arXiv:2301.00950v3 [cs.CV] 9 Jan 2024

类别连续条件生成神经辐射场

金志旭tom919@cau.ac.kr李珉赫*mlee@cau.ac.kr

电气与电子工程学院 中央大学 首尔 06974，大韩民国

摘要

三维感知图像合成的重点在于生成具有精细细节的高分辨率图像，同时保持空间一致性。近年来，神经辐射场（NeRF）已成为一种强大的新视角合成方法，具有较低的计算成本和出色的性能。尽管现有的生成式 NeRF 方法已取得显著成果，但它们无法在生成过程中处理条件特征操控和连续特征操控。在本研究中，我们提出了一种名为类别连续条件生成 NeRF（C^3G-NeRF）的新模型，通过将条件特征投射到生成器和判别器上，合成经条件操控的照片级真实感三维一致图像。我们在三个图像数据集上评估了所提出的 C^3G-NeRF：AFHQ、CelebA 和 Cars。我们的模型展现出强大的 3D一致性、精细细节、 360° 生成能力，以及条件特征操控中的平滑插值。例如，C^3G-NeRF 在三维感知人脸图像合成中实现了 7.64 的弗雷歇 Inception 距离（FID），分辨率为 128^2 。此外，我们提供了数据集中各类别生成的 3D感知图像的 FID，并展示了 C^3G-NeRF 合成类别条件图像的能力。

1 引言

已有许多方法 [20, 43]可用于合成新颖图像。生成对抗网络（GAN） [11] 通过学习数据集的分布，在图像生成方面取得了出色成果，从而合成出逼真的实例。此外，许多研究提升了生成高分辨率图像的能力[5, 6, 18, 19]。尽管已有研究取得了进展，但由于大多数数据集基于二维信息，生成对抗网络在合成单个物体的多视角图像时仍存在局限，这可能导致 3D一致性不稳定。

为了解决这一问题，研究人员对基于三维的 GAN [30, 31, 44]进行了研究，这类方法采用体渲染方法。然而，这些方法需要大量计算资源和内存才能有效训练。近年来，Mildenhall 等人 [26]提出了神经辐射场（NeRF），作为传统基于体素的体渲染方法 [10, 14, 17]的替代方案。NeRF 大幅降低了计算复杂度

*通讯作者。© 2023。本文档版权归作者所有，可免费以纸质或电子形式原样分发。

与其他方法相比，所需的记忆量更少。因此，神经辐射场已扩展应用于基于3D的生成对抗网络研究 [4, 9, 32, 39]，并取得了出色的结果，同时复杂度较低。

然而，现有的基于神经辐射场的生成对抗网络无法通过条件标签连续地控制图像生成，因为它们没有将必要的条件信息纳入生成对抗网络的生成器中。为基于神经辐射场的生成对抗网络添加条件至关重要，因为许多行业应用都要求在生成过程中进行精确操控，例如在新兴的元宇宙中选择虚拟形象的细节。GIRAFFE [32]，是一种基于神经辐射场的生成模型，无需额外信息（如传统神经辐射场所需的相机位置和方向），即可生成具有 3D 感知能力的图像。尽管 GIRAFFE 取得了令人瞩目的成果，尤其是在 360° 真实世界数据集上的生成方面，与 pi-GAN 和高效几何感知 3D 生成对抗网络（EG3D）等其他最先进的模型相比， [3, 4], 它仍无法解耦各种特征， [2, 24, 31] 而这些特征是生成具有指定属性的图像所必需的。

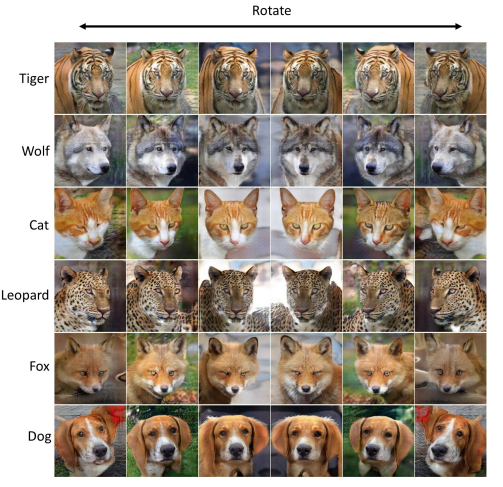


图 1：我们的模型生成的 AFHQ 各类别图像（分辨率为 256^2 ）。每行展示一个物体在不同旋转输入向量下的图像。请注意，不同类别的图像由单个模型通过不同的条件输入向量生成。我们的模型能够生成不同物体的多种视角，同时保持强大的 3D一致性。

Jo 等人. [16] 尝试通过提供图像类型和文本等条件信息来解决这一问题。然而，这种方法无法有效地表示条件强度，因为文本与数值相比含义较为模糊，而数值能够更直观地设定强度。控制条件强度至关重要，因为大多数元宇宙用户都希望用特定的眼型或发色定制自己的虚拟形象，而选择代表各项条件强度的数值即可实现这一点。

本文提出一种新方法，用于解决 3D 感知条件图像生成任务。该任务要求条件标签值 [27, 28, 33] 能够控制生成图像的特征，如图 1所示；这些数值还必须是连续的，才能使相应条件的强度平滑变化。据我们所知，本文首次研究这一任务。

我们提出了类别连续条件生成神经辐射场（C^3G-NeRF），专注于 3D 感知图像生成中的条件特征连续操控。我们采用 GIRAFFE 作为骨干模型，它无需额外的相机参数即可生成图像；由于应用范围广，这也是 3D 感知生成中的一项核心任务。尽管 EG3D 和 pi-GAN 在生成质量方面优于 GIRAFFE，但它们并不适用于真实世界数据集上的 360° 生成。EG3D 依赖额外的相机参数，因此不适合在缺少这些参数的真实世界数据集上进行学习。此外，我们使用弗雷歇 Inception 距离（FID）和核初始距离（KID）指标，将 GIRAFFE 在真实世界汽车数据集 [1] 上的生成表现与 pi-GAN 进行比较，以验证 GIRAFFE 的必要性。

我们观察到 th 条件GIRAFFE若不使用残差模块，难以学习数据分布

中的精细细节，因此残差模块对于训练条件GIRAFFE至关重要。为解决这一问题，我们将残差模块 [12, 13] 融入模型架构，以支持训练并提升图像合成效果。

针对在 3D感知图像中生成多视角实例这一挑战，C^3G-NeRF 能够在多个视角之间实现精细的 3D一致性。我们展示了来自三个数据集的结果：AFHQ [7], CelebA [23], 和 Cars [1], , 并展示了如何通过平移、旋转以及在单张图像中添加物体来控制图像合成。

本文的贡献如下：

- 我们提出 C^3G-NeRF 模型，以解决一项新的任务：在 3D 感知图像生成中进行有条件的连续特征操控。
- 通过将残差模块融入神经辐射场架构，我们缩短了训练时间并提升了性能。
- 我们使用多个数据集展示了 3D 感知图像生成中的有条件连续特征操控：AFHQ、CelebA 和 Cars。
- 由于我们的模型能够生成类别条件的 3D感知图像，因此我们提供 AFHQ 和 Cars 数据集中每个标签的 FID 分数。

2 相关工作

隐式神经表示与渲染：近年来，利用深度学习技术[21] 表示三维空间受到了广泛关注。在提出的各种方法中，隐式神经表示[36, 40, 42] 展现出尤为可观的前景。NeRF 将隐式神经表示与体渲染相结合， [10] 从而能够合成训练数据中未明确表示的新视图。因此，NeRF 能够合成具有精细细节且保持 3D 一致性的图像。然而，使用 NeRF 的一个缺点是训练期间需要高度受限的图像作为监督。另一个问题是，每个 NeRF 实例只能表示单个物体，而不能同时表示多个物体。研究者提出，生成式NeRF或许能够缓解这些问题。

生成式NeRF： 近年来，基于NeRF的方法在从二维无约束图像数据集合成3D感知图像方面取得了一些进展。这些方法通过训练生成模型来确保连续的三维几何一致性。例如，GRAF [39] 和 pi-GAN [4] 都提出了生成式神经辐射场，并取得了令人鼓舞的成果。GIRAFFE 是另一种与我们工作更为相关的方法，它通过将物体与其背景场景分离来改进 GRAF。然而，这些方法都无法控制图像的条件生成，而这能够实现对生成图像的多种特征操控。为解决这一问题，Jo 等人 [16] 提出了 CG-NeRF。具体而言，CG-NeRF 使用两种类型的条件信息。第一种是条件数据形式，用于将一张图像转换为另一张图像；而我们的研究旨在直接从噪声向量生成新图像。第二种是 CG-NeRF 使用 CLIP（一种自然语言处理技术） [34]中的文本作为条件；而我们的模型使用数值条件，这些条件可以通过改变数值进行插值，从而在图像合成中实现连续特征操控，以表示与给定数值相关的强度变化。

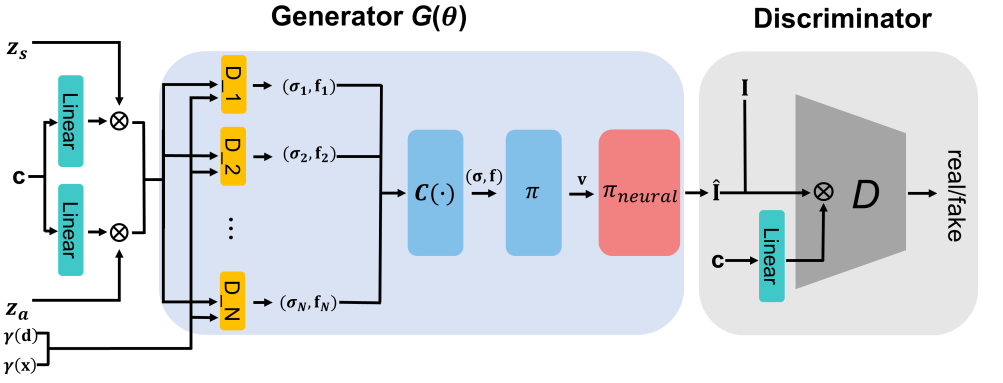


图 2：所提出的 C^3G-NeRF 的整体概览。由于我们的模型受到 GIRAFFE 架构的启发，因此模型使用 $N-1$ 个解码器和一个合成算子来生成物体和背景。D_i 表示第i个解码器，$C(\cdot)$ 表示合成算子。解码器接收经过位置编码 $\gamma(\mathbf{x})$ 的三维坐标向量和观察方向 $\gamma(\mathbf{d})$，其中 γ表示位置编码函数。此外，解码器还接收由线性层编码的条件向量 **c**、形状编码**z**s和外观编码**z**a。利用合成算子$C(\cdot)$ 合成各解码器的输出，然后对结果进行体积渲染，从而生成合成特征向量**v** 。特征向量**v** 经过神经渲染模块 π_{neural}。在此过程中，生成器$G(\theta)$ 合成一张伪造图像 $\hat{\mathbf{I}}$。判别器D 接收一张真实图像**I** 或一张由条件标签**c**投影的伪造图像 $\hat{\mathbf{I}}$。

3 方法

我们的目标是构建一个可进行条件控制的3D感知图像合成框架，确保以连续数值表示条件强度。给定一个带标签的真实世界二维图像数据集，3D感知图像生成器 G 接收条件和相机位姿，分别记为**x** 和 **d**，以及用于表示形状和外观的潜向量，即 **z**s 和 **z**a。随后，**G** 生成与输入条件相对应的图像 **I**。训练时，将数据集中的真实图像和 **I** 输入判别器**D**。图 2 展示了我们模型的概览。

3.1 条件神经辐射场

传统的神经辐射场使用全连接层，将三维坐标$\mathbf{x}=(x,y,z)$ 和观察方向$\mathbf{d}=(\theta,\phi)$映射为体密度 σ以及视角相关的 RGB 颜色值 R、G 和 B。然而，若干研究 [35] 观察到，深度学习技术难以表示高频细节，尤其是在输入为低维时。为使输入更加多样化，神经辐射场在全连接层之前对输入 [41] **x** 和 **d**进行位置编码：

$$\gamma(p,\ L)=(\sin(2^0\pi p),\cos(2^0\pi p),\sin(2^1\pi p),\cos(2^1\pi p),\cdots,\sin(2^{L-1}\pi p),\cos(2^{L-1}\pi p)) \tag{1}$$

其中，$\gamma(\cdot)$表示位置编码，L 是位置编码的维度，而p 是**x** 和 **d**的一个标量分量。要扩展到生成式神经辐射场，

形状和外观编码$\mathbf{z_s}$和$\mathbf{z_a}$按如下方式输入多层感知机：

$$(\gamma(\mathbf{x}),\gamma(\mathbf{d}),\mathbf{z_s},\mathbf{z_a}) \mapsto (\sigma,\mathrm{R},\mathrm{G},\mathrm{B}). \tag{2}$$

在我们的模型中，通过将原本为三维的颜色 [32], 扩展到维度为M_f的特征空间，将生成式神经辐射场替换为生成式神经特征场，如下所示：

$$\mathrm{h}_\theta : \mathbb{R}^{L_\mathbf{x}} \times \mathbb{R}^{L_\mathbf{d}} \times \mathbb{R}^{M_\mathbf{s}} \times \mathbb{R}^{M_\mathbf{a}} \mapsto \mathbb{R} \times \mathbb{R}^{M_f}, \quad (\gamma(\mathbf{x}),\gamma(\mathbf{d}),\mathbf{z_s},\mathbf{z_a}) \mapsto (\sigma,\mathbf{f}), \tag{3}$$

其中，h_θ表示生成式神经特征场，$L_\mathbf{x}$和$L_\mathbf{d}$分别表示位置编码输出的维度，而$M_\mathbf{s}$和$M_\mathbf{a}$分别表示形状和外观的潜在编码维度。

在这项工作中，我们通过逐元素乘积将条件投影到潜向量 [28], $\mathbf{z_s}$和$\mathbf{z_a}$上。投影前，维度为$M_\mathbf{c}$的条件向量$\mathbf{c}$经全连接层编码，使其维度与$\mathbf{z_s}$和$\mathbf{z_a}$的维度相同。形状和外观的条件编码$\mathbf{c_s}$和$\mathbf{c_a}$构造如下：

$$\mathbf{c_s} = L_\mathbf{s}(\mathbf{c}) * \mathbf{z_s}, \quad \mathbf{c_a} = L_\mathbf{a}(\mathbf{c}) * \mathbf{z_a}, \quad L_\mathbf{s} : \mathbb{R}^{M_\mathbf{c}} \mapsto \mathbb{R}^{M_\mathbf{s}},\ L_\mathbf{a} : \mathbb{R}^{M_\mathbf{c}} \mapsto \mathbb{R}^{M_\mathbf{a}}, \tag{4}$$

其中，$L_\mathbf{s}$和$L_\mathbf{a}$分别是形状和外观的编码层，$*$表示逐元素乘法。我们采用这些条件投影来替代生成式特征场中的传统潜向量：

$$h_\theta : \mathbb{R}^{L_\mathbf{x}} \times \mathbb{R}^{L_\mathbf{d}} \times \mathbb{R}^{M_\mathbf{s}} \times \mathbb{R}^{M_\mathbf{a}} \mapsto \mathbb{R} \times \mathbb{R}^{M_f}, \quad (\gamma(\mathbf{x}),\gamma(\mathbf{d})),\mathbf{c_s},\mathbf{c_a} \mapsto (\sigma,\mathbf{f}). \tag{5}$$

3.2 场景合成

受 GIRAFFE 的启发，我们的模型能够利用多个特征向量 [32]分离场景和单个物体。因此，当图像中存在$N-1$个物体和一个背景场景时，模型中会有N个生成式特征场。为了合成N个实体，合成算子$C(\cdot)$会合成这N个实体的所有特征场。每个单独的实体$h^i_{\theta i}$都会输出一个体密度 θ_i和一个特征向量$\mathbf{f}_i$。按照 GIRAFFE 中的方法，我们使用基于密度加权均值的合成来合成实体的各个组成部分。合成算子的数学表达式可表示为

$$C(\mathbf{x},\mathbf{d},\mathbf{c}) = \left(\sigma, \sum_{i=1}^{N} \frac{\sigma_i \mathbf{f}_i}{\sigma}\right), \tag{6}$$

其中$\sigma = \sum_{i=1}^{N} \sigma_i$。

3.3 体渲染与神经渲染

对于场景渲染，我们的模型将相机射线 $r(t) = \mathbf{o} + t\mathbf{d}$渲染为特征向量[32], ，随后通过神经渲染将特征向量转换为合成图像。我们的方法基本遵循神经辐射场中使用的离散化体渲染方法[26]；不过，具体方法采用 GIRAFFE 的做法，将特征渲染为图像，可表示为

$$\mathbf{v} = \sum_{j=1}^{N_\mathbf{s}} T_j(1 - e^{-\sigma_j \delta_j}) f_j, \quad \text{where} \quad T_j = \prod_{k=1}^{j-1} e^{-\sigma_j \theta_j}, \tag{7}$$

其中 $\mathbf{v}$ 表示最终特征向量，T 表示沿投射射线累积的透射率，$N_{\mathbf{N}}$ 表示在任意相机位姿 ξ下，沿投射射线的采样点数量。通过沿相机射线采样，我们为每个点使用一个特征向量 f_i和一个密度σ_i 。此外，出于成本效益考虑，特征图像的 16^2 分辨率为 $H_V \times W_V$ 。包括 GRAF 和神经辐射场在内的现有研究 [26, 39],通常采用体渲染方法进行 3D到2D投影。然而，在我们的模型中，由于体渲染合成的是特征图像，而非彩色高分辨率图像，因此还需要额外的 2D 神经渲染网络。这个额外的神经渲染网络，

$$\pi_{neural} : \mathbb{R}^{H_V \times W_V \times M_f} \mapsto \mathbb{R}^{H \times W \times 3}, \tag{8}$$

通过对特征图进行上采样，将体渲染的输出映射为合成图像 $\hat{\mathbf{I}} \in \mathbb{R}^{H \times W \times 3}$。神经渲染网络的架构与现有研究中的神经渲染算子相似 [32]；不过，我们采用残差模块而非传统卷积网络，以提升训练速度和性能。

3.4 训练细节

训练时，我们从先验分布 $\mathbf{p_s}$、$\mathbf{p_a}$ 和 $\mathbf{p}_\xi$ 中随机采样潜向量 $\mathbf{z_s}$ 和 $\mathbf{s_a}$，以及相机位姿 ξ。先验分布 $\mathbf{p_s}$ 和 $\mathbf{p_a}$ 定义为高斯分布，相机位姿分布 ξ 则设为均匀分布。在生成器 G 中，我们将条件标签投影到潜向量 [28]$\mathbf{z_s}$ 和 $\mathbf{z_a}$。类似地，在判别器 D 中，条件标签 $\hat{}$ 被投影到真实图像 I 或合成图像 I。真实图像 I 从训练数据集随机采样，该数据集服从分布 $\mathbf{p_I}$。我们使用带 R1梯度惩罚 [25] 的 GAN损失，如下所示：

$$\begin{aligned}\mathcal{L}(G,D) =& \mathbb{E}_{\mathbf{z_s}\sim p_\mathbf{s}, \mathbf{z_a}\sim p_\mathbf{a}, \xi\sim p_\xi}\left[-\log(D(G(\mathbf{z_s},\mathbf{z_a},\xi,\mathbf{c})))\right] \\ &+ \mathbb{E}_{\mathbf{I}\sim p_\mathbf{I}}\left[-\log(1-D(\mathbf{I})) + \lambda\|\nabla D(\mathbf{I})\|^2\right].\end{aligned} \tag{9}$$

我们通过使用上述损失函数进行零和博弈，让所提出模型的生成器和判别器相互竞争，从而对其进行训练。我们分别使用 RMSprop优化器 [38]，学习率为 3×10^{-4}，并采用 1×10^{-4} 来训练生成器和判别器。模型在一块内存为 48GB 的 A6000 GPU 上训练，批量大小为 32。对于 64^2 分辨率和 128^2 分辨率，我们设置 $M_f = 128$；对于 256^2 分辨率，我们设置 $M_f = 256$ 。我们在模型中使用 ReLU激活 [29] 作为激活函数，但神经渲染器的最后一层除外，该层使用 Sigmoid函数 [37]。

3.5 使用残差模块加速训练

我们发现，如果不使用残差模块，就无法利用 GIRAFFE 实现复杂的条件生成 [12, 25]。在本研究的实验中，我们证明，使用普通网络进行条件生成时，性能表现落后。我们认为，这是因为经过条件投影的潜向量距离判别器过远，导致梯度消失。因此，残差模块有助于将信息从判别器传递到经过条件投影的潜向量。此外，由于我们将条件标签投影到潜向量，潜向量的变化范围也随之扩大，而普通网络难以学习这一点。我们在模型的解码器、神经渲染网络和判别器中使用残差模块。我们在第 4 节验证了这一改进的有效性。



(a) (b

)图3：C^{3}G-神经辐射场在 CelebA 和 Cars 数据集上训练后生成的类别条件合成物体旋转。在 (a) 中，每一行表示 CelebA 中一个具有相同潜向量的物体。每一列表示不同的旋转角度。在 (a) 左侧的图中，我们将输入条件设为一名秃头男性；在 (a) 右侧的图中，则将条件设为一名金发微笑女性。在 (b) 中，通过控制水平平移和深度平移，展示了物体与背景的解耦。在仅含单个物体的非结构化二维图像上训练后，我们可以生成$N-1$个物体，并通过复制 N 个解码器，将它们放入同一场景中，如“添加物体”所示。所有图像的分辨率为 128^2。使用 C^{3}G-神经辐射场，可在给定条件下成功实现三维一致性图像生成。

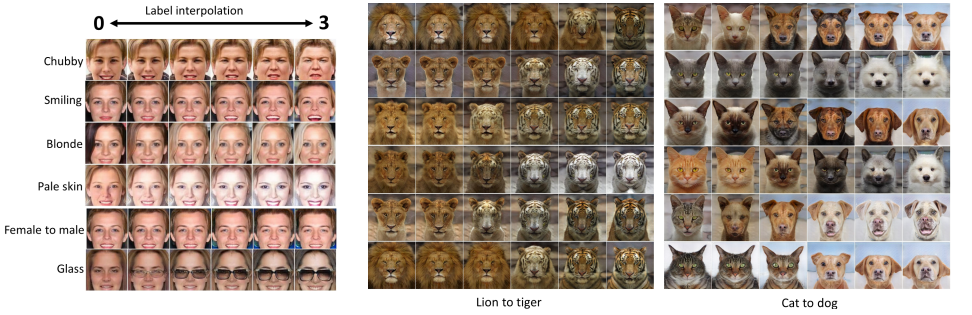
4 实验

本节中，我们在三个真实世界数据集上评估 C^{3}G-NeRF：CelebA [23],AFHQ [7], 和 Cars [1]。我们首先评估条件控制的三维一致性图像生成。随后，我们通过 Fréchet初始距离 [15]评估生成质量。最后，我们评估残差模块，以验证模型中所采用残差模块的有效性。

4.1 三维物体生成中的可控特征

模型评估包括进行物体旋转、水平平移、深度平移和物体添加。评估结果如图 1和 3所示。这些图表明，对于 AFHQ、CelebA 和 Cars 数据集，C^{3}G-NeRF 均能有效学习 3D一致性，即使引入变换，也能保持空间一致性。此外，模型还能够成功捕捉条件输入特征，并据此生成图像。

此外，在图 3 中，我们使用测试时深度和平移水平方向变化的分布外图像，评估 C^{3}G-NeRF。这意味着，C^{3}G-NeRF 可以通过使用超出训练集范围的旋转和平移值，生成超出训练图像分布的图像。此外，我们可以通过 C^{3}G-NeRF 精细控制生成图像中的每个物体和场景；例如，在场景中添加物体时，可以通过在三维空间中平移和旋转来控制每个物体。



(a) (b)

图4：在 CelebA 和 AFHQ 上对条件输入值进行插值和外推。每一行和每一列分别表示 AFHQ 中相同的潜向量（相同物体）以及相同的类别条件值。在 (a) 中，我们展示了条件输入值从零到三范围内对应的条件生成结果。请注意，在训练时，特征仅使用零和一这两个值进行训练，它们表示相应特征是否存在。人脸图像中的特征会随插值和外推输入值平滑变化，这表明连续条件学习已得到充分推进。通过对每个类别的值进行插值，不同类别的特征会在类别条件值的中间状态共存。

4.2 3D物体生成中的连续可控特征

我们通过在 CelebA 数据集中对 40 个条件二进制值进行插值，展示模型操控单独条件的能力，例如丰满、微笑、金发和白皙肤色。尽管由于 CelebA 数据集采用二进制编码，训练过程中的取值范围为零到一，但如图 4 所示，我们将测试过程中的取值范围扩展到零到三。此实验用于评估每种面部特征是否映射到生成器的输入标签。对于不同的条件标签值，C^3G-NeRF 在每种条件的插值和外推方面都表现出色。例如，丰满和微笑条件 [23] 会随条件值平滑变化，无论外推范围如何。这一结果凸显了模型生成分布外图像的能力，例如生成输入标签值高于一时夸张的特征。

我们还考察了与 AFHQ 类别对应的非特征性（隐式）标签。由于其图像特征不够明显，因此更难学习。我们按照图4所示，使用插值后的类别条件输入值生成类别间图像。我们观察到，不同类别的特征会在类别条件值的中间状态共存，这表明特征操控可应用于隐式类别标签。

4.3 定量评估

我们采用传统方法，通过 KID 和 FID 评估图像质量，并对比 20,000 张随机抽取的真实图像和生成图像。评估结果展示了每个数据集和标签的生成质量，并评估了使用随机条件输入在不同分辨率和数据集上生成的图像。

表1 显示，C^3G-NeRF 在条件式 3D 一致生成中取得了出色的 FID 和 KID，无论分辨率如何，都优于条件GIRAFFE，并表明其在高分辨率下具有稳健性。条件GIRAFFE的 FID 较高，说明训练可能存在问题。例如，在 CelebA人脸图像生成中，采用 128^2 分辨率时，

模型	AFHQ（FID）			CelebA（FID）		Cars（FID）		
	64^2	128^2	256^2	64^2	128^2	64^2	128^2	256^2
基线	212.74	226.51	239.01	55.90	83.89	228.66	244.55	266.51
OURS	**26.72**	**25.79**	**28.58**	**5.60**	**7.64**	**44.79**	**43.29**	**31.63**
模型	AFHQ（KID）			CELEBA（KID）		CARS（KID）		
	64^2	128^2	256^2	64^2	128^2	64^2	128^2	256^2
基线	0.200	0.352	0.323	0.119	0.296	0.096	0.166	0.303
OURS	**0.077**	**0.059**	**0.051**	**0.046**	**0.023**	**0.057**	**0.039**	**0.043**

表1：在三个数据集上使用 FID（↓）和 KID（↓）进行定量比较。基线设为使用普通网络的条件GIRAFFE。 64^2, 128^2, 和 256^2 分别为生成图像和真实图像的分辨率。

类别	CARS			类别	AFHQ		
	64^2	128^2	256^2		64^2	128^2	256^2
Peykan	73.54	67.47	68.57	CAT	10.38	13.65	15.48
QUIK	58.73	62.71	49.64	CAT	10.38	13.65	15.48
Samand	54.63	50.53	61.01	DOG	31.64	43.37	51.73
标致 Pars	65.34	66.27	46.45	豹	17.20	14.98	13.42
标致-207I	71.17	71.06	52.99	FOX	25.97	28.01	22.50
PRIDE-111	67.91	62.28	68.84	LION	8.76	12.20	7.61
PRIDE-131	68.70	57.70	65.43	老虎	12.78	8.96	5.93
TIBA2	60.80	61.79	55.57	WOLF	28.39	30.82	14.85
雷诺-L90	67.79	65.51	76.84				
日产-ZAMIAD	78.17	88.83	133.77				
标致-206	63.53	61.55	128.10				
标致-405	71.79	66.15	71.45				
马自达-2000	71.75	77.23	104.53				

表 2：AFHQ 和 Cars 各类别的 FIDs（↓）定量比较。 64^2,128^2, 和 256^2 分别为生成图像和真实图像的图像分辨率。

C^3G-NeRF 的 FID 达到 7.64，较基线降低了 90.9%。

表 2 展示了各个标签对应的 FIDs，反映了 C 所实现的条件解耦[3] G-NeRF。各个标签的 FIDs 与所有标签合并后的结果相近或更优，这表明 C^3G-NeRF 能够有效地学习不同数据集中各个条件对应的特征。此外，即便样本数量很少（例如狐狸，

模型	Cars（KID）		Cars（FID）	
	64^2	128^2	64^2	128^2
pi-GAN	0.105	0.083	137.34	104.51
Ours	**0.057**	**0.039**	**44.79**	**43.29**

表 3：在 Cars 数据集上，将分辨率设为 64^2 和 128^2，时， 360° 生成结果的 FIDs（↓）和 KIDs（↓）定量比较，并与 pi-GAN 进行比较。

AFHQ 中的狮子、老虎），相近的 Fréchet 初始距离表明，C^3G-NeRF 擅长从数据中学习，不受其未对齐分布的影响。

表3显示，我们的模型在 Cars 数据集上优于 pi-GAN，能够有效解耦类别条件。通常，条件模型的表现不如无条件模型，因此 pi-GAN 不适用于 360° 生成。此外，EG3D 在 Cars 数据集上的表现不佳，因为它需要与样本相对应的相机位置，而 EG3D 可以通过独立算法获取面部数据集的相机位置 [8, 22]。因此，考虑到 360° 在真实世界数据集上生成的通用性，我们采用 GIRAFFE。

4.4 残差模块评估

我们评估了带有残差模块的 C^3G-NeRF 生成图像的质量，并将其与采用普通网络的条件 GIRAFFE 进行比较。该实验突显了在 C [12, 25] G-NeRF^{3}架构中引入残差模块的影响。对三个数据集的 FID 和 KID 比较（表1）表明，C^3G-NeRF 显著优于采用普通网络的条件 GIRAFFE，这支持了我们的假设：残差模块能够提升生成质量，并且对条件式 3D 感知生成至关重要。残差模块有助于梯度从判别器的输出流向生成器的条件输入，从而促进条件信息的学习。图 12 展示了定性比较，详见补充材料。

5 结论

我们提出了一种新颖模型 C^3G-NeRF，用于 3D 感知图像生成中的条件式连续特征操控。我们的方法通过编码层将条件标签投影到生成器的潜向量和判别器中间层，从而解耦数据集特征。C^3G-NeRF 利用残差模块优化 3D 感知条件训练。通过对条件输入值进行插值和外推，我们实现了精确且保持 3D 一致性的图像生成与特征操控。

6 致谢

本研究获得韩国国家研究基金会（NRF）资助，资助经费来自韩国政府（MSIT）（编号：RS-2023-00251528）。

参考文献

[1]Yousef Ashrafi。伊朗汽车，2022年8月。网址 https://www.kaggle.com/datasets/usefashrfi/iran-used-cars-dataset。

[2] Yoshua Bengio、Aaron Courville 和 PascalVincent。表征学习：综述与新视角。*IEEE*模式分析与机器智能汇刊，35(8):1798–1828，2013年。

[3] Eric R. Chan、Connor Z. Lin、Matthew A. Chan、Koki Nagano、Boxiao Pan、Shalini De Mello、Orazio Gallo、Leonidas Guibas、Jonathan Tremblay、Sameh Khamis、Tero Karras 和 Gordon Wetzstein。高效几何感知3D生成对抗网络。载于*arXiv*，2021年。

[4] Eric R Chan、Marco Monteiro、Petr Kellnhofer、Jiajun Wu 和 Gordon Wetzstein。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，第5799–5809页，2021年。

[5] Xi Chen、Yan Duan、Rein Houthooft、John Schulman、Ilya Sutskever 和 Pieter Abbeel。InfoGAN：通过信息最大化生成对抗网络实现可解释表征学习

对抗网络。《神经信息处理系统进展》，29:2172–2180，2016年。

[6] Yunjey Choi , Minje Choi , Munyoung Kim , Jung-Woo Ha , Sunghun Kim , 和 Jaegul Choo。Stargan：用于多域图像到图像平移的统一生成对抗网络。载于*IEEE*计算机视觉与模式识别会议论文集，第8789–8797页，2018年。[7] Yunjey Choi、Youngjung Uh、Jaejun Yoo和Jung-Woo Ha。Stargan v2：面向多个域的多样化图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8188–8197页，2020年。[8] Yu Deng、Jiaolong Yang、Sicheng Xu、Dong Chen、Yunde Jia和Xin Tong。弱监督学习实现精准三维人脸重建：从单张图像到图像集。载于*IEEE*计算机视觉与模式识别研讨会，2019年。[9] Yu Deng、Jiaolong Yang、Jianfeng Xiang和Xin Tong。Gram：用于3D 感知图像生成的生成式辐射流形。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10673–10683页，2022年。[10] Robert A Drebin、Loren Carpenter和Pat Hanrahan。体渲染。*ACM SIGGRAPH*计算机图形学，22(4)：65–74，1988年。[11] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sher-jil Ozair、Aaron Courville和Yoshua Bengio。生成对抗网络。《神经信息处理系统进展》，27：2672–2680，2014年。[12] Kaiming He、Xiangyu Zhang、Shaoqing Ren和Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[13] Kaiming He、Xiangyu Zhang、Shaoqing Ren和Jian Sun。深度残差网络中的恒等映射。载于欧洲计算机视觉会议，第630–645页。Springer，2016年。[14] Philipp Henzler、Niloy J Mitra和Tobias Ritschel。走出柏拉图洞穴：通过对抗渲染重建三维形状。载于*IEEE/CVF*国际计算机视觉会议，第9984–9993页，2019年。[15] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler和Sepp Hochreiter。采用双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。《神经信息处理系统进展》，30：6626–6637，2017年。[16] Kyungmin Jo、Gyumin Shim、Sanghun Jung、Soyoung Yang和Jaegul Choo。CG-NeRF：条件生成式神经辐射场。*arXiv*预印本 *arXiv:2112.03517*，2021年。[17] James T Kajiya和Brian P Von Herzen。体密度的光线追踪。*ACM SIG-GRAPH*计算机图形学，18(3)：165–174，1984年。[18] Tero Karras、Timo Aila、Samuli Laine和Jaakko Lehtinen。通过GAN渐进式增长提升质量、稳定性和多样性。*arXiv*预印本 *arXiv:1710.10196*，2017 年。

[19] Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen和Timo Aila。分析并提升StyleGAN的图像质量。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8110–8119页，

2020年.[20] Diederik P Kingma 和 Max Welling. 自编码变分贝叶斯. arXiv 预印本 arXiv:1312.6114, 2013年.[21] Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton. 深度学习. 《自然》, 521(7553): 436–444, 2015年.[22] Taehee Brad Lee. 猫_hipsterizer, 2018年10月. URL https://github.com/ kairess/cat_hipsterizer. [23] 刘子维、罗平、王晓刚和汤晓鸥. 野外面部属性的深度学习. 收录于 IEEE国际计算机视觉会议论文集, 页码 3730–3738, 2015年.[24] Francesco Locatello、Stefan Bauer、Mario Lucic、Gunnar Raetsch、Sylvain Gelly、Bern- hard Schölkopf 和 Olivier Bachem. 挑战无监督学习解耦表示中的常见假设. 收录于国际机器学习会议, 页码 4114–4124. PMLR, 2019年.[25] Lars Mescheder、Andreas Geiger 和 Sebastian Nowozin. 哪些生成对抗网络训练方法确实能够收敛？收录于国际机器学习会议, 页码 3481–3490. PMLR, 2018年.[26] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ra-mamoorthi 和 Ren Ng. 神经辐射场：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, 2021 年 .[27] Mehdi Mirza 和 Simon Osindero. 条件生成对抗网络. arXiv 预印本 arXiv:1411.1784, 2014年.[28] Takeru Miyato 和 Masanori Koyama. 使用投影判别器的条件生成对抗网络. arXiv 预印本 arXiv:1802.05637, 2018年.[29] Vinod Nair 和 Geoffrey E Hinton. 修正线性单元改进受限玻尔兹曼机. 收录于国际机器学习会议, 页码 807–814, 2010年.[30] Thu Nguyen-Phuoc、Chuan Li、Lucas Theis、Christian Richardt 和杨永亮. HoloGAN：从自然图像中无监督学习三维表示. 收录于 IEEE/CVF国际计算机视觉会议论文集, 页码 7588–7597, 2019年.[31] Thu H Nguyen-Phuoc、Christian Richardt、Long Mai、杨永亮和 Niloy Mitra. BlockGAN：从未标注图像中学习三维物体感知场景表示. 《神经信息处理系统进展》, 33:6767–6778, 2020年.[32] Michael Niemeyer 和 Andreas Geiger. GIRAFFE：将场景表示为组合式生成式神经特征场. 收录于 IEEE/CVF计算机视觉与模式识别会议论文集, 页码 11453–11464, 2021年。

[33] Augustus Odena、Christopher Olah 和 Jonathon Shlens. 使用辅助分类器生成对抗网络进行条件图像合成. 收录于 国际机器学习会议, 页码 2642–2651. PMLR, 2017年.

[34] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。发表于国际机器学习会议，第 8748–8763 页。PMLR，2021年。[35] Nasim Rahaman、Aristide Baratin、Devansh Arpit、Felix Draxler、Min Lin、Fred Hamprecht、Yoshua Bengio 和 Aaron Courville。论神经网络的谱偏置。发表于国际机器学习会议，第 5301–5310 页。PMLR，2019年。[36] Sai Rajeswar、Fahim Mannan、Florian Golemo、Jérôme Parent-Lévesque、David Vazquez、Derek Nowrouzezahrai 和 Aaron Courville。Pix2shape：基于视图的表示，从图像中无监督学习三维场景。《国际计算机视觉杂志》，128(10):2478–2493，2020年。[37] Prajit Ramachandran、Barret Zoph 和 Quoc V Le。搜索激活函数。*arXiv* 预印本 *arXiv:1710.05941*，2017年。[38] Sebastian Ruder。梯度下降优化算法综述。*arXiv* 预印本 *arXiv:1609.04747*，2016年。[39] Katja Schwarz、Yiyi Liao、Michael Niemeyer 和 Andreas Geiger。GRAF：用于三维感知图像合成的生成式辐射场。《神经信息处理系统进展》，33:20154–20166，2020年。[40] Vincent Sitzmann、Michael Zollhöfer 和 Gordon Wetzstein。场景表示网络：连续三维结构感知神经场景表示。《神经信息处理系统进展》，32:1121–1132，2019年。[41] Matthew Tancik、Pratul Srinivasan、Ben Mildenhall、Sara Fridovich-Keil、Nithin Raghavan、Utkarsh Singhal、Ravi Ramamoorthi、Jonathan Barron 和 Ren Ng。傅里叶特征使网络能够在低维域中学习高频函数。《神经信息处理系统进展》，33:7537–7547，2020年。[42] Shubham Tulsiani、Nilesh Kulkarni 和 Abhinav Gupta。从未标注图像集合中进行隐式网格重建。*arXiv* 预印本 *arXiv:2007.08504*，2020年。[43] Aäron Van Den Oord、Nal Kalchbrenner 和 Koray Kavukcuoglu。像素循环神经网络。发表于国际机器学习会议，第 1747–1756 页。PMLR，2016年。[44] Jiajun Wu、Chengkai Zhang、Tianfan Xue、Bill Freeman 和 Josh Tenenbaum。通过三维生成对抗建模，学习物体形状的概率潜在空间。《神经信息处理系统进展》，29:82–90，2016年。