基于局部空间自注意力的 3D 感知图像合成

宋成刚[1)], 袁霞[1)], 吴涛[1)], 杨智鹏[2)], 刘伟[3)], 符颖[1)*]

[1)] (成都信息工程大学计算机学院 成都 610225)
[2)] (成都信息工程大学电子工程学院 成都 610225)
[3)] (上海交通大学自动化系 上海 200240)
(fuying@cuit.edu.cn)

摘 要: 神经辐射场(neural radiance fields, NeRF)拥有将 3D 场景建模为连续函数的强大隐式表征能力, 促进了生成模型在非结构化的 2D 图像集合上完成 3D 感知图像合成任务. 然而, 由于忽略了真实世界的 3D 本质或依赖表达能力不足的网络架构, 当前方法在图像合成方面仍然存在保真度低、一致性差和不可控制等挑战. 为此, 提出一种基于自注意力机制的局部空间神经表示方法对 NeRF 生成器进行改良, 实现可控的高质量 3D 感知图像合成. 首先通过结合各对象的位姿信息和预定义的 3D 边界框, 将待合成场景进行解耦表示; 然后利用自注意力局部表示模块为各对象单独建模, 定义组合算子得到完整场景表示; 最后经过渲染模块处理得到最终的 RGB 图像. 实验结果表明, 所提方法能够在对空间点低频采样的情况下完成视角一致的、可控的高质量 3D 感知图像合成任务; 且在高复杂度的 CompCars 公开数据集上进行 64×64 和 256×256 像素分辨率的合成实验, 该方法获得的 FID 相较于基线方法分别提升 4.69 和 2.51; 同样在 Churches 公开数据集上 FID 分别提升 1.93 和 3.56, 进一步突出了该方法的有效性.

关键词: 神经辐射场; 隐式表征; 注意力机制; 3D 感知图像合成
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024-00073

3D-Aware Image Synthesis via Local Spatial Self-Attention

Song Chenggang[1)], Yuan Xia[1)], Wu Tao[1)], Yang Zhipeng[2)], Liu Wei[3)], and Fu Ying[1)*]

[1)] (*School of Computer Science, Chengdu University of Information Technology, Chengdu* 610225)
[2)] (*College of Electronic Engineering, Chengdu University of Information Technology, Chengdu* 610225)
[3)] (*Department of Automation, Shanghai Jiao Tong University, Shanghai* 200240)

Abstract: The neural radiance field (NeRF), endowed with the implicit representation capability to model 3D scenes as continuous functions, facilitates the generation model in accomplishing 3D-aware image synthesis tasks on unstructured 2D image sets. Nevertheless, current approaches face challenges in image synthesis, including low fidelity, poor consistency and uncontrollability, stemming from their inability to fully capture the three-dimensional essence of the real world or their reliance on insufficiently expressive network architectures. To address these limitations, we introduce a locally spatial neural representation method grounded in a self-attention mechanism. This approach augments the NeRF generator, empowering it to achieve controllable high-quality 3D-aware image synthesis. Initially, the scene to be synthesized is decomposed by incorporating the pose information of each object and the predefined 3D boundary box. Subsequently, the self-attention local representation

收稿日期: 2024-02-02; 修回日期: 2024-09-29. **基金项目**: 国家自然科学基金(62402318); 四川自然科学基金(2023NSFSC0469); 四川省重点研发项目(2024YFG0012, 2023YFS0428); 四川省科技厅项目(2023JDZH0034); 成都信息工程大学科技创新项目(KYQN202219); 成都信息工程大学人才引进项目(KYTZ202267); 四川省科技创新苗子工程项目(MZGC20240061). **宋成刚**(1998—), 男, 硕士研究生, 主要研究方向为计算机视觉、3D 重建; **袁霞**(1987—), 女, 博士, 副研究员, 硕士生导师, 主要研究方向为计算机图形学、3D 重建; **吴涛**(1984—), 女, 博士, 教授, 硕士生导师, 主要研究方向为智能计算、高性能并行计算; **杨智鹏**(1985—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为图像处理及应用; **刘伟**(1989—), 男, 博士, 副教授, 博士生导师, 主要研究方向为计算机视觉、计算机图形学; **符颖**(1988—), 女, 博士, 副教授, 硕士生导师, 论文通信作者, 主要研究方向为信号与图像处理、生成模型在机器视觉中的应用.

module models each object distinctly. A defined combination operator is then utilized to obtain the complete scene representation. Finally, the rendering module processes the resulting RGB image. Experimental results show that the proposed method accomplishes the task of high-quality 3D-aware image synthesis with multi-view consistency and controllability, even under conditions of low-frequency spatial point sampling. Furthermore, when conducting synthesis testing with resolutions of 64×64 and 256×256 pixels on the highly complex CompCars public dataset, our proposed method achieved a 4.69 and 2.51 improvement in terms of FID scores compared to the baseline method, respectively. Similarly, on the Churches public dataset, the FID scores were improved by 1.93 and 3.56, respectively, further highlighting the effectiveness of the method.

Key words: neural radiance field; implicit representation; attention mechanism; 3D-aware image synthesis

3D 感知图像合成[1-4]旨在生成具有明确可控的相机姿势的 3D 一致图像. 与 2D 合成图像相比, 3D 感知图像合成需要理解 2D 图像背后的几何结构, 它通常是通过将 3D 表示纳入生成对抗网络(generative adversarial networks, GAN)[5]这类生成模型来实现的. 给定一个非结构化的 2D 图像集合, GAN 能够从中学习一种无监督的 3D 场景表示, 进而通过该表示从新的相机位姿合成多视角一致的图像[1-4,6-8].

目前, 已有许多工作围绕 3D 感知图像合成任务展开, 并取得了令人鼓舞结果, 但其图像质量仍然远远落后于传统的 2D 图像合成[9-14]. 基于体素的表示方法[6]虽然可以生成可解释的、真实的 3D 表示, 但受高内存和高计算成本的限制, 只能实现低分辨率的粗糙细节生成; 具有深度体素表示的卷积方法[7-8]可以创建精细的图像, 但对于黑盒渲染的依赖性使得它们无法保证多视角一致性. 近年来, 研究人员利用神经辐射场(neural radiance fields, NeRF)[15]隐式表示的方法[1-4,16]将场景或对象的几何和外观参数化为神经网络的权重, 采用一种显式的、基于物理的体渲染范式[17-18]进行图像渲染合成, 能够有效地确保合成图像的多视角一致性和显式的相机控制. 然而, 这些方法通常依赖多层感知机(multi-layer perceptron, MLP)网络进行参数化, 而 MLP 架构庞大的参数量会带来计算复杂度和内存需求过高等问题, 将极大地限制它们对场景重建时的表达能力. 由于 MLP 的固有限制, 这些方法在细节表达方面面临困难, 影响合成图像的质量. 在实践中, 这种依赖于表达能力有限的表示网络架构使得生成的图像缺乏真实感、细粒度和一致性, 与传统的 2D 图像合成相比, 仍然存在较大差距. 因此, 对表达能力的进一步提升, 成为解决当前图像合成质量欠佳问题的关键.

进一步, 生成模型还应该具备简单且一致的控制能力, 不局限于完成静态、不可控的多视角合成任务. 换句话说, 在保持其他属性不变的情况下, 希望能够对感兴趣的属性(如物体的几何形状、大小和姿态)进行控制. 为了实现这一目标, 一些工作专注于在没有显式监督的情况下学习解耦表示[19-20]的方法. 然而, 大多数方法忽视了场景的组成性质, 仅限于 2D 层面的操作, 忽略了真实世界的 3D 本质, 可能导致属性间的表示相互纠缠, 并需要在后验的潜在空间中寻找控制机制才能间接地实现对特定属性的约束. 还有一些工作致力于将体素[6-8]、基本元素[21]和辐射场[1,3]等 3D 表示直接合并至生成模型中. 尽管这些方法通过直接在生成过程中控制特定属性获得令人印象深刻的结果, 但它们在处理更高分辨率、更复杂和更逼真的图像(如物体位于非中心位置或背景杂乱的场景)时, 结果的一致性有待提高.

本文将组合的 3D 表示直接合并至生成模型[2,4], 以提高对场景的可控性, 其核心是将前景对象的几何和外观从其背景的几何和外观中分离出来, 允许在潜在空间中对这些属性实现简单操控, 这种分离操作则是通过分别为前景对象和背景构建特征场网络来表征实现的. 因此, 特征场网络的表达能力决定了在多大程度上能够精确地捕捉和再现场景的细节和复杂性.

近年来, 自注意力机制[22]因其强大的表达能力广泛应用于各大领域[23-29]. 本文将其用于构建对象和背景特征场表征网络, 以缓解因表达能力不足带来的合成图像质量欠佳问题. 自注意力网络能够有效地实现全局上下文信息的捕捉和细粒度特征的表达, 在 3D 结构的精确建模和视角一致性渲染中发挥关键作用. 然而, 由于 3D 数据具有

更高的维度和复杂性, 处理和生成 3D 感知图像需要更多的计算资源. 由于自注意力机制需要计算每对位置之间的关联程度, 因此大大增加了辐射场隐式表示的计算复杂度. 受到 2D 图像生成中 patch 自注意力方法[30-32]的启发, 本文提出一种基于自注意力机制的局部空间隐式表示的 3D 感知图像合成方法. 该方法利用基于自注意力的 NeRF 表示来鼓励多视角一致性, 允许从广泛的相机姿势进行图像渲染; 采用自注意力机制进行局部隐式场景表示比全连接网络更能表达精细细节, 合成结果也更加清晰. 完整像素投射产生的辐射场被划分为多个非重叠均匀的 3D 子空间辐射场, 每个子空间用于局部场景的表征. 通过在每个局部空间内应用自注意力机制, 可以在局部范围内实现全局感知, 并大大降低计算复杂度. 特别地, 由于背景复杂度较低, 本文在对象特征场网络结构的基础上进行简化设计, 使背景特征场网络能够有效地适应不同场景下的复杂性, 保持或提升合成图像的质量.

本文还设计一种新颖的 2D 神经渲染器, 并将其与 3D 表示相结合, 从空间和通道 2 个层面对体渲染输出逐步细化整合, 以提升分辨率. 其中, 空间层面通过细化像素级别的渲染细节, 增强图像的局部清晰度和纹理质量; 通道层面则通过整合和强化多维特征, 提升图像的语义信息和整体视觉效果. 这种结构设计在有效地提升输出分辨率的同时, 还降低计算开销, 并提升了渲染速度.

本文在单对象公开数据集 FFHQ[10], Cats[33], Chairs[34]和 CelebA[35]上以及复杂度更高的现实数据集 CompCars[36]和 Churches[37]上进行实验, 以验证本文方法的合理性, 并获得了 3D 感知图像合成结果. 实验结果表明, 本文方法除了可以推广到新的视点完成新视角合成任务外, 其强大的表达能力还允许在低频采样(16 个采样点)的情况下实现视角一致的、高保真度的图像合成; 同时, 将组合场景合并至生成模型的设计, 实现了对场景的简单控制.

1　相关工作

1.1　神经表示和渲染

近年来, 神经隐式表示在 3D 建模和图像合成中得到快速发展[15,38-39]. 与网格[40]、点云[41-42]和体素[43-44]等经典的离散模型相比, 神经隐式表示使用神经网络对场景进行建模, 具有优越的空间连续性、分辨率无关性和表现力. 作为神经渲染中最流行的隐式表示, NeRF[15]能够以高保真度和精确几何图形渲染复杂的 3D 场景. NeRF 使用隐式坐标函数表示场景, 该函数将每个 3D 坐标位置和其 2D 视角方向映射为与视角相关的颜色(亮度)和与视角无关的体积密度, 并使用体渲染方法生成新视角的图像. 鉴于 NeRF 强大的表征能力, 本文使用其生成变体作为对象级表示. 与多视图监督训练方式不同, 本文从非结构化的单目图像数据集合上训练一个生成模型, 缓解密集捕获单场景带来的过拟合及泛化性差等问题. 该模型允许对生成的场景进行可控的逼真图像合成.

1.2　自注意力表示

人类视觉系统以快速扫描全局图像的方式获取所有需要着重注意的信息, 并将更多的注意力注入这些区域, 在获得更多有效的信息的同时抑制无关信息. 类似地, 深度学习中的自注意力机制[22]在处理特征信息时, 也会从众多的信息中关注对当前任务更有帮助的特征信息. 这种机制的有益性能已经在 2D 视觉领域得到了广泛讨论和证明[23-29]. 最近, 一些工作将其与 NeRF 相结合, 以促进 3D 视觉领域的发展. Wang 等[45]提出一种光线自注意力方法 IBRNet, 能够动态地关联来自多个源视图的外观信息; NerFormer[46]在光线的不同视图之间采用自注意力机制聚合特征, 并从聚合的特征中学习辐射场; GNT[47]采用自注意力机制, 通过直接回归颜色合成视图, 不需要 NeRF 的体渲染技术. 与上述方法不同, 本文采用全自注意力构建的生成网络编码场景的几何和外观, 以实现视角一致的逼真图像合成. 为了降低计算成本, 本文在局部空间内进行自注意力隐式表示, 确保了局部范围内的全局感知.

1.3　生成式的 3D 感知图像合成

扩散模型(diffusion models, DMs)[48]作为近年来生成模型中的新兴技术, 已经在图像生成领域展现出强大的表现力[49-51], 但其推理过程的迭代性质不仅使得潜在空间的控制变得复杂和难以预测, 而且带来了计算上的高要求. 相比之下, GAN[5]只需要执行一次前馈通道来生成, 提供了更直接和高效的潜在空间控制机制[52-54], 这种优良特性决定了它在 3D 感知图像合成方面的普适性.

GAN 由于缺乏对场景参数的显式控制, 并且不能保证 3D 一致性, 因此限制了其在 3D 感知图像

合成方面的应用. 为此, 许多工作将 GAN 与 3D 表示相结合, 以改善3D感知图像的合成结果. 一些工作[55-56]选择直接在 3D 数据上进行训练, 这是不切实际的, 因为获取这样的 3D 数据属于资源密集型任务; VON[57]和 PrGANs[58]首次提出体素化作为 3D 表示进行图像渲染; 在此基础上, HoloGAN[7]和 BlockGAN[8]结合了隐式的3D表示: 然而, 由于体素分辨率的限制, 这些方法的图像质量和一致性较差. Henzler 等[6]和 Szabó 等[59]分别学习生成显式的 3D 体素和网格, 但生成的形状和图像质量有限.

本文采用基于坐标的隐式体积表示, 并提出一种全自注意力机制实现的生成器网络来定义这种 3D 表示. 这种直接在 3D 空间中对图像的形成过程进行建模的方式, 可以更好地实现对象间的解耦和更可控的合成.

2 本文方法

在没有任何额外监督的情况下, 本文方法可以从原始图像集合中学习辐射场表示. 在训练时, 生成的图像被定向到传统的卷积判别器进行对抗训练; 在测试时, 辐射场可以从任意相机姿势渲染, 以产生视角一致的、可控的高质量图像.

本文方法的架构如图 1 所示, 分为 4 部分: 对象解耦模块对场景中包含背景在内的各对象定义仿射变换操作, 实现对对象的可控操作; 局部空间场景表征模块采用自注意力机制对各对象进行局部隐式表示, 提高对精细细节的表征能力; 特征场组合模块将各对象特征场进行整合, 得到整个场景的 3D 表征; 渲染模块先将场景体渲染为分辨率较低的特征图, 以减小计算量和渲染时长, 然后通过 2D 神经渲染器渲染特征场得到最终合成结果.

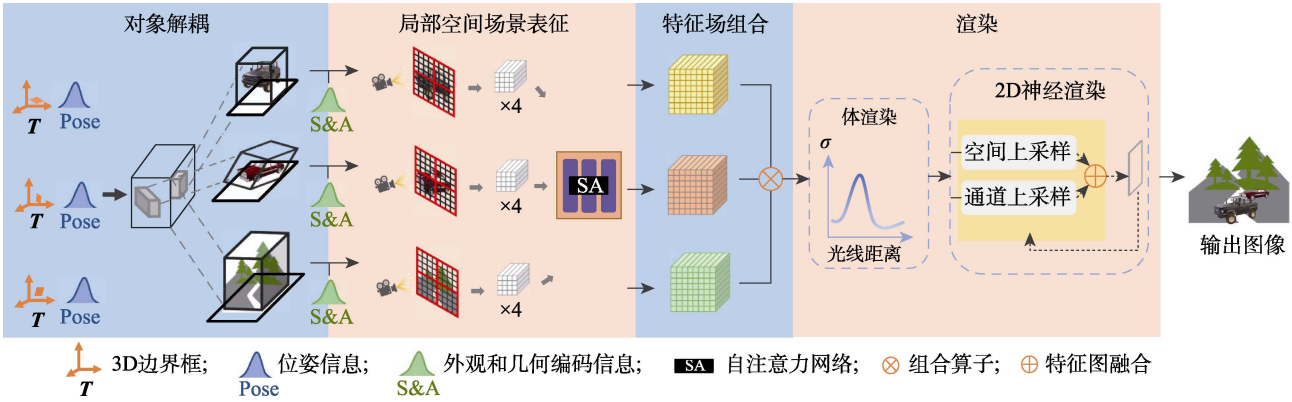


图 1 本文方法架构

2.1 场景解耦

在NeRF及其后续诸多改进工作[1,3,15]中, 由于仅用单个模型表达整个场景, 用户不能控制场景中单个感兴趣的对象(物体), 如控制其位姿、大小和外观. 因此可预先为场景中的每个对象(背景也视为对象)定义一个3D边界框, 以捕获各对象在相机空间中的范围、位置和旋转. 边界框定义 $\boldsymbol{T}=\{\boldsymbol{s},\boldsymbol{t},\boldsymbol{R}\}$, 它从数据集的相关分布中采样得到. 其中, $\boldsymbol{s},\boldsymbol{t}\in\mathbb{R}^3$ 分别表示对象在场景空间中 x,y,z 这 3 个维度上的尺寸缩放和平移参数; $\boldsymbol{R}\in O(3)$ 表示对象在场景空间中的旋转矩阵. 特别地, 将背景的缩放参数和平移参数固定为跨越整个场景, 并以场景空间原点为中心.

2.2 对象特征场表示

图 2 所示为特征场的结构, 包括左上部分的对象特征场, 左下部分的背景特征场和右上部分的自注意力块结构.

(1) NeRF. 将场景定义为一个连续的体积函数, 以空间位置 $\boldsymbol{x}\in\mathbb{R}^3$ 和光线方向 $\boldsymbol{d}\in S^2$ 作为输入, 输出发射颜色 $\boldsymbol{c}\in\mathbb{R}^3$ 和体积密度 $\boldsymbol{\sigma}\in\mathbb{R}^+$. 该过程形式化为 $G_\theta\left(\gamma_{\mathrm{p}}(\boldsymbol{x}),\gamma_{\mathrm{p}}(\boldsymbol{d})\right)\rightarrow(\boldsymbol{c},\boldsymbol{\sigma})$. 其中, G_θ 表示参数化为 θ 的 MLP 网络; $\gamma_{\mathrm{p}}(\cdot)$ 表示位置编码函数, 用来将输入的每个分量映射到更高维度的频率空间. 这样做的好处在于促进神经网络克服其频谱偏差, 以再现输入信号的高频细节, 公式为

$$\gamma_{\mathrm{p}}(t,L)=\left(\sin\left(2^0t\pi\right),\cos\left(2^0t\pi\right),\cdots,\right.$$
$$\left.\sin\left(2^{L-1}t\pi\right),\cos\left(2^{L-1}t\pi\right)\right).$$

其中, t 表示标量输入($\boldsymbol{x}$ 和 $\boldsymbol{d}$ 的分量); 超参数 L 表示映射中使用的最高频率, 可用于控制场景函数的平滑度[60].

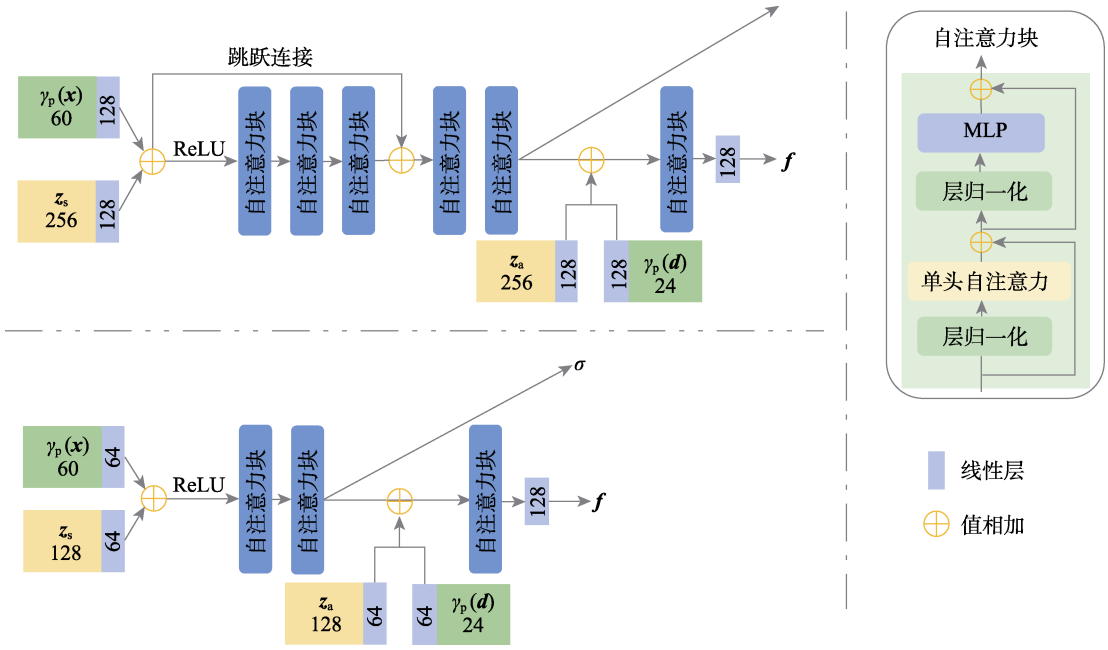


图 2　特征场的结构

(2) 生成 NeRF. 作为将 NeRF 与生成模型结合的首个工作，GRAF[3]从未知位姿信息的图像集合中训练模型，并将 MLP 条件化为对象或背景的形状和外观编码 $\boldsymbol{z}_s, \boldsymbol{z}_a \sim \mathcal{N}(0, \boldsymbol{I})$. 该过程形式化为

$$G_\theta\left(\gamma_p(\boldsymbol{x}), \gamma_p(\boldsymbol{d}), \boldsymbol{z}_s, \boldsymbol{z}_a\right) \rightarrow (\boldsymbol{c}, \boldsymbol{\sigma}).$$

与 GRAF 不同，本文需要对解耦后的各个对象单独建模，然后将其整合得到完整场景. 因此，对于每个查询的 3D 空间点 $\boldsymbol{x}$，获取的是一个体积密度 $\boldsymbol{\sigma} \in \mathbb{R}^+$ 和一个多通道特征 $\boldsymbol{f} \in \mathbb{R}^{M_f}$，而不再是 RGB 颜色；其中，$M_f$ 表示特征通道的数量.

(3) 自注意力表示. 对第 i 个对象特征场进行隐式表示的过程为 $G_\theta\left(\gamma_p(\boldsymbol{x}), \gamma_p(\boldsymbol{d}), \boldsymbol{z}_s, \boldsymbol{z}_a\right) \rightarrow (\boldsymbol{f}, \boldsymbol{\sigma})$.

设给定相机姿态下的渲染图像为 $\hat{I} \in \mathbb{R}^{H \times W}$，即需要对每个对象特征场建模的光线数量为 $H \times W$. 为此，根据针孔相机模型，定义从相机原点 $\boldsymbol{o}$ 沿方向 $\boldsymbol{d}$ 投射产生的光线 $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$. 相应地，近边界和远边界分别定义为 t_n, t_f.

Step1. 3D 点采样. 遵循 NeRF 中的分层采样技术近似难以处理的体积投影积分. 每条光线被划分为 N_s 个均匀间隔的箱子，并在每个箱子内均匀采样一个样本点. 第 i 个采样点定义为

$$\boldsymbol{x}_i = \boldsymbol{r}(t_i),$$

其中，$t_i \sim u\left(t_n + i - 1/N_s(t_f - t_n), t_n + i/N_s(t_f - t_n)\right)$.

Step2. 将 $\boldsymbol{x}$ 转换至该对象对应的 3D 边界框(场景空间)中，公式为

$$k(\boldsymbol{x}) = \boldsymbol{R} \cdot \begin{bmatrix} s_1 & & \\ & s_2 & \\ & & s_3 \end{bmatrix} \cdot \boldsymbol{x} + \boldsymbol{t}.$$

Step3. 位置编码函数升高维度 $\gamma_p\left(k^{-1}(\cdot)\right)$.

Step4. 3D 子空间场分割. 如图 1 中的局部场景表征模块所示，首先定义大小为 p 的窗口，将 2D 像素平面分割为多个非重叠均匀像素块；然后对每个局部像素块投射产生的辐射场单独表示. 该过程形式化为

$$(H \times W \times N \times D) \rightarrow (H/p \times W/p, p \times p \times N \times D).$$

Step5. 对第 j 块 $\left(j = [1, 2, \cdots, H/p \times W/p]\right)$ 局部像素投射产生的子辐射场，可以获取所有采样点 $N_j = (p \times p \times N_s)$ 和对应的特征 $F_j^k \in \mathbb{R}^D$，$k = [1, 2, \cdots, N_j]$；其中，D 表示此时每个采样点的特征通道数.

Step6. 对子辐射场中各采样点对应特征的通道级进行自注意力操作. 将查询矩阵 $\boldsymbol{Q}$，键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$ 定义为

$$\begin{cases} \boldsymbol{Q} = \left(\mathrm{LN}\left(F_j^k\right)\right) \\ \boldsymbol{K} = \left(\mathrm{LN}\left(F_j^k\right)\right), \\ \boldsymbol{V} = \left(\mathrm{LN}\left(F_j^k\right)\right) \end{cases}$$

LN 表示层归一化(layer normalization, LN)操作.

Step7. 自注意力得分计算. 公式为

$$\hat{\boldsymbol{F}}_j^k = \boldsymbol{V} \otimes \mathrm{softmax}\left((\boldsymbol{Q}, \boldsymbol{K}) \times s\right).$$

其中，$\otimes$ 表示矩阵乘法；s 表示缩放因子.

Step8. 每个子辐射场经过自注意力网络表征后再将其整合，得到该对象的完整特征场表征. 在实践中，对于不同的实体(对象或背景)，通过堆叠不同数量的自注意力块对其进行表征.

在实践中，本文在场景空间中进行体渲染，并在图像空间中评估特征场的操作，允许在场景中排列多个前景对象；所有对象特征场共享其权重.

2.3 特征场组合

将场景解耦定义为 N 个实体的组合，其中，前 $N-1$ 个是场景中的对象物体，最后一个表示背景. 为了将各对象特征场及其背景特征场整合得到完整场景表示，定义一种有效的组合算子. GIRAFFE[2]指出，密度加权平均的组合算子 C 是简单且合理有效的，它能够确保梯度向着密度大于 0 的所有实体流动.

本文遵循相同的方式对 N 个实体的特征场进行组合. 对于单个实体预测特征场，得到给定点 $\boldsymbol{x}$ 和视角方向 $\boldsymbol{d}$ 的密度 $\boldsymbol{\sigma}_i \in \mathbb{R}^+$ 和特征向量 $\boldsymbol{f}_i \in \mathbb{R}^{M_f}$. 完整的场景表示为

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

其中，$\boldsymbol{\sigma}=\sum_{i=1}^{N}\boldsymbol{\sigma}_i$；$\boldsymbol{x}$ 处的总体密度直接通过将单个密度求和得到，$(\boldsymbol{x},\boldsymbol{d})$ 处的特征则通过密度加权平均获得.

2.4 渲　染

为了提高合成效率，本文遵循 GIRAFFE[2]中的设置，从 16×16 像素的分辨率渲染图像；然后通过 2D 神经渲染器，将低分辨率的特征映射上采样至更高分辨率的 RGB 图像. 因此，已有的将各像素点投射光线上的采样点直接体渲染到 RGB 颜色值的设置不再适合本文的任务. 为了得到高质量的合成结果，对体渲染公式进行拓展：首先将每条光线上的采样点映射到该像素的高维特征向量 $\boldsymbol{f}$，以保留更多细节信息；然后在 $\boldsymbol{f}$ 上进一步利用 2D 神经渲染器获取最终的合成结果.

(1) 3D 体渲染. 对于给定的相机外参 ξ，设给定像素沿相机光线 $\boldsymbol{d}$ 的采样点为 $(\boldsymbol{x}_j)_{j=1}^{N_s}$，$(\boldsymbol{\sigma}_i,\boldsymbol{f}_i)=C(\boldsymbol{x}_j,\boldsymbol{d})$ 为相应场的密度和特征向量. 体渲染算子 $\boldsymbol{\omega}$ [18]拓展为

$$\begin{cases}\boldsymbol{\omega}:(\mathbb{R}^+\times\mathbb{R}^C)^{N_s}\to\mathbb{R}^{M_f}\\ \{\boldsymbol{\sigma}_j,\boldsymbol{f}_j\}_{j=1}^{N_s}\to\boldsymbol{f}\end{cases}.$$

采用文献[15]中的数值积分法，可得

$$\boldsymbol{f}=\sum_{j=1}^{N_s}\tau_j\alpha_j\boldsymbol{f}_j \tag{2}$$

其中，$\tau_j=\sum_{k=1}^{j-1}(1-\alpha_k)$ 表示透射率；$\alpha_j=1-\mathrm{e}^{-\sigma_i\delta_j}$ 表示 $\boldsymbol{x}_j$ 的 alpha 值，$\delta_j=\|\boldsymbol{x}_{j+1}-\boldsymbol{x}_j\|_2$ 表示相邻采样点之间的距离. 通过计算每个像素的 $\boldsymbol{\omega}$ 值，即可得到整幅特征图像 $I_V \in \mathbb{R}^{H_V\times W_V\times M_f}$.

(2) 2D 神经渲染器. 旨在将低分辨率的高维特征图映射上采样至更高分辨率的 RGB 图像，该过程可以形式化为

$$U_\theta^{\text{neural}}: I_V \in \mathbb{R}^{H_V\times W_V\times M_f}\to \hat{I}\in\mathbb{R}^{H\times W\times 3}.$$

如图 3 所示，通过将 2D 卷积神经网络与上采样技术巧妙结合构建神经渲染器. 在设计中充分地综合空间层面和通道层面的因素，以逐渐升高图像分辨率. 其中，主分支将卷积与空间上采样技术结合，强调对图像空间信息的敏感性，卷积操作捕获像素间的局部结构和空间关系，以保证空间上采样的空间增强和细节保留；次分支则突出对通道信息的处理，引入超分辨率中提出的 pixelshuffle 技术[61]，通过重新排列通道的方式实现高分辨率的转换. 卷积操作先将某一阶段的特征图映射到 RGB 图像，然后 pixelshuffle 单独对各通道进行重新排列，实现 2 倍上采样. 在每个空间分辨率下，次分支输出都会和主分支输出融合，并添加至下一次输出. 实验发现，神经渲染器的创新设计使得模型能够更全面地捕获和学习图像特征，同时保证更强大的梯度流动和避免不必要的信息丢失. 在逐渐提高分辨率的过程中，通过卷积操作和上采样技术的协同作用，模型能够逐渐细化图像细节，生成更真实和精细的渲染结果.

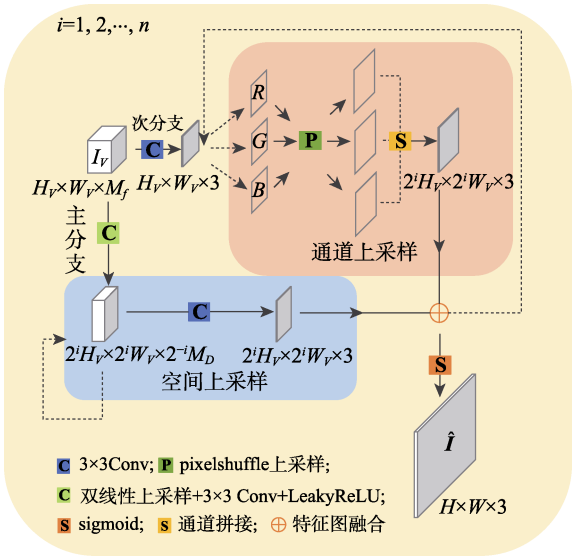


图 3　神经渲染器流程

2.5 网络架构

2.5.1 生成器架构

本文使用自注意力网络分别参数化对象和背景特征场, 将输入点 $\boldsymbol{x}\in\mathbb{R}^3$ 和视角方向 $\boldsymbol{d}\in S^2$, 以及潜在的形状和外观编码 $\boldsymbol{z}_{\mathrm{s}},\boldsymbol{z}_{\mathrm{a}}\sim\mathcal{N}(0,\boldsymbol{I})$ 映射到一维密度 σ 和 M_f 维特征向量 $\boldsymbol{f}$ (如图 2 所示). 与 NeRF 类似, 将 $\boldsymbol{x}$ 和 $\boldsymbol{d}$ 传递给网络之前对其进行位置编码, 将它们分别嵌入到 $L_{\boldsymbol{x}}=2\times3\times10=60$ 和 $L_{\boldsymbol{d}}=2\times3\times4=24$ 维空间中. 对于对象特征场, 使用 6 个自注意力块(中间维度为 128), 并在第 3 层处添加一个跳跃连接, 这样做已被证明可以改善 3D 重建的结果[39]. 对于背景特征场, 由于其复杂度较低, 因此使用 3 个自注意力块(中间维度为 64). 对象特征场的潜在编码从 256 维高斯分布中采样; 背景特征场的潜在编码从 128 维高斯分布中采样.

2.5.2 判别器架构

本文遵循 GIRAFFE[2]的设置, 将判别器参数化为具有 LeakyReLU 激活的全卷积网络, 并分别为 64×64 和 256×256 像素分辨率使用 5 层和 7 层网络[62].

2.6 训练策略

在训练时, 从 p_{ξ} 中随机采样得到相机姿态 $\boldsymbol{\xi}$; 从分布 $\mathcal{N}(0,\boldsymbol{I})$ 和 p_T 中随机采样获得各对象的潜在编码 $\boldsymbol{z}_{\mathrm{s}}^i,\boldsymbol{z}_{\mathrm{a}}^i$ 和边界框 $\boldsymbol{T}_i$; 每个数据集的姿态分布和边界框分布都是先验已知的均匀分布; 相机位置的设置遵循 GIRAFFE[2]. 真实图像 I 从分布为 p_{D} 的训练集中采样. 本文使用具有 R_1 正则化[63]的非饱和 GAN 损失[1], 公式为

$$L(\theta,\phi)=E_{\boldsymbol{z}_{\mathrm{s}}^i,\boldsymbol{z}_{\mathrm{a}}^i\sim\mathcal{N},\boldsymbol{\xi}\in p_{\boldsymbol{\xi}},T_i\in p_T}\left[f\left(D_\phi\left(G_\theta\left(\left\{\boldsymbol{z}_{\mathrm{s}}^i,\boldsymbol{z}_{\mathrm{a}}^i,\boldsymbol{T}\right\}_i,\xi\right)\right)\right)\right]+$$
$$E_{I\sim p_D}\left[f\left(-D_\phi(I)\right)-\lambda\left\|\nabla D_\phi(I)\right\|^2\right].$$

其中, $f(t)=-\ln(1+\mathrm{e}^{-t})$; $\lambda=10$.

本文在 GAN 框架中训练模型, 其中, 生成器和判别器在零和博弈中竞争. 生成器在试图最小化 L 的同时判别器试图最大化 L.

3 实验结果和分析

3.1 数据集

在公开单对象数据集 Cats[33], Chairs[34], CelebA[35]和 FFHQ[10]上对本文方法进行评估, 验证其有效性. 由于这些数据集复杂度是有限的, 其背景是纯白色的或只占图像的一小部分, 因此在更具挑战性的公开单对象数据集 CompCars[36]和 Churches[37]上, 也对合成结果进行评估, 由于对象(物体)并不总是位于中心位置, 背景也更加复杂多变且占据图像的更大一部分, 因此对这 2 个数据集的合成任务更加困难. 对于 CompCars, 本文通过随机裁剪图像实现图像中物体位置的更多变化. 6 个数据集的详细信息如表 1 所示.

3.2 实验设置和评估指标

本文实验的平台和系统设置如下: Ubuntu20.04 操作系统; Gen Intel(R) Core(TM) i7-13700K CPU; NVIDIA GeForce RTX 3090 GPU, 使用 Pytorch1.13 深度学习框架, Python 版本为 3.8. 分别对 64×64 和 256×256 像素分辨率的图像进行合成, 2 种像素分辨率下设置的数据集参数一致. 所有实验中, 相机被定义在半径为 2.732 的球体上, 分别使用 $t_{\mathrm{n}}=0.5$ 和 $t_{\mathrm{f}}=6$ 作为近边界和远边界; 局部表示的窗口大小 p 设定为 4×4, 并沿每条光线采样 $N_{\mathrm{s}}=16$ 个点. 优化器采用 $\beta_1=0$, $\beta_2=0.99$ 的 Adam, batch size=32. 对于 64×64 像素分辨率的实验, 生成器和判别器的学习率分别设置为 6×10^{-4} 和 1×10^{-4}; 对于 256×256 像素分辨率的实验, 修改 $M_f=256$, 生成器学习率为 3×10^{-4}. 为了确定停止训练的时间, 本文遵循常规操作, 每执行 10 000 次迭代进行一次 FID(Fréchet inception distance)度量评估.

表 1　6 个数据集的相关参数

数据集	名称	数据集类型	数量	角度范围/(°)			平移		对象缩放	视场角/(°)
				对象旋转	背景旋转	仰角	横向	纵向		
Cats[33]	猫	真实	9 407	70	10	10			1	10
Chairs[34]	椅子	合成	152 680	360	0	90			1	49
CelebA[35]	人脸	真实	202 599	70	10	10			1	10
Churches[37]	教堂	真实	126 227	360	0	0	−0.15~0.15	−0.15~0.15	0.8~1.0	30
CompCars[36]	车	真实	136 726	360	10	10	−0.12~0.12	−0.22~0.22	0.8~1.0	10
FFHQ[10]	高清人脸	真实	70 000	70	10	10			1	10

本文采用 FID[64]作为度量指标，以此量化合成图像的质量. 使用 20000 个真实和生成样本计算 FID 得分, FID 值越小，合成效果越好.

3.3 实 验

3.3.1 生成结果

针对不同数据集，本文方法在 2 种像素分辨率下通过随机采样获得的部分合成结果分别如图 4 和图 5 所示. 可以看出，本文方法是有效的，其能够从任意的相机姿势进行渲染，合成高质量图像.

为了进一步验证本文方法的有效性，与不同基线方法评估的 FID 基准分数和可视化结果对比分别如表 2 和表 3 所示. 对比方法包括: GIRAFFE[2]方法、



a. Cats[33]



b. Chairs[34]



c. CelebA[35]



d. CompCars[36]



e. Churches[37]

图 4 在 5 个数据集上本文方法通过随机采样生成 64×64 像素分辨率的图像结果



a. FFHQ[10]



b. CompCars[36]



c. Churches[37]

图 5　在 3 个数据集上本文方法通过随机采样生成 256×256 像素分辨率的图像结果

表 2　7 种方法在 64×64 像素分辨率下的 FID 得分对比

方法	数据集				
	Cats	Chairs	CelebA	Churches	CompCars
2D-GAN[63]	18.00	59.00	15.00	19.00	16.00
BlockGAN[8]	47.00	41.00	69.00	28.00	41.00
HoloGAN[7]	27.00	59.00	25.00	31.00	17.00
GRAF[3]	26.00	34.00	25.00	38.00	39.00
GIRAFFE[2]	8.00	27.00	**6.00**	17.00	16.00
FMC[4]				18.75	15.82
本文	**7.08**	**20.34**	**6.09**	**15.07**	**11.13**

注. 粗体表示最优值.

表 3　5 种方法在 256×256 像素分辨率下的 FID 得分对比

方法	数据集		
	FFHQ	Churches	CompCars
HoloGAN[7]	192.00	58.00	34.00
GRAF[3]	59.00	87.00	95.00
GIRAFFE[2]	32.00	30.00	26.00
FMC[4]	29.30	27.01	22.08
本文	**26.23**	**23.45**	**19.57**

注. 粗体表示最优值.

基于辐射场的 GRAF[3]、特征掩码耦合(feature mask coupling, FMC)方法[4]、基于体素的 HoloGAN[7]和 BlockGAN[8]方法，以及基于 ResNet[65]的 2D-GAN 方法[63].

从表 2 可以看出，在 64×64 像素分辨率下，本文方法在所有数据集上获得的 FID 得分显著低于 HoloGAN[7]和 BlockGAN[8]方法，表明与基于体素化 3D 表示的方法相比，本文采用基于神经隐式表示的方法能更有效地捕捉和表达图像中的复杂结构和细节；与同样采用基于神经隐式表示的 GRAF[3]，GIRAFFE[2]和 FMC[4]方法相比，本文方法获得了相似或更好的 FID 分数，进一步证明，采用自注意力网络结构改良的隐式 GAN 生成器在表达能力上优于传统的 MLP 架构. 类似地，从表 3 可以看出，在 256×256 像素分辨率下，本文方法的有效性得到了验证.

在 CompCars 数据集上，在 2 种像素分辨率下，将 5 种方法在复杂场景进行可控合成实验，结果如表 4 所示. 其中，旋转对象 360°，FMC[4]方法中只

提供了 64×64 像素分辨率结果. 可以看出: (1) 虽然所有方法都允许在有限复杂性的数据集上进行可控图像合成, 但在具有杂乱背景的复杂场景上, HoloGAN[7]和 GRAF[3]方法的合成结果是不一致的, 其前景对象和背景之间存在相互纠缠; 在不同视角下, 前景对象要么几何上存在明显伪影扭曲, 要么外观上各不相同. (2) GIRAFFE[2], FMC[4]和本文方法由于将场景解耦表示, 因此能够在固定背景的情况下旋转对象. (3) 与 GIRAFFE[2]和 FMC[4]方法相比, 本文方法在不同视角下能够更好地保证合成对象的一致性, 在 64×64 像素分辨率下, 随着角度的变化, GIRAFFE[2]和 FMC[4]方法合成对象的几何或外观方面发生了明显变化, 注意车辆形状和车身颜色在旋转角度时的差异; 在 256×256 像素分辨率下, GIRAFFE[2]方法合成对象在几何上同样存在差异, 具体体现在旋转过程中车辆几何形变和车标特征缺失. (4) 无论在几何还是外观方面, 本文方法合成图像均获得了更好的一致性, 进一步表明, 本文采用基于自注意力网络的神经隐式表示, 更能保证合成结果的一致性.

表 4 在 CompCars 数据集上 5 种方法可控合成结果对比

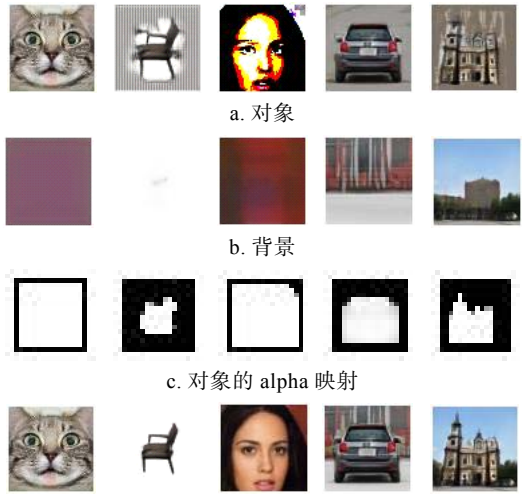
方法	像素分辨率	可控合成结果
HoloGAN[7]	64×64	
	256×256	
GRAF[3]	64×64	
	256×256	
GIRAFFE[2]	64×64	
	256×256	
FMC[4]	64×64	
本文	64×64	
	256×256	

定量和定性结果表明, 本文方法在多个方面取得了显著的改进. 定量分析展示了在 FID 得分上的明显降低, 进一步证明了本文方法在图像合成任务中的优越性; 通过图像定性分析可以看出, 本文方法生成的图像在细节、多样性和场景一致性等方面都表现出卓越的性能, 充分地验证了本文方法的有效性和实用性.

3.3.2 可控场景生成

为了验证对象和背景是否被正确地解耦, 利用组合算子这一简单加法操作(式(1)), 在 64×64 像素分辨率合成任务中, 分别呈现各场景下的解耦对象、解耦背景和对象的 alpha 映射(式(2))结果, 如图 6 所示. 可以看出, 本文方法在没有任何监督的情况下, 成功地实现了对象和背景解耦表示; 对于

Cats 和 CelebA 这些训练数据中几乎仅包含带有对象的图像(即仅包含猫或人脸)时, 学习到的模型也能够生成一个合理可信的背景, 从而隐式地发挥修补效果.



a. 对象

b. 背景

c. 对象的 alpha 映射

d. 最终合成结果

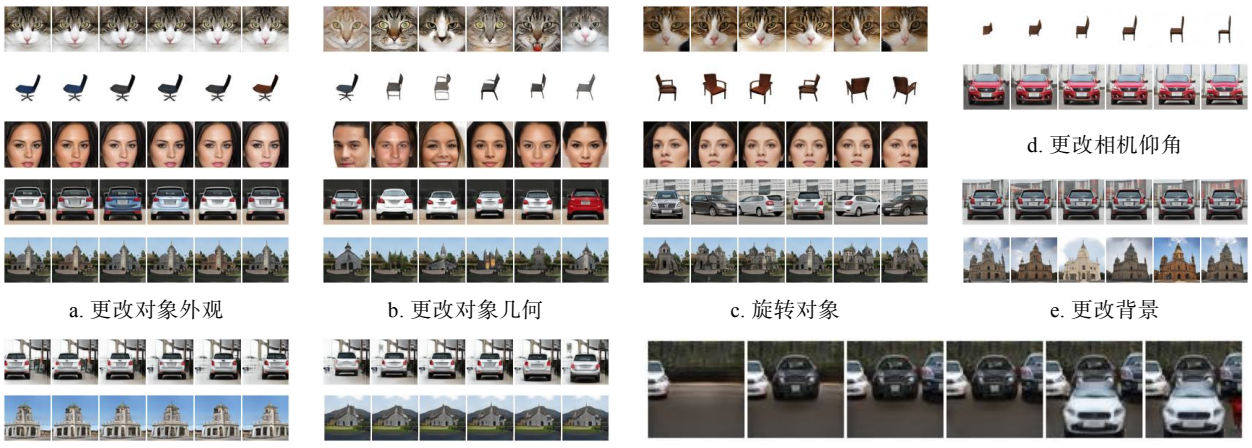
图 6　本文方法在 64×64 像素分辨率上的解耦结果

进一步, 对将组合的 3D 表示合并到生成模型中带来多大程度的可控性进行验证, 从各数据集上多方面地对对象的属性进行控制分析. 在各场景合成任务中, 通过采集输入不同的潜在编码, 模型能够为每个实体对象建模不同的几何和外观, 同时在控制对象几何不变的情况下改变对象的外观. 根据表 1 中的参数设定, 针对不同的数据集在图像合成过程中, 本文实现了对场景进行不同程度的旋转、平移、缩放, 以及改变相机仰角等操作, 完成高质量的可控图像合成. 本文方法在 Cats, Chairs, CelebA, CompCars 和 Churches 数据集上, 对 64×64 像素分辨率的图像进行可控合成的可视化, 结果如图 7 所示. 可以看出: (1) 图 7a 中, 在保持背景不变的情况下, 通过为前景对象采集相同的几何潜在编码 z_s 和不同的外观潜在编码 z_a, 本文方法实现了对场景中的对象进行几何固定的同时, 更改其外观; 同理, 图 7b 中, 实现了同时更改场景中对象的几何和外观的操作; 图 7c 和图 7d 中, 通过在对象旋转范围或相机仰角范围之内进行不同角度的合成, 实现了对场景中的对象进行旋转和更改仰角操作. (2) 图 7f~图 7h 中, 对于具有杂乱背景的高复杂度 CompCars 和 Churches 数据集, 进一步实现了在 3D 空间中对感兴趣的单个实体对象进行横向和纵向平移, 以及更换场景背景操作. 这充分地表明了本文方法具有出色的可控性能. (3) 图 7e 中, 当控制实体对象不变而对背景进行更改时, 场景中相应的实体对象会自适应地随背景亮度变化而产生变化, 进一步表明本文方法在处理复杂场景下具有良好的鲁棒性和逼真性. (4) 图 7h 中, 在 CompCars 数据集上探索了多个对象合成性能, 将组合的 3D 表示合并至生成模型, 促进了所提模型能够在场景中添加多个分布之外的对象. 类似地, 在 FFHQ, CompCars 和 Churches 数据集上, 对 256×256 像素分辨率的图像进行可控合成的可视化结果如图 8 所示.

综上所述, 本文方法除了能够实现多视角一致的逼真图像合成之外, 还支持对场景中感兴趣的对象进行多方面的控制, 进一步促进生成的图像更具多样性和可定制性.

3.3.3　消融实验

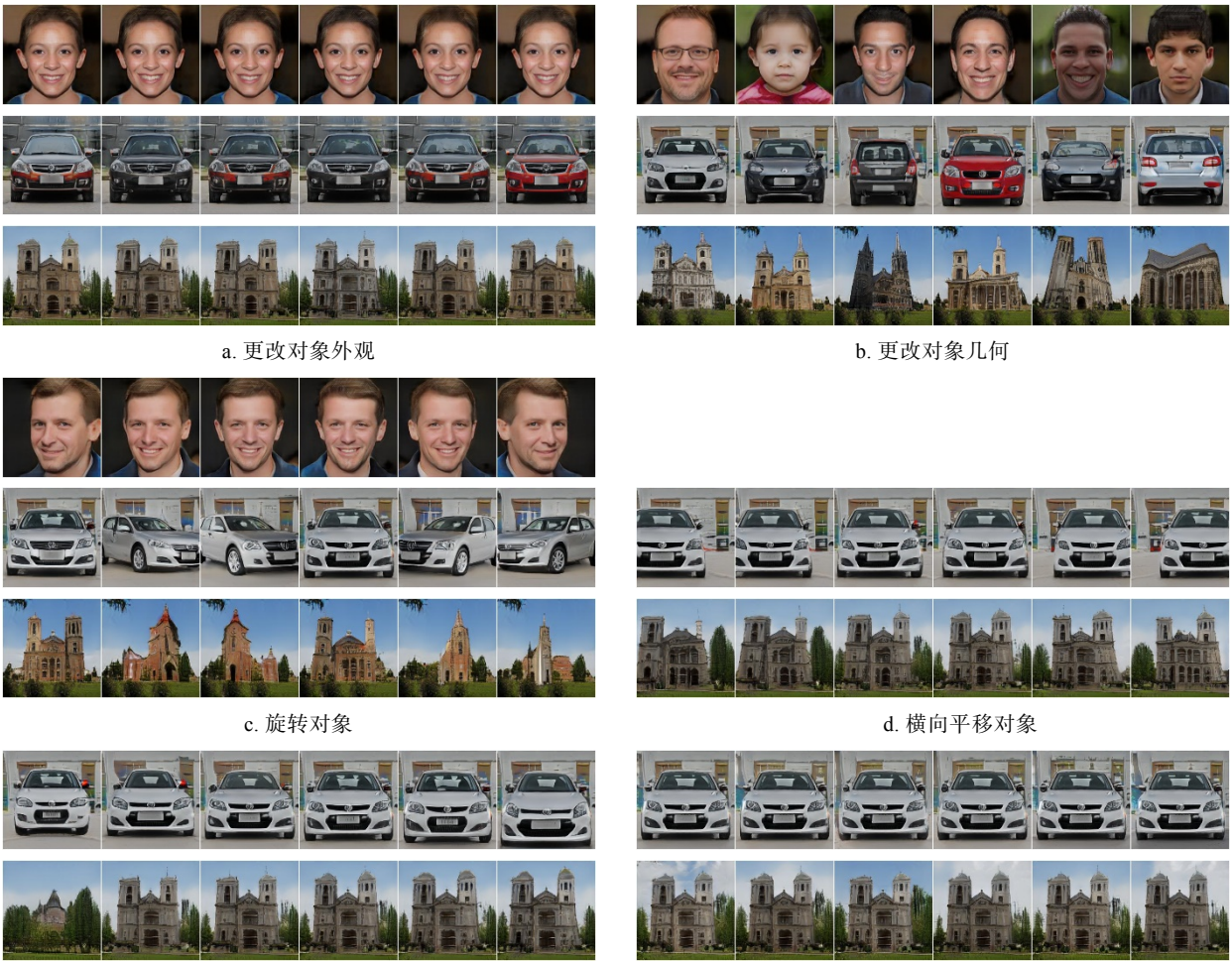
在 64×64 像素分辨率下, 分别在复杂度各不



a. 更改对象外观　b. 更改对象几何　c. 旋转对象　d. 更改相机仰角　e. 更改背景

f. 横向平移对象　g. 纵向平移对象　h. 添加多个对象

图 7　在 64×64 像素分辨率上对对象进行可控操作的结果



a. 更改对象外观　b. 更改对象几何　c. 旋转对象　d. 横向平移对象　e. 纵向平移对象　f. 更改背景

图 8　在 256×256 像素分辨率上对对象进行可控操作的结果

相同的 Cats, Chairs 和 Churches 数据集上，对背景特征场与对象特征场的结构完全相同或简化情况下的性能差异进行实验，结果如表 5 所示. 可以看出，在背景特征场与对象特征场网络结构完全相同的情况下，与简化结构相比，Cats 和 Chairs 数据集上的 FID 指标分别升高 1.19 和 0.73，表明对于背景仅占图像小部分区域或背景纯白的合成场景来说，复杂的表征网络并没有促进合成效果的提升，反而会因为对简单特征的过度拟合造成合成性能的降低；进一步，对于背景占据图像更大一部分的 Churches 数据集，这种过度拟合带来的合成性能影响则显得更加突出. 实验结果表明，简化背景特征场结构不仅有助于减少计算资源浪费，还能适应不同场景下的背景复杂性，保持或提升合成质量.

在 64×64 像素分辨率、其他配置不变的情况下，通过逐一删除各个模块的方式进行消融实验，进一步分析本文架构中的 RGB 跳跃连接和渲染模块对合成性能的改善是积极的. 以 CompCars 和 Churches 复杂数据集为例，实验结果如表 6 所示. 可以看出，去除 RGB 跳跃连接模块时，2 个数据集上的 FID 指标较完整架构分别升高 2.88 和 4.29，表明 RGB 跳跃连接在将低级特征信息传递到网络更深层时，能够有效地帮助模型保留细节和结构信息，提高生成图像的质量，也验证了 RGB 跳跃连接在本文架构中的有效性和可行性.

表 5　在 3 个数据集上 2 种结构的 FID 对比

数据集	相同结构	简化结构
Cats[33]	8.27	**7.08**
Chairs[34]	21.07	**20.34**
Churches[37]	22.01	**15.07**

注. 粗体表示最优值.

在渲染模块的设计中，本文方法与 GRAF[3]的关键区别是将体渲染与 2D 神经渲染相结合，这与 GIRAFFE[2]和 FMC[4]的设计思想是相似的. 表 2 和表 3 的定量比较，以及表 4 的定性比较已经表明，

表 6　在 2 个数据集上消融实验之后的 FID 对比

数据集	完整架构	−RGB 跳跃连接	−pixelshuffle
CompCars[36]	11.13	14.01	13.93
Churches[37]	15.07	19.36	18.74

注. “−”表示删除.

这种设计可以获得更好的合成结果, 尤其是对于复杂的真实数据, 模型的表现力更突出; 此外, 这种设计使得渲染速度得到提升. 与 GRAF[3]相比, 在 64×64 像素分辨率下, 渲染一幅图像的总渲染时长从 33.5 ms 减少到 2.6 ms; 在 256×256 像素分辨率下, 渲染一幅图像的总渲染时长从 552.8 ms 减少到 3.5 ms. 在神经渲染器的构造上, 本文方法不再像 GIRAFFE[2]和 FMC[4]那样仅考虑空间层面的因素影响, 即仅使用最近邻或双线性上采样, 而是引入了通道上采样 pixelshuffle 技术. 通过分析表 5 可知, 当去除 pixelshuffle 上采样技术时, 在 2 个数据集上的合成性能降低, FID 指标分别升高 2.8 和 3.67, 进一步证明了本文方法将通道上采样技术与空间上采样技术相结合的有效性, 成功地改善了合成性能.

3.4　局限性

如果数据中存在固有偏差, 本文方法就难以识别变异因素. 如图 9 所示, 在 FFHQ 场景合成任务中, 当对象进行旋转操作时, 生成图像中的眼睛和头发区域未能保持其原始的相对位置, 而是随着视角的改变而发生相应变化. 这一现象的原因是 FFHQ 数据集中的眼睛和头发通常朝向相机, 而不是与面部的旋转角度保持一致. 因此, 在图像合成过程中, 这些特征会根据数据集的固有分布模式进行自适应的调整, 这可能限制了模型对于视角变化的泛化能力.

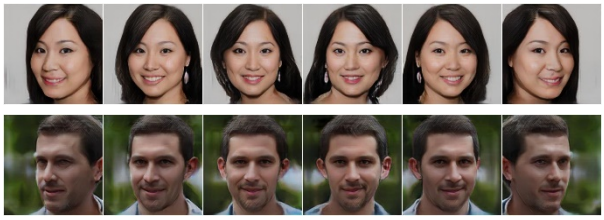


图 9　数据集偏差导致眼睛和头发随视角变化

在处理复杂度更高的 CompCars 和 Churches 场景合成任务时, 本文方法偶尔会出现解耦失败的情况. 如图 10 所示, 当解耦前景对象时, 生成的表示有时会包含背景元素; 反之, 解耦背景时会包含前景对象(车或教堂). 这种解耦失败可能源于几个因素: (1) 场景本身的高复杂性增加了模型处理的难度; (2) 模型在训练过程中对相机姿势和对象转换的假设是均匀分布的, 这与实际场景中这些因素的分布可能存在显著差异, 这种假设与实际的不匹配, 可能导致模型在实际应用中的性能下降.



a. 解耦前景对象时出现背景元素



b. 解耦背景时出现前景对象

图 10　解耦失败案例——前景对象与背景之间相互纠缠

4　结　语

针对现有方法缺乏底层的 3D 表示和依赖表达能力不足的网络架构, 导致生成图像保真度低、不可控和一致性差等问题, 本文提出一种基于自注意力机制的局部空间 3D 感知图像合成方法. 对象解耦模块通过为各个对象定义单独 3D 边界框, 有效地解决对象间的纠缠, 保证了对场景中感兴趣的单个对象进行可控操作; 局部空间场景表征模块使用自注意力网络进行局部空间表征, 促进局部空间细节的表达, 大大降低了计算代价; 渲染模块中, 神经渲染器对每个空间分辨率下的特征图采用空间上采样技术和通道上采样技术并行处理并融合, 逐渐升高分辨率. 实验结果表明, 本文方法可以在对空间点低频采样的情况下, 实现多视角一致的高质量图像合成; 进一步, 在复杂度更高的真实场景数据集上, 该方法的表达能力更突出.

未来, 将从以下 3 个方面改进本文工作: (1) 探索一种在训练过程中逐渐改变局部空间表征大小的策略, 进一步提高合成质量; (2) 尝试从数据中学习相机姿态的分布, 有益于进一步对合成场景进行精确控制; (3) 结合易于获得的监督(如预测的对象掩码信息), 进一步扩展到更复杂的多对象场景.

参考文献(References):

[1] Chan E R, Monteiro M, Kellnhofer P, *et al*. pi-GAN: periodic

implicit generative adversarial networks for 3D-aware image synthesis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 5795-5805

[2] Niemeyer M, Geiger A. GIRAFFE: representing scenes as compositional generative neural feature fields[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 11448-11459

[3] Schwarz K, Liao Y Y, Niemeyer M, *et al*. GRAF: generative radiance fields for 3D-aware image synthesis[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.1692

[4] Gupta S, Keshari A, Das S. Controllable image synthesis via feature mask coupling using implicit neural representation[C] //Proceedings of the 13th Indian Conference on Computer Vision, Graphics and Image Processing. New York: ACM Press, 2023: Article No.45

[5] Goodfellow I J, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680

[6] Henzler P, Mitra N, Ritschel T. Escaping plato's cave: 3D shape from adversarial rendering[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9983-9992

[7] Nguyen-Phuoc T, Li C, Theis L, *et al*. HoloGAN: unsupervised learning of 3D representations from natural images[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7587-7596

[8] Nguyen-Phuoc T, Richardt C, Mai L, *et al*. BlockGAN: learning 3D object-aware scene representations from unlabelled images[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.568

[9] Karras T, Aila T, Laine S, *et al*. Progressive growing of GANs for improved quality, stability, and variation[C] //Proceedings of the 6th International Conference on Learning Representations. Red Hook: Curran Associates Inc, 2018: 1-26

[10] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4217-4228

[11] Karras T, Laine S, Aittala M, *et al*. Analyzing and improving the image quality of StyleGAN[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8107-8116

[12] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[C] //Proceedings of the 7th International Conference on Learning Representations. Red Hook: Curran Associates Inc, 2019: 1-35

[13] Choi Y, Uh Y, Yoo J, *et al*. StarGAN v2: diverse image synthesis for multiple domains[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8185-8194

[14] Zhu Yan, Wang Kai, Wang Linbo, *et al*. Pose and shape estimation of occluded humans with attention and conditional GAN[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(1): 142-151(in Chinese)
(朱妍, 汪楷, 汪鄯波, 等. 联合注意力和条件 GAN 的被遮挡人体姿态和体形估计方法[J]. 计算机辅助设计与图形学学报, 2024, 36(1): 142-151)

[15] Mildenhall B, Srinivasan P P, Tancik M, *et al*. NeRF: representing scenes as neural radiance fields for view synthesis[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 405-421

[16] Xie Zhifeng, Zheng Jiaheng, Wang Ji, *et al*. Speech-driven facial reenactment based on implicit neural representations with structured latent codes[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(10): 1616-1624(in Chinese)
(谢志峰, 郑迦恒, 王吉, 等. 基于结构化潜码引导 NeRF 的语音驱动人脸重演[J]. 计算机辅助设计与图形学学报, 2024, 36(10): 1616-1624)

[17] Drebin R A, Carpenter L, Hanrahan P. Volume rendering[J]. ACM SIGGRAPH Computer Graphics, 1988, 22(4): 65-74

[18] Kajiya J T, Von Herzen B P. Ray tracing volume densities[J]. ACM SIGGRAPH Computer Graphics, 1984, 18(3): 165-174

[19] Locatello F, Bauer S, Lucic M, *et al*. Challenging common assumptions in the unsupervised learning of disentangled representations[C] //Proceedings of the 36th International Conference on Machine Learning. Red Hook: Curran Associates Inc, 2019: 4114-4124

[20] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[21] Liao Y Y, Schwarz K, Mescheder L, *et al*. Towards unsupervised learning of generative models for 3D controllable image synthesis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 5870-5879

[22] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6000-6010

[23] Zhang Y L, Li K P, Li K, *et al*. Image super-resolution using very deep residual channel attention networks[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 294-310

[24] Zhang H, Goodfellow I J, Metaxas D N, *et al*. Self-attention generative adversarial networks[C] //Proceedings of the 36th International Conference on Machine Learning. Red Hook: Curran Associates Inc, 2019: 7354-7363

[25] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[26] Zheng S X, Lu J C, Zhao H S, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6877-6886

[27] Beal J, Kim E, Tzeng E, *et al*. Toward transformer-based object

detection[OL]. [2024-02-02]. https://arxiv.org/abs/2012.09958

[28] Han K, Xiao A, Wu E, *et al*. Transformer in transformer[C] //Proceedings of the Neural Information Processing Systems. Curran Associates Inc, 2021: 15908-15919

[29] Srinivas A, Lin T Y, Parmar N, *et al*. Bottleneck transformers for visual recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 16514-16524

[30] Jiang Y F, Chang S Y, Wang Z Y. TransGAN: two pure transformers can make one strong GAN, and that can scale up[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.1130

[31] Parmar N, Vaswani A, Uszkoreit J, *et al*. Image transformer[C] //Proceedings of the 35th International Conference on Machine Learning. Red Hook: Curran Associates Inc, 2018: 4055-4064

[32] Chen Y L, Wang Z C, Peng Y X, *et al*. Cascaded pyramid network for multi-person pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7103-7112

[33] Zhang L, Curless, Hertzmann, *et al*. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo[C] //Proceedings of the 9th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2003: 618-625

[34] Oechsle M, Niemeyer M, Reiser C, *et al*. Learning implicit surface light fields[C] //Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2020: 452-462

[35] Liu Z W, Luo P, Wang X G, *et al*. Deep learning face attributes in the wild[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 3730-3738

[36] Yang J L, Li H D. Dense, accurate optical flow estimation with piecewise parametric model[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1019-1027

[37] Yu F, Seff A, Zhang Y D, *et al*. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop[OL]. [2024-02-02]. https://arxiv.org/abs/1506.03365

[38] Mescheder L, Oechsle M, Niemeyer M, *et al*. Occupancy networks: learning 3D reconstruction in function space[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4455-4465

[39] Park J J, Florence P, Straub J, *et al*. DeepSDF: learning continuous signed distance functions for shape representation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 165-174

[40] Bagautdinov T, Wu C L, Saragih J, *et al*. Modeling facial geometry using compositional VAEs[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3877-3886

[41] Fan H Q, Su H, Guibas L. A point set generation network for 3D object reconstruction from a single image[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2463-2471

[42] Yang Y Q, Feng C, Shen Y R, *et al*. FoldingNet: point cloud auto-encoder via deep grid deformation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 206-215

[43] Girdhar R, Fouhey D F, Rodriguez M, *et al*. Learning a predictable and generative vector representation for objects[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 484-499

[44] Tatarchenko M, Dosovitskiy A, Brox T. Octree generating networks: efficient convolutional architectures for high-resolution 3D outputs[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2107-2115

[45] Wang Q Q, Wang Z C, Genova K, *et al*. IBRNet: learning multi-view image-based rendering[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4688-4697

[46] Reizenstein J, Shapovalov R, Henzler P, *et al*. Common objects in 3D: large-scale learning and evaluation of real-life 3D category reconstruction[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 10881-10891

[47] T M V, Wang P H, Chen X X, *et al*. Is attention all that NeRF needs?[C] //Proceedings of the 11th International Conference on Learning Representations. Red Hook: Curran Associates Inc, 2023: 1-22

[48] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, *et al*. Deep unsupervised learning using nonequilibrium thermodynamics[C] //Proceedings of the 32nd International Conference on International Conference on Machine Learning. Red Hook: Curran Associates Inc, 2015: 2256-2265

[49] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.574

[50] Song Y, Sohl-Dickstein J, Kingma D P, *et al*. Score-based generative modeling through stochastic differential equations[C] // Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: 37799-37812

[51] Nichol A Q, Dhariwal P, Ramesh A, *et al*. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models[C] //Proceedings of the 39th International Conference on Machine Learning. New York: PMLR, 2022: 16784-16804

[52] Berrada T, Verbeek J, Couprie C, *et al*. Unlocking pre-trained image backbones for semantic image synthesis[OL]. [2024-02-02]. https://arxiv.org/abs/2312.13314

[53] Pan X G, Tewari A, Leimkuhler T, *et al*. Drag your GAN: interactive point-based manipulation on the generative image manifold[C] //Proceedings of the ACM SIGGRAPH 2023

Conference Proceedings. New York: ACM Press, 2023: Article No.78

[54] Xiao Z S, Kreis K, Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs[C] //Proceedings of the 10th International Conference on Learning Representations. Red Hook: Curran Associates Inc, 2022: 1-28

[55] Chen Z Q, Kim V G, Fisher M, *et al*. DECOR-GAN: 3D shape detailization by conditional refinement[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 15735-15744

[56] Wu J J, Zhang C K, Xue T F, *et al*. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2016: 82-90

[57] Zhu J Y, Zhang Z T, Zhang C K, *et al*. Visual object networks: image generation with disentangled 3D representations[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2018: 118-129

[58] Gadelha M, Maji S, Wang R. 3D shape induction from 2D views of multiple objects[C] //Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2017: 402-411

[59] Szabó A, Meishvili G, Favaro P. Unsupervised generative 3D shape learning from natural images[OL]. [2024-02-02]. https://arxiv.org/abs/1910.00287

[60] Tancik M, Srinivasan P P, Mildenhall B, *et al*. Fourier features let networks learn high frequency functions in low dimensional domains[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.632

[61] Shi W Z, Caballero J, Huszár F, *et al*. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1874-1883

[62] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[OL]. [2024-02-02]. https://arxiv.org/abs/1511.06434

[63] Mescheder L M, Geiger A, Nowozin S. Which training methods for GANs do actually converge?[C] //Proceedings of the 35th International Conference on Machine Learning. Red Hook: Curran Associates Inc, 2018: 3481-3490

[64] Heusel M, Ramsauer H, Unterthiner T, *et al*. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6629-6640

[65] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778