测试时，未对齐的潜在特征会导致明显的多视角不一致性，严重损害生成结果的可靠性。

训练步骤 2 中的特征校正我们进一步研究训练步骤 1 中不理想的图像修复是否
会对最终的新视角合成结果产生负面影响。为此，我们使用训练步骤 1 的图像复原模型，从退化的输入中恢复高分辨率图像，并将其与我们的完整 NVB-Face 模型生成的最终新视角输出进行比较。如图 6 所示，即使训练步骤 1 恢复的图像在视觉上并不理想——这表明图像编码器未能提取所有关键信息——我们的训练步骤 2 仍然能够纠正不完善的特征，并生成更加逼真的新视角图像。这得益于我们的视点变换模型，该模型在训练步骤 2 中独立进行训练。因此，即使输入特征存在部分缺陷，该模型仍能生成稳健的新视角特征。如前文所述，当修复质量不佳时，传统的两阶段方法往往会在新视角合成过程中放大误差。相比之下，我们的单阶段方法能够减轻此类误差传播，从而获得更加稳定可靠的输出。

5 结论

本文提出了 NVB-Face，这是首个能够直接从单张盲人脸图像生成新视角的框架。NVB-Face 将整体任务拆解为两个经过精心设计、可独立优化的步骤，同时引入定制的损失函数，使这两个组件能够无缝衔接。与先进行人脸复原、再进行新视角合成的传统两阶段流程不同，我们的端到端方法显著减少了误差累积，并能更好地保留身份及其他重要的图像属性。此外，通过在三维潜空间中构建新视角特征，我们的方法有效地确保了生成图像之间的多视图一致性。大量实验表明，NVB-Face 能够为任意真实场景中的盲人脸图像可靠地合成高质量且一致的新视角。

6 致谢

本研究部分基于国家情报总监办公室（ODNI）通过情报高级研究计划局（IARPA）提供的支持，项目编号为[2022-21102100003]。本文所述观点和结论仅代表作者立场，不应被解读为必然代表 IARPA 或美国政府明示或暗示的官方政策。尽管其中含有任何版权声明，美国政府仍有权为政府用途复制和分发转载件。

参考文献

An, S.; Xu, H.; Shi, Y.; Song, G.; Ogras, U. Y.; and Luo, L. 2023年。Panohead：360度几何感知三维全头部合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，20950–20959。

Bhattarai, A. R.; Nießner, M.; and Sevastopolsky, A. 2024年。Triplanenet：用于EG3D反演的编码器。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，3055–3065。

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; 等。2022年。高效的几何感知三维生成对抗网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，16123–16133。

Chan, E. R.; Monteiro, M.; Kellnhofer, P.; Wu, J.; 和 Wet- zstein, G. 2021年。pi-gan：用于三维感知图像合成的周期性隐式生成对抗网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，5799–5809。

Deng, J.; Guo, J.; Xue, N.; 和 Zafeiriou, S. 2019a。Arcface：用于深度人脸识别的加性角度间隔损失。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，4690–4699。

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; 和 Tong, X. 2019b。基于弱监督学习的精准三维人脸重建：从单张图像到图像集。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，0–0。

Dib, A.; Thebault, C.; Ahn, J.; Gosselin, P.-H.; Theobalt, C.; 和 Chevallier, L. 2021年. 利用自监督学习和光线追踪实现具有丰富反射率的高保真单目人脸重建。载于*IEEE/CVF*国际计算机视觉会议论文集，12819–12829。

Ding, K.; Ma, K.; Wang, S.; 和 Simoncelli, E. P. 2020年. 图像质量评估：统一结构与纹理相似度。*IEEE*模式分析与机器智能汇刊，44(5)：2567–2581。

Esser, P.; Rombach, R.; 和 Ommer, B. 2021年. 驯服变换器以实现高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，12873–12883。

Gu, J.; Liu, L.; Wang, P.; 和 Theobalt, C. 2021年。StyleNeRF：一种用于高分辨率图像合成的基于风格的三维感知生成器。*arXiv* 预印本 *arXiv:2110.08985*。

Gu, Y.; Tran, P.; Zheng, Y.; Xu, H.; Li, H.; Karmanov, A.; 和 Li, H. 2025年。DiffPortrait360：用于360度视图合成的一致人像扩散。*arXiv* 预印本 *arXiv:2503.15667*。Gu, Y.; Xu, H.; Xie, Y.; Song, G.; Shi, Y.; Chang, D.; Yang, J.; 和 Luo, L. 2024年。DiffPortrait3D：用于零样本人像视图合成的可控扩散。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，10456–10465。Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; 和 Hochreiter, S. 2017年。采用双时间尺度更新规则训练的生成对抗网络会收敛到局部纳什均衡。神经信息处理系统进展，30。

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; 和 Tan, H. 2023年. 大规模重建模型：用于单图像转3D的大规模重建模型。*arXiv*预印本 *arXiv:2311.04400*。

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; 等. 2022年。LoRA：大语言模型的低秩适配。国际学习表征会议，1(2)：3。

Huang, G. B.; Mattar, M.; Berg, T.; 和 Learned-Miller, E. 2008。野外标记人脸：一个用于研究非受限环境下人脸识别的数据库。发表于真实场景图像研讨会：检测、对齐与识别。

Karras, T.; Aila, T.; Laine, S.; 和 Lehtinen, J. 2017年。通过渐进式生成生成对抗网络，提升图像质量、稳定性和多样性。*arXiv* 预印本 *arXiv:1710.10196*。

Karras, T.; Laine, S.; 和 Aila, T. 2019年。用于生成对抗网络的基于风格的生成器架构。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，4401–4410。

Kirschstein, T.; Qian, S.; Giebenhain, S.; Walter, T.; and Nießner, M. 2023年。NeRSemble：人头多视角辐射场重建。*ACM*图形学汇刊（*TOG*），42(4)：1–14。

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Qiao, Y.; Ouyang, W.; and Dong, C. 2024年。DiffBIR：利用生成式扩散先验实现盲图像复原。载于欧洲计算机视觉会议，430–448。Springer。

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023年。Syncdreamer：从单视角图像生成多视角一致图像。*arXiv*预印本 *arXiv:2309.03453*。

Luo, S.; Tan, Y.; Patil, S.; Gu, D.; von Platen, P.; Passos, A.; Huang, L.; Li, J.; 和 Zhao, H. 2023年。Lcm-lora：一种通用的Stable Diffusion加速模块。*arXiv*预印本 *arXiv:2311.05556*。Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; 和 Ng, R. 2021年。NeRF：将场景表示为用于视图合成的神经辐射场。*ACM*通讯，65(1)：99–106。Papantoniou, F. P.; Lattas, A.; Moschoglou, S.; Deng, J.; Kainz, B.; 和 Zafeiriou, S. 2024年。Arc2face：一种用于生成身份一致人脸的基础模型。载于欧洲计算机视觉会议，241–261。Springer。

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; 和 Vetter, T. 2009年。一种用于姿态和光照不变人脸识别的三维人脸模型。载于*2009*年第六届*IEEE*国际先进视频与基于信号的监控会议，296–301。IEEE。Peebles, W.; 和 Xie, S. 2023年。基于Transformer的可扩展扩散模型。载于*IEEE/CVF*国际计算机视觉会议论文集，4195–4205。Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; 等。2021年。从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议，8748–8763。PmLR。Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; 和 Ommer, B. 2022年。利用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，10684–10695。Song, J.; Meng, C.; 和 Ermon, S. 2020年。去噪扩散隐式模型。*arXiv*预印本 *arXiv:2010.02502*。Sun, L.; Wu, R.; Zhang, Z.; Yong, H.; 和 Zhang, L. 2024年。提升扩散模型在内容一致性超分辨率中的稳定性。*CoRR*。Trevithick, A.; Chan, M.; Stengel, M.; Chan, E.; Liu, C.; Yu, Z.; Khamis, S.; Chandraker, M.; Ramamoorthi, R.; 和 Nagano, K. 2023年。用于单图人像视图合成的实时辐射场。*ACM*图形学汇刊（*TOG*），42(4)：1–15。Wang, J.; Gong, J.; Zhang, L.; Chen, Z.; Liu, X.; Gu, H.; Liu, Y.; Zhang, Y.; 和 Yang, X. 2024a年。OSDFace：用于人脸复原的单步扩散模型。*arXiv*预印本 *arXiv:2411.17163*。Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; 和 Loy, C. C. 2024b年。利用扩散先验实现真实世界图像超分辨率。国际计算机视觉期刊，132(12)：5929–5949。Wang, X.; Li, Y.; Zhang, H.; 和 Shan, Y. 2021a年。借助生成式人脸先验实现真实世界盲人脸修复。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，9168–9178。Wang, X.; Xie, L.; Dong, C.; 和 Shan, Y. 2021b年。Real-ESRGAN：利用纯合成数据训练真实世界盲超分辨率模型。载于*IEEE/CVF*国际计算机视觉会议论文集，1905–1914。Wang, Z.; Bovik, A. C.; Sheikh, H. R.; 和 Simoncelli, E. P. 2004年。图像质量评估：从误差可见性到结构相似性。*IEEE*图像处理汇刊，13(4)：600–612。Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; 和 Luo, P. 2022年。Restoreformer：通过未退化键值对实现高质量盲人脸修复。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，17512–17521。Wang, Z.; Zhang, J.; Chen, T.; Wang, W.; 和 Luo, P. 2023a年。RestoreFormer++：利用未退化键值对实现真实世界盲人脸修复。*IEEE*模式分析与机器智能汇刊。

Wang, Z.; Zhang, Z.; Zhang, X.; Zheng, H.; Zhou, M.; Zhang, Y.; and Wang, Y. 2023b. Dr2：基于扩散的盲人脸修复鲁棒退化去除器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，1704–1713。Wu, F.; Bao, L.; Chen, Y.; Ling, Y.; Song, Y.; Li, S.; Ngan, K. N.; and Liu, W. 2019. Mvf-net：多视图三维人脸可变形模型回归。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，959–968。Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2025. 用于真实世界图像超分辨率的一步高效扩散网络。神经信息处理系统进展，37：92529–92553。Yang, J.; Cheng, Z.; Duan, Y.; Ji, P.; and Li, H. 2024. Consistnet：为多视图图像扩散强制实施三维一致性。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，7079–7088。Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. 用于野外场景盲人脸修复的嵌入GAN先验的网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，672–681。Yin, F.; Zhang, Y.; Wang, X.; Wang, T.; Li, X.; Gong, Y.; Fan, Y.; Cun, X.; Shan, Y.; Oztireli, C.; et al. 2023. 具有面部对称先验的三维GAN反演。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，342–351。Yuan, Z.; Zhu, Y.; Li, Y.; Liu, H.; and Yuan, C. 2023. 通过几何与遮挡感知编码，让编码器在三维GAN反演中再次强大。载于*IEEE/CVF*国际计算机视觉会议论文集，2437–2447。Yue, Z.; and Loy, C. C. 2024. Diffface：利用扩散误差收缩进行盲人脸修复。*IEEE*模式分析与机器智能汇刊。Yue, Z.; Wang, J.; and Loy, C. C. 2023. Resshift：通过残差平移实现用于图像超分辨率的高效扩散模型。神经信息处理系统进展，36：13294–13307。Zhang, J.; Li, X.; Wan, Z.; Wang, C.; and Liao, J. 2022. Fdnerf：用于人脸重建和表情编辑的少样本动态神经辐射场。载于*SIGGRAPH Asia 2022*会议论文集，1–9。Zhang, L.; Rao, A.; and Agrawala, M. 2023. 为文生图扩散模型添加条件控制。载于*IEEE/CVF*国际计算机视觉会议论文集，3836–3847。Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. 深层特征作为感知度量的非凡有效性。载于*IEEE*计算机视觉与模式识别会议论文集，586–595。Zhou, S.; Chan, K.; Li, C.; and Loy, C. C. 2022. 通过码本查找变换器实现稳健的盲人脸修复。神经信息处理系统进展，35：30599–30611。Zhuang, Y.; Zhu, H.; Sun, X.; and Cao, X. 2022. Mofan-erf：可变形人脸神经辐射场。载于欧洲计算机视觉会议，268–285。Springer。

A 我们的架构设计详述

我们的图像编码器遵循（Esser、Rombach 和 Ommer，2021年；Zhou 等，2022）的设计（见图7），并接收大小为 $512 \times 512 \times 3$ 的输入图像，生成大小为 $8 \times 8 \times 1024$ 的输出特征图。随后，这些提取的特征会通过交叉注意力输入SD模型（训练步骤 1），或输入视点变换模型（训练步骤 2）。为确保与SD模型的内部表示保持对齐，我们遵循（Rombach 等，2022）的策略，将扩散过程所使用的相同时间步插入每个残差块中。该设计确保了编码器输出与扩散模型去噪轨迹之间的时间一致性和空间一致性。

在训练步骤 2 中，为了高效实现三维感知视图变换模型，我们采用了一种节省内存的设计。具体而言，我们将空间位置嵌入 $\mathcal{V}_{in}$ 设为大小为 8×8×8×1024，输出的 3D特征体 $\mathcal{V}_{out}$ 保持与8×8×8×1024相同的分辨率。遵循（Liu 等人，2023；Yang 等，2024）提出的方法，我们将该3D特征体变换为与目标相机视点 $\mathcal{C}_i$对应的2D特征视锥 $\mathcal{F}^i_{frustum}$，深度维度为 12。最后，我们通过深度聚合将沿深度方向的特征融合为紧凑的2D表示，以用于新视角合成。

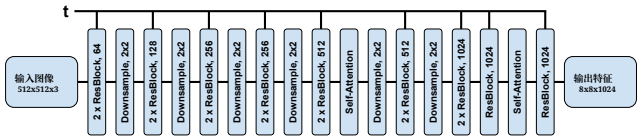


图7：我们的时间感知图像编码器架构

B 盲人脸修复性能

我们在训练步骤 1 中评估了盲人脸修复（BFR）流水线，并将其与 CelebA-Test 数据集上的 CodeFormer（Zhou 等，2022）（基于 GAN）和 DiffBIR（Lin 等，2024）（基于扩散）进行比较。结果见表2。如表中所示，我们的BFR模块性能与这些经典方法相当，但并未展现出显著优势。这是预期之中的，因为我们的主要目标并非优化BFR本身，而是将BFR与新视角合成整合到一个统一框架中。我们没有针对修复任务引入任何专门的设计或改进。尽管如此，这一比较进一步验证了我们单阶段流水线的优越性：即使我们的BFR性能与其他方法相当，最终的新视角合成结果仍显著优于基于这些方法的两阶段流水线所生成的结果。这证实了我们的方法能够有效缓解误差传播，同时保留传统的

两阶段方法往往会在后续合成阶段放大复原误差。



图 8：LFW-Test 数据集上基于 CodeFormer 的两阶段方法、基于 DiffBIR 的方法与我们的方法的定性比较

C 与基于 DiffBIR 的新视角合成方法的比较

我们进一步将本方法与基于 DiffBIR（Lin 等人，2024年）的两阶段流水线进行比较。如表 3 所示，即使在第一阶段复原中使用基于扩散的 DiffBIR 模型，而非 CodeFormer，我们的方法仍能取得更优的性能。此外，通过比较表中基于 CodeFormer 和基于 DiffBIR 的变体结果可以看出，DiffBIR 在大多数情况下的表现不如 CodeFormer。图 8 中的定性结果也进一步印证了这一点：在第一阶段使用 DiffBIR 会导致视觉质量较差，尤其是对于退化更严重的输入。我们推测，这是因为 DiffBIR 未在退化程度更高的图像上进行训练，因此在此类场景中的鲁棒性有限。因此，在全文所有两阶段基线比较中，我们始终在第一阶段使用 CodeFormer 作为复原模块，以确保结果公平且具有竞争力。

D 更多结果

我们在图 10 和图 11 中展示了更多结果，其中所有输入图像均取自 LFW（Huang 等人，2008）数据集。图 10 展示了退化程度相对较轻的情况，其中一些案例（例如最后两行）输入的是原始清晰图像。我们的方法始终能够生成高保真度的新视角图像，并保持较强的身份相似度和多视图一致性，表现与现有最先进的新视角合成方法不相上下。图 11 展示了模型在极度退化条件下的表现。即使在如此困难的情形下，我们的方法仍能稳定地生成合理且逼真的新视角图像。虽然可以观察到一些身份和表情偏移，但考虑到输入图像中信息的严重丢失，我们认为这些偏差可以接受。这些结果进一步证实，与两阶段方法相比，我们的单阶段流程在应对具有挑战性的真实世界退化时展现出显著更强的鲁棒性。

	峰值信噪比↑	结构相似性↑	学习感知图像块相似度↓DISTS↓	弗雷歇起始距离↓		多尺度图像质量↑	自然图像质量评价↓	C-IQA↑	M-IQA↑
CodeFormer	20.98	**0.6174**	0.4750	0.2636	61.44	66.06	5.6992	0.5608	0.4775
DiffBIR	**22.00**	0.6000	**0.4631**	0.2340	32.54	**75.61**	6.2855	**0.7895**	0.6256
Ours	20.05	0.5574	0.4969	**0.2133**	**18.43**	74.71	**5.2747**	0.7431	**0.6536**

表2：与其他经典盲人脸修复（BFR）方法的比较。粗体表示最佳结果，下划线表示次佳结果。

	结构相似性↑	学习感知图像块相似度↓	DISTS↓	FID↓	ID↑	姿态↓
PanoHead-PTI	0.63(0.73)	0.54(0.48)	0.27(0.26)	85.31(83.19)	0.30(0.32)	0.0484(0.0446)
GOAE	0.62(0.73)	0.50(0.45)	0.27(0.25)	108.65(90.00)	0.27(0.27)	0.0392(0.0425)
TriplaneNet	0.70(0.75)	0.52(0.51)	0.29(0.32)	136.21(111.84)	0.29(0.30)	0.0484(0.0456)
Ours	**0.78**	**0.17**	**0.10**	**5.67**	**0.77**	**0.0084**

表 3：与基于 DiffBIR 的两阶段方法在 NeRSemble 数据集上的定量比较。括号中的数值表示基于 CodeFormer 修复的结果。带下划线的数值表示基于 CodeFormer 和 DiffBIR 的变体中表现更好的结果。

E 局限性与未来工作

如附录 B 所述，尽管我们的图像修复模块的性能与其他盲人脸修复（BFR）方法相当，但其结果并不特别突出。这进一步印证了训练步骤 2 的有效性，但我们也承认，在训练步骤 1 中获得更好的修复质量，有助于生成质量更高的新视角图像。未来，我们计划进一步优化训练步骤 1 的流程，力求在提升修复性能的同时，确保其与新视角合成任务无缝集成。

此外，尽管我们的模型能够忠实地生成高质量且视角一致的人脸图像，但其生成非面部内容（如服装和背景）的能力有限。如图 9 所示，模型错误地将帽子视为背景的一部分，导致输出在视觉上不够真实。此外，由于背景场景复杂，我们的模型有时无法生成时间稳定的背景，导致不同视角之间出现明显闪烁。这些局限源于当前模型主要关注面部区域，而未对非面部元素进行显式建模或优化。未来工作中，我们计划更深入地探索这些挑战，拓展生成范围。

使我们的模型能够以一致且逼真的方式更好地处理配饰、服装和背景。



图 9：我们模型的局限性

F 社会影响

我们提出的 NVB-Face 模型旨在根据任意质量的输入生成高质量的新视角人脸图像。这项工作的初衷并非用于欺诈或冒充等恶意用途，而是解决若干具有挑战性且有益于社会的任务。例如，我们的模型可用于修复低质量人脸图像并生成多视角表示，从而提升多视角人脸识别系统的性能。在现实场景中，它可以协助将监控录像或低分辨率证据帧转换为高分辨率多视角人脸图像，提高嫌疑人识别和身份验证的效率。此外，它还有望通过改善人脸匹配，为搜救工作中寻找失踪人员提供帮助。不过，我们也意识到存在潜在的滥用风险，例如用于深度伪造生成——比如根据一张社交媒体图像创建高质量的多视角人脸化身。为降低此类风险，我们提议在生成的输出中嵌入可见水印，以明确标示其合成来源。我们致力于负责任的人工智能开发，并将在研究不断推进的过程中，持续向公众通报最新进展、安全保障措施和技术手段。



图 10：更多简单案例的新视角结果



图 11：更多困难案例的新视角结果