GIRAFFE: 将场景表示为叠加生成式的神经特征场

Michael Niemeyer[1,2] Andreas Geiger[1,2]
[1] Max Planck Institute for Intelligent Systems, Tubingen [2]University of Tubingen

摘要

深度的生成模型可以生成高分辨率的真实图像。但在很多应用场景中，这还远远不够：我们希望能够控制生成的图像内容。虽然近来有几个研究工作想要消除由于数据变化所带来的潜在影响，但是他们中的大部分都只在二维空间来进行实验，却忽略了现实世界是三维的。此外，只有少数的研究考虑到了场景的可叠加性。我们的关键假设是将 3D 构图场景引入到生成模型中，来合成更可控的图像。把场景表示成生成式的组合神经特质场能够让我们将一个或多个物体从背景中结耦出来。同时，我们的模型还可以将不同形状与外观的物体互相结耦。这些都可以仅在原图像集上完成（相机拍摄位置随意的，场景物体也随意的那种），不需要任何额外的监督信号。将场景表示与神经渲染管道相结合能够得到一种快速生成真实图像的生成模型。我们的实验表明，我们的模型能够单独的物品，并且能够在场景中对让它们进行旋转变换，和视角的变换 。

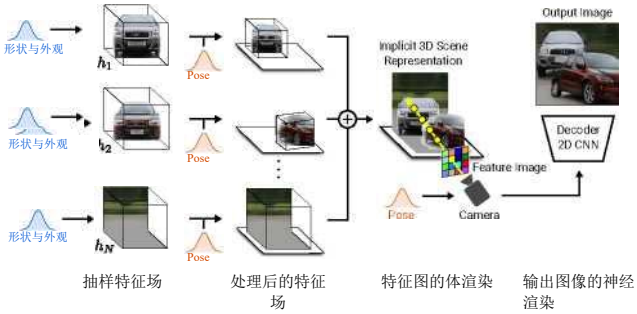


图 1: **概述.** 我们将场景表示成叠加生成的神经特征场，对于随机抽样得到的一张图像，我们基于场景本身的特征场来渲染出一张特征图像。一个二维的神经渲染网络能够将特征图转换成 RGB 图像。虽然我们仅在原图像集上进行训练，但我们能够在不需要考虑相机和物体的移动、物体的形状与外观下，控制图像的生成过程。此外，我们的模型可以生成训练集里没有的数据，例如我们能够生成拥有比训练图像上更多物体的场景。注意，为了能够更清晰地观察到结果，我们将颜色可视化，而不是特征可视化。

1. 介绍

计算机视觉和图形学的一个长期目标是使计算机拥有生成并且操纵图像的能力。现在计算机图像技术已经在游戏与电影场景制作方面达到了很好的效果。但是，这些依赖于昂贵的相机设备和大量的人力来进行场景内容的创建和排布。

近几年，计算机视觉领域已经在生成真实图像中取得了巨大的进展，特别是在生成模型 GANs[24] 的出现之后。GANs 能够合成 1024*1024 分辨率或者更高分辨率的真实图像。 [6,14,15,39,40].

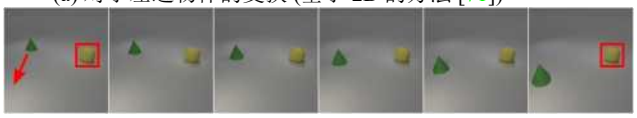
尽管 GANs 取得了一些成功，合成真实的 2 维图像并不是生成模型所应用的唯一方向。生成图像的过程需要用一个简单有效的方式来控制。为此，有许多工作[9,25,39,43,44,48,54,71,74,97,98]研究了在没有监督信号的条件下通过数据来学到结耦特征。特征结耦的定义各有不同[5,53]，但是基本上都是指能够在不改变其他特征属性的条件下控制一个特征的变化（例如：物体的形状，大小，姿态等等）。然而，目前的大部分方法都没有考虑到场景的可叠加性，并且只在二维图像上进行实验，忽略我们的现实世界是三维的。这经常造成场景表示的混乱(图. 2) 。控制场景生成的机制也不是我们设定的，而是在潜在空间上发现的。但往往能够进行场景的表示和控制生成的场景内容对于应用程序来说是至关重要的。例如：电影制作中需要合成复杂物体的运动轨迹。

因此，近来有数项工作研究如何合成三维图像：体素[32,63,64]、原语[46]、辐射场 [77]，并直接将它们用到生成模型上。

虽然这些方法可以合成可控的图像并取得了很好的效果，但它们大多都局限在单对象的场景里，对于更高分辨率和更真实复杂的场景中（例如物体不在图像中心或是背景杂乱的场景），结果往往差异很大。

贡献: 在本文中，我们介绍了 *GIRAFFE*，一种在原始图像

(a) 对于左边物体的变换 (基于 2D 的方法 [71])



(b) 对于左边物体的变化 (我们的模型)

(c) 旋转变换 (我们的模型)　　(d) 增加物体数 (我们的模型)

图 2: **可控图像生成。**因为大多数生成模型只在二维空间上进行操作，我们在那些模型上加入了可叠加的三维场景表示。这会使得生成的图像趋于一致。但要注意，在二维图像上复现我们的方法时，改变一个物体可能会导致其他物体也改变。改进后的模型还能实现像旋转变换或者是测试时在场景中增加更多物体等复杂操作。这些方法都是在双对象的原始图像集上进行训练的。

集上训练后就能够合成可控的真实图像的新方法。主要有两个关键点: 第一，直接将叠加后的的 3 维场景图与生成模型相结合，能够得到更为可控的合成图像。第二：将场景表示和神经渲染管线相结合，能够更快地生成更真实的图像。为此，我们将场景表示成组合的神经特质场(图. 1)。为了减少时间和计算量，我们在低分辨率的特征图像上进行体积渲染。神经渲染器处理特征图像并将最后的渲染图输出。这样，我们就可以得到高质量的图像，并把它们与现实世界的场景进行对应。我们发现我们的模型虽然仅在原训练集上进行训练，但在多物体场景中生成可控场景的效果与在单物体场景上的一样好。代码与数据集请参阅 https://github.com/autonomousvision/giraffe.

2. 相关研究

基于 GAN 的图像合成: 生成对抗网络（GANs）[24] 已经被证实可以生成 1024*1024 甚至更高分辨率的真实图像 [6,14,15, 39,40]。为了能够更好地控制图像合成过程，许多工作研究了如何在无监督信号的条件下消除变量的影响。他们要不修改网络训练目的[9,40,71]，要不修改网络结构[39]，或者研究预先训练好的生成模型的隐空间[1,16, 23,27,34,78,96]。然而，这些方法全都不能表现出场景的叠加特性。因此，近期的工作都在主要研究怎样才能可控地合成物体 [3,4,7,18,19,26,45,86,90]。此外，上述模型虽然都能合成二维的真实图像，但是却没考虑到现实世界是三维的。在本文中，为了更好地结耦和更可控地合成图像，我们倡导大家模拟 3D 建模过程来实现这个操作。

隐函数：在基于学习的三维重构中，非常流行使用隐函数来表示三维物体的几何形状[11, 12, 22, 59, 60, 65, 67, 69, 76]，并将此方法运用到了场景重建中[8, 13, 35,72, 79]。为了克服 3D 监督的需要，很多团队 [50, 51, 66, 81, 92] 提出不同的渲染方法。Mildenhall 等人[61] 提出了神经辐射场（NeRFs）的概念，将隐式的神经模型和用于合成复杂场景的体积渲染相结合。由于 NeRFs 的可表示性，我们使用它的变体来表达模型。与我们的方法相比，NeRFs 需要用不同姿态的相机所拍摄的多视角图像作为监督，每个场景都要用一个网络来训练，并且不能生成新的场景（即未在训练集上出现的场景）。而我们能够从原图像集上学到一个生成模型来合成可控的、逼真的场景图像。

3D 感知图像合成:一些工作研究了如何将三维表示作为一个误差量加入到生成模型中[21,29-32,46,55,63,64,75,77]。有很多方法需要用到额外的监督信号[2,10, 87,88,99]，但我们的方法仅需在原图像集上进行操作即可。

Henzler 等人[32] 用可微分渲染学到基于体素的表达，这个方法使得生成图像在三维空间上是可控的，但是随着立方体存储所需空间的增长，体素的分辨率受到了限制，这会导致图像出现伪影。Nguyen-Phuoc 等人 [63,64] 提出了体素网格，能够将图像重建后呈现在 2 维上。虽然取得的结果很好，但是在数据集上的训练变得不太稳定，并且在高维空间上的结果缺乏一致性。Liao 等人 [46] 用抽象特征将原语和可微分渲染结合在一起，虽然这个方法可以用于多物体场景，但是它要求图片是纯背景形式的，这个要求在真实世界的场景中是很难达到的。Schwarz 等人[77] 提出生成的神经辐射场（GRAF），虽然能够合成高分辨率的可控图像，但是此方法仅能用在单物体场景，在更复杂的真实世界场景中效果不好。相比于上述方法，我们将三维场景的组成结构与生成模型相结合，以便更好地处理多物体场景问题。此外，将此方法与神经渲染管线相结合 [20,41,42,49,62,80,81,83,84]，我们的模型能够适应更复杂的真实世界数据。

3. 方法

我们的目标是无需额外的监督信号，单在原图像集上训练就能获得一段可控的图像序列（视频）。下面我们将讨论我们方法的主要内容。首先，我们的模型将单个物体建模成神经特质场 (Sec. 3.1)。接下来，我们利用特质场的叠加属性将多个单独对象叠加到同一个场景(Sec. 3.2)。为了达到更好的渲染效果，我们研究了一个能够将体素和神经渲染技术相结合的方法(Sec. 3.3)。最后，我们讨论了如何在原图像集上训练我们的模型(Sec. 3.4)。具体方法概述请看图.3。

3.1. 用神经特质场来表示物体

神经辐射场: 辐射场是一个连续函数f，它将三维点坐标$\mathbf{x} \in \mathbf{R}^3$与视角$d \in \mathbf{S}^2$映射成体积密度与视相关 RGB 颜色值。[61,82]两篇论文中都有一个关键的发现，当f是神经网络的

参数时，低维的输入 x，d 需要被映射成高维的特征，才能表示复杂的信号。更特殊的是，一个预定义的位置编码能够被应用到 x，d 中的任何一个分量中：

$$\gamma(t, L) = (\sin(2^0 t\pi), \cos(2^0 t\pi), \ldots, \sin(2^L t\pi), \cos(2^L t\pi)) \quad (1)$$

其中，t 是一个输入的标量，比如 **x**，**d** 中的任一参数；L 是频率。在生成模型的背景下，我们发现这一种表示方法的另一种好处：它引入了一个归纳偏置来学习特定方向而不是任意方向上的三维形状表示（具体看图.11）

根据图形学中的隐式表示 [12，59，69]，Mildenhall 等人[61]提出用多层感知机（MLP）参数化 f 来学习神经辐射场（NeRFs）：

$$f_\theta : \mathbb{R}^{L_\mathbf{x}} \times \mathbb{R}^{L_\mathbf{d}} \to \mathbb{R}^+ \times \mathbb{R}^3 \quad (\gamma(\mathbf{x}), \gamma(\mathbf{d})) \mapsto (\sigma, \mathbf{c}) \quad (2)$$

其中θ是网络参数，Lx 和 Ld 是 x，d 位置编码的输出维数。

生成神经特质场: 因为参数θ [61]只能够用于单场景的多视角图像，Schwarz 等人[77] 提出用神经辐射场来训练原图像集的生成模型（GRAF）。为了学习到 NeRFs 的隐空间，他们将形状Z_s与外观Z_a作为多层感知机的条件，Z_s和Z_a均服从均值为 0，方差为 I 的高斯分布。

$$g_\theta : \mathbb{R}^{L_\mathbf{x}} \times \mathbb{R}^{L_\mathbf{d}} \times \mathbb{R}^{M_s} \times \mathbb{R}^{M_a} \to \mathbb{R}^+ \times \mathbb{R}^3 \quad (\gamma(\mathbf{x}), \gamma(\mathbf{d}), \mathbf{z}_s, \mathbf{z}_a) \mapsto (\sigma, \mathbf{c}) \quad (3)$$

其中 Ms，Ma 是隐码的维数。

在本次实验中，我们探讨了一种将体素和神经渲染结合起来的更有效的方法。我们将 GRAF 公式中输出的三通道颜色 c 替换成更通用的 M_f 维的特征 f，并用神经特质场来表示物体：

$$h_\theta : \mathbb{R}^{L_\mathbf{x}} \times \mathbb{R}^{L_\mathbf{d}} \times \mathbb{R}^{M_s} \times \mathbb{R}^{M_a} \to \mathbb{R}^+ \times \mathbb{R}^{M_f} \quad (\gamma(\mathbf{x}), \gamma(\mathbf{d}), \mathbf{z}_s, \mathbf{z}_a) \mapsto (\sigma, \mathbf{f}) \quad (4)$$

对象表示: GRAF 与 NeRF 中都有一个关键限制条件：整个场景只能用一个模型来表示。当我们想要分离场景中的不同实体时，我们需要控制住这个物体的摆放位置，形状和外观（我们认为背景也算是一个物体），因此，我们对每一个对象都用一个特征场和一个仿射变换来表示。

$$\mathbf{T} = \{\mathbf{s}, \mathbf{t}, \mathbf{R}\} \quad (5)$$

其中，s 是指示向量，t 是仿射变换的参数，R 是一个旋转矩阵。使用这种表示方法后，我们将物体的特征点转化到场景空间中，如下：

$$k(\mathbf{x}) = \mathbf{R} \cdot \begin{bmatrix} s_1 & & \\ & s_2 & \\ & & s_3 \end{bmatrix} \cdot \mathbf{x} + \mathbf{t} \quad (6)$$

在实践中，我们对场景中进行体绘制，并计算对象空间的特征场。 (详见 图. 1):

$$(\sigma, \mathbf{f}) = h_\theta(\gamma(k^{-1}(\mathbf{x})), \gamma(k^{-1}(\mathbf{d})), \mathbf{z}_s, \mathbf{z}_a) \quad (7)$$

这就能允许我们在一个场景中生成多种对象，所有的对象特征场都共享它们的权重，并且 T 是从一个与数据集相关的分布里抽样得到的。 (详见 Sec. 3.4).

3.2. 场景叠加

正如先前所讨论的，我们将 N 个实体所组成的图像称为场景，并把前 N -1 个对象描述成在场景中，最后一个对象作为背景。我们考虑到两种情况：第一，N 在数据集上是不变的，这就意味着图像中总有 N-1 个物体和另外一个作为背景的对象。第二：N 在数据集上是可变的。在实验中，我们对背景与物体使用相同的表示方法，此外我们固定参数 Sn 与 Tn 的规模与变换形式来张成空间，并以场景空间中的原点作为变换中心。

合成操作：为了定义合成操作 C，让我们回顾一下：单个实体的特征场能够预测定点 X 和视角 d 的密度和特征向量。当组合无实体物体时，点 X 处的总体密度是用将（x，d）处所有的特征密度加起来，并用密度加权平均值的方法来计算：

$$C(\mathbf{x}, \mathbf{d}) = \left(\sigma, \frac{1}{\sigma}\sum_{i=1}^{N}\sigma_i \mathbf{f}_i\right), \text{where} \quad \sigma = \sum_{i=1}^{N}\sigma_i \quad (8)$$

除了简单与直观之外，选择合成操作 C 还有其他好处：我们能保证梯度能流经所有密度大于 0 的实体。

3.3. 场景渲染

三维体渲染: 之前的研究[47,57,61, 77]将 RGB 三通道颜色值进行体渲染，我们将此方式扩展成渲染 M_f 维的特征向量 f。

对于给定的相机外参ξ，设$\{X_j\}_{j=1}^{N_s}$是给定像素相机沿射线方向 d 的一个采样点。并且$(\sigma_j, \mathbf{f}_j) = C(\mathbf{x}_j, \mathbf{d})$是特征场对应的密度和特征向量。体渲染操作[37] 将计算后的结果映射到像素最后的特征向量 f 上：

$$\pi_{\text{vol}} : (\mathbb{R}^+ \times \mathbb{R}^{M_f})^{N_s} \to \mathbb{R}^{M_f}, \quad \{\sigma_j, \mathbf{f}_j\}_{j=1}^{N_s} \mapsto \mathbf{f} \quad (9)$$

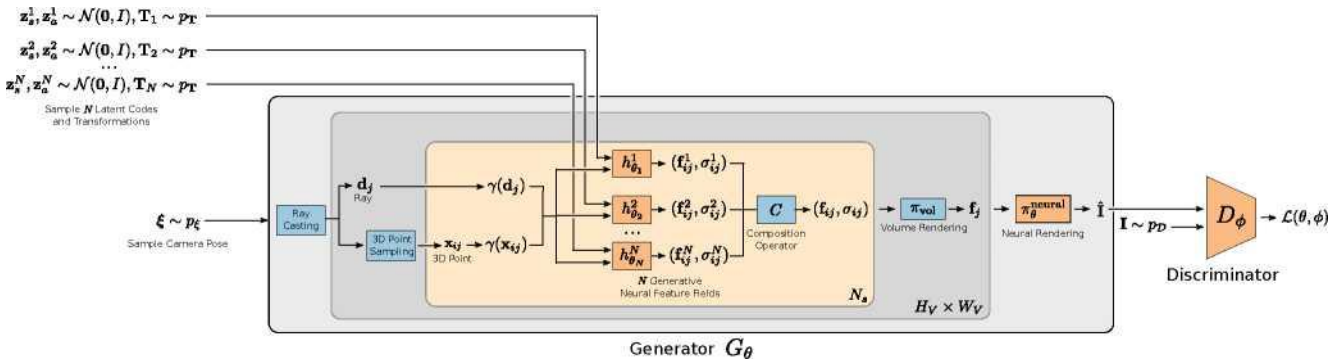


图 3: **GIRAFFE. 我们将相机姿态**ξ和 N 个对象的形状掩码Z_S^i与外观掩码Z_a^i、仿射变换 $\mathbf{T}_i$ 作为生成器的输出，然后生成一个具有 N - 1 个对象和 1 个背景的生成图像。判别器D_ϕ以生成图像$\hat{I}$和真实图像 **I** 作为输入，我们的模型用 GAN 的标准损失函数来进行训练。测试时，我们能够控制生成图像中相机位姿，物体的形状与外观，以及物体在场景中的姿态。橙色表示可学习操作，蓝色表示不可学习操作。

利用 NeRFs 中数值计算后的结果，我们可以得到 f：

$$\mathbf{f}=\sum_{j=1}^{N_s}\tau_j\alpha_j\mathbf{f}_j \quad \tau_j=\prod_{k=1}^{j-1}(1-\alpha_k) \quad \alpha_j=1-e^{-\sigma_j\delta_j} \tag{10}$$

其中，T_j 是透射率，a_j 是 $\mathbf{X}_j$ 的 alpha 值，$S_j=\|\mathbf{x}_{j+1}-x_j\|_2$ 是邻近的两个点之间的距离（欧氏距离）。我们可以通过计算每个像素的来获得全部的特征图像。为了高效性，我们渲染特征图像的分辨率是 16*16 的，远低于输出图像的分辨率 64*64 和 256*256。然后，我们通过 2 维的神经渲染将低分辨率的特征图上采样成更高分辨率的 RGB 图像。结果表明，这个方法有两个优点：加快渲染速度和提高图像质量。

2D 神经渲染： 含权重θ**的**神经渲染操作

$$\pi_\theta^{\text{neural}}:\mathbb{R}^{H_V\times W_V\times M_f}\rightarrow\mathbb{R}^{H\times W\times 3} \tag{11}$$

将特征图映射成最后的合成图像。我们参数化 ，使之作为一个二维的含 leaky ReLU 激活函数[56,89]的卷积神经网络(CNN)(Fig. 4)。为了提高图像的空间分辨率，我们还将最近邻插值和 3*3 的卷积核相结合起来。为了在提高输出图像分辨率的同时避免图像合成过程中场景的主要含义发生改变，我们选择小的卷积核，并且不在网络中加中间层，只允许在空间上做小的变化。这些方法可以避免图像生成时发生解纠缠，同时还可以提高输出分辨率。受[40],的启发。我们将特征图像映射到不同分辨率的 RGB 图像上，并用双线性上采样的方法将该分辨率下的 RGB 图像复合到下一分辨率的图像上。上下分辨率所输出图像之间的连接使得一个比较大的梯度值能够流经整个特征场。然后在最后一个 RGB 层中用 sigmoid 激活函数，我们就能获得最终预测的合成图像。我们也用消融实验来验证了我们的方法(Tab. 4)。

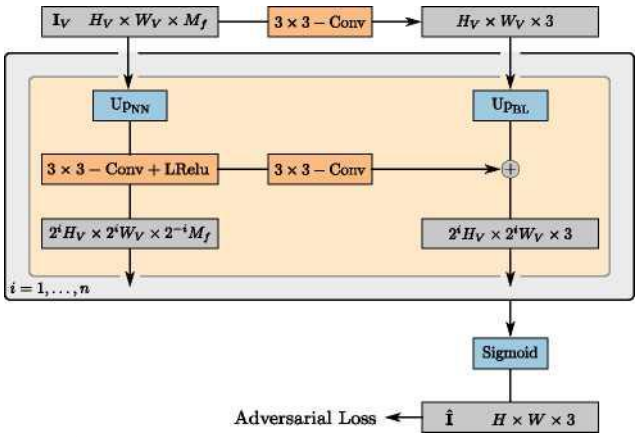


图 4: **神经渲染操作。**特征图像I_V经过 n 个最近邻上采样块和一个含 leaky ReLU 的激活函数的 3*3 卷积核来处理。在每个分辨率下，我们将特征图像映射到一个具有 3*3 卷积核的 RGB 图像上并进行卷积操作，并通过双线性上采样将其添加到之前的输出中。最后再用个 sigmoid 激活函数来获得最终的图像$\hat{I}$。灰色表示输出，橙色表示可学习的操作，蓝色表示不可学习到的操作。

3.4. 训练过程

生成器: 我们将完整的生成过程表示成:

$$G_\theta(\{\mathbf{z}_s^i,\mathbf{z}_a^i,\mathbf{T}_i\}_{i=1}^N,\boldsymbol{\xi})=\pi_\theta^{\text{neural}}(\mathbf{I}_V)$$
$$\text{where}\quad \mathbf{I}_V=\{\pi_{\text{vol}}(\{C(\mathbf{x}_{jk},\mathbf{d}_k)\}_{j=1}^{N_s})\}_{k=1}^{H_V\times W_V} \tag{12}$$

其中N是场景中的实体数目，N_s是每条射线所经过的采样点的数目，d_K表示第 k 个像素所引发的射线，X_{jk}表示第 K 个像素的第 j 个采样点。

判别器: 我们用含 Leaky ReLU 激活函数的 CNN[73]作为判别器D_ϕ。

训练过程:在训练过程中，我们对场景中的实体数目进行随机采样，隐码Z_s^i服从均值为 0，方差为 I 的高斯分布。相机姿态ξ服从分布p_ξ，对象的变换T_i服从分布p_T。在实验中，我们定义分布p_ξ和p_T分别为依赖数据集图像的相机仰角和物体变换的均匀分布。这样做的原因是，在大部分现实场景中，物体的安置角度是任意的，但由于重力的原因，物体不可能是倾斜着的。观察者（相机也算）却可以自由地改变观察角度。

我们用非饱和的[24] 、能够进行梯度惩罚 [58]的 GAN 来训练我们的模型

$$\mathcal{V}(\theta,\phi) = \mathbb{E}_{\mathbf{z}_s^i,\mathbf{z}_a^i\sim\mathcal{N},\boldsymbol{\xi}\sim p_\xi,\mathbf{T}_i\sim p_T}\left[f(D_\phi(G_\theta(\{\mathbf{z}_s^i,\mathbf{z}_a^i,\mathbf{T}_i\}_i,\boldsymbol{\xi}))\right] + \mathbb{E}_{\mathbf{I}\sim p_\mathcal{D}}\left[f(-D_\phi(\mathbf{I})) - \lambda\|\nabla D_\phi(\mathbf{I})\|^2\right] \tag{13}$$

其中$f(t) = -\log(1+\exp(-t))$，λ =10, p_D表示数据分布。

3.5. 实验细节

所有的对象特征域共享它们的权重，我们使用带 ReLU 函数的多层感知机来训练参数。感知机有 8 层，包括一个 128 维的隐藏层，1 维和 128 维的主要特征及其特征密度。对于背景特征域，我们使用的感知机层数减半，隐藏层的维数也减半。我们使用 2*3*10 作为三维点的位置编码维数，2*3*4 作为视角的位置编码维数。在每个视角方向上我们随机采样 64 个点，并以 16*16 像素来渲染特征图像。我们使用带 0.999 衰减的滑动平均[93]来计算生成器的权重。还使用块大小为 32，学习率为 0.0001（对于判别器）、0.0005（对于生成器）的 RMSprop 优化器[85] 。对于 256*256 的图像，我们设置特征图的维数为 256，生成器的学习率设为 0.00025。

4. 实验

数据集: 我们在一些常用的单对象数据集（如：Chairs[68]，Cats[95]，CelebA[52]，CelebA-HQ[38]）上进行了实验。首先，我们对椅子的图像集进行合成渲染[70]，之后我们用猫和人脸的混合图像集进行实验。由于图像背景要么是纯白的，要么是仅在图像中占一小部分，大大地限制了数据集的复杂性。进一步，我们在更具有挑战性的单对象真实场景数据集（如：*CompCars*[91]，*LSUN Churches*[94]，*FFHQ*[39]）中进行实验。对于数据集 *CompCars*，我们任意裁剪图像集中的图像，以实现物体在图像中位置的多样性。对于这些数据集来说，将这些物体从背景中分离变得更困难了，因为物体并不总在图像中心，背景更加地复杂且在图像中占比变大。为了在多对象场景中测试我们的模型，我们使用[36] 的带 2、3、4、5 个随机物体(*Clevr- N*)来进行场景渲染。为了测试我们的模型在不同数目对象的场景图的效果，我们还把这些数据集随机混合起来进行实验。 *(Clevr-2345).*

参照物（最基础的模型）: 我们的模型与基于体素的 PlatonicGAN[32]，BlockGAN[64]，HoloGAN[63]，基于辐射场的 GRAF [77]（具体模型相关讨论请看 Sec.2）进行比较。此外，我们的模型还和 HoloGAN 进行比较，HoloGAN 是能处理更高分辨率的图像的一种 GAN[63] 的变体，在[77]中被提出。另外我们还对比了基于 ResNet[28] 的二维 GAN [58] 。

量化评估方法: 我们通过 FID 分值[33]来量化评估图像的质量，我们用 20000 个真样本与假样本来计算 FID 值。

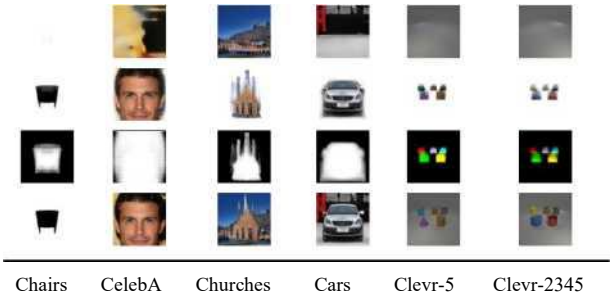


图 5: 结耦**场景。我们**从上到下显示了单背景图，单对象图对象颜色编码的 alpha 映射图, 以及 64*64 分辨率的最终合成图像。在无监督的条件下进行结耦，即使训练集上只有含对象的图像（意思是无纯背景图像），模型也会学到如何生成合理的背景。

图 6: **训练过程。**我们对 256*256 分辨率的 *Clevr-2345* 图像集上的一张图像分别进行 0，1，2，3，10 及 1 百万次渲染，然后发现在训练的最开始就产生了图像结耦。

	Cats	CelebA	Cars	Chairs	Churches
2D GAN [58]	18	15	**16**	59	19
Plat. GAN [32]	318	321	299	199	242
BlockGAN [64]	47	69	41	41	28
HoloGAN [63]	27	25	17	59	31
GRAF [77]	26	25	39	34	38
Ours	**8**	**6**	**16**	**20**	**17**

表 1: **定量比较。**我们比较了我们的模型与基准模型在 64*64 分辨率图像上的 FID 分值。

	CelebA-HQ	FFHQ	Cars	Churches	Clevr-2
HoloGAN [63]	61	192	34	58	241
w/o 3D Conv	33	70	49	66	273
GRAF [77]	49	59	95	87	106
Ours	**21**	**32**	**26**	**30**	**31**

表 2: **定量比较。** 我们比较了我们的模型与基准模型在 256*256 分辨率图像上的 FID 分值。

2D GAN	Plat.GAN	BlockGAN	HoloGAN	GRAF	Ours
1.69	381.56	4.44	7.80	0.68	0.41

表 3: 网络参数**比较。**我们比较了网络参数中生成器所需的存储量。

4.1. 可控场景生成

结耦图像生成：我们首先分析了我们的模型学到了多少结耦场景合成的能力，其中，我们对使物体从背景当中结耦出来特别感兴趣。为了达到这个目的，我们对图像的操作仅是简单的加法运算(Eq. 8)，渲染对象的特征并对对象进行 alpha 映射(Eq. 10)。请注意，训练时我们通常用 16*16 的分辨率来渲染特征图，但在测试时我们可以选择任意分辨率。

图. 5 表明我们的方法能从背景中结耦物体，注意这个结耦是在无监督信号的条件下完成的，而且我们的模型在没有纯背景的图像集上学到了如何生成合理的背景，隐式地解决了合成不存在场景图像的问题。我们进一步观测到，我们的模型能在多对象场景中正确的分离开单个对象。更进一步，我们发现无监督结耦场景是我们模型的一大特征，而且这个特征在训练最开始就已经出现了(图. 6)。注意观察我们的模型如何在输出背景之前合成单个物体。

可控场景生成：由于场景中的各个对象都已经被正确地结耦出来，我们就能分析它们到底可控到哪种地步。具体来说，我们好奇是否能对单个物体进行旋转变换，还有控制它们形状外观的改变。图 7 ,我们展示了我们的模型控制生成场景一些的例子。我们在 3 维空间里对单个物体进行旋转变换，或者是改变相机的仰角。通过用不同的隐码对每一个实体的形状和外形进行建模，我们能只改变物体的外观而不改变其形状。

生成训练集里没有的图像：模型学到的对于组合场景的表示能够让我们去生成训练集里所没有的图像。举个例子，我们能够扩展对象的范围或者增加比训练集里更多的对象(图. 8)。

4.2. 与基准模型的比较

相比于基准模型，我们的模型能够在分辨率为 64*64(表. 1)和 256*256(表. 2) 的图像上获得相近或更高的 FID 分数。定性来说，我们观察到虽然这两种方法都能在有限复杂度的数据集上合成可控图像，但是基准模型在背景杂乱的复杂图像上得到的结果缺少一致性。但是，由于我们的模型将物体从背景中独立出来，我们就可以单独控制物体了(图. 9)。

接着我们又注意到，我们的模型比起基于 2 维 GAN 的 ResNet 来说，在网络参数更少的同时，还能够达到相近或者更好的 FID 分数（0.41M 与 1.69M 的参数所需内存比较）。

这个发现验证了我们最初的预想：使用三维表示作为归纳偏置能够获得更好的输出结果。为了实验的准确和公正性，我们只对比相似的网络大小和训练时间的模型(详见 表. 3)。

4.3. 消融实验

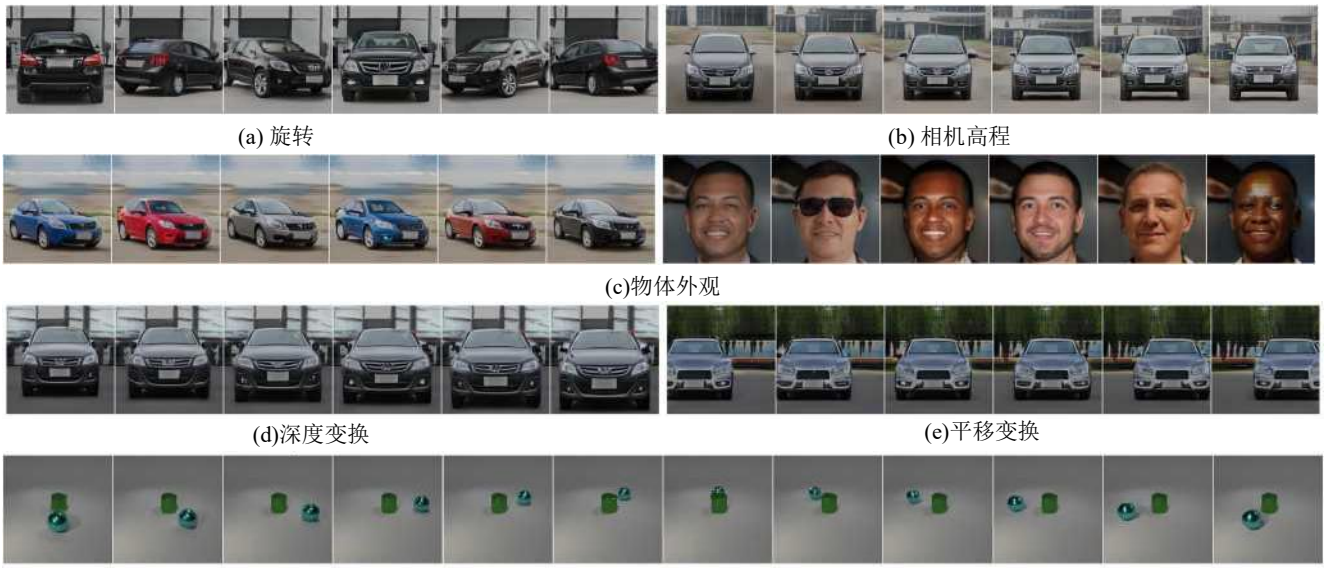
单分量的重要性: 表. 4 中的消融实验向我们展示了 RGB 残差连接的选择、最终激活函数的选择、上采样方法等都可以让模型产生更高的 FID 分数。

神经渲染的作用: 和 [77] 相比，我们方法的一个关键不同点是我们将体素与神经渲染相结合。从定量 (表. 1 和 2) 定性分析 (图. 9)来看，我们的方法在生成图像上效果更好，特别是在复杂、真实的场景数据中。我们的模型表达能力更强，更能够把握好真实场景的复杂性。 (图. 10)观察神经渲染器是如何让对象的外观适应场景而变。接下来，我们发现神经渲染器加速了图像合成：相比于[77]的模型，我们的模型在 64*64 像素的图像下，渲染时间从 110.1ms 降至 4.8ms，256*256 像素的图像下，渲染时间从 1595.0ms 降至 5.9ms。

位置编码: 我们对输入点坐标和视角使用轴对齐的位置编码(Eq. 1)，惊喜地发现它能让模型学到一种规范的表示方法。这种位置编码会产生一种偏差来使对象轴与标准坐标轴相对齐，这样子模型能够利用对象的对称性来进行操作。(图. 11)

4.4. 局限性

数据集的误差: 如果数据集中存在某种固有误差，我们的模型就很难消除由这个误差产生的影响。我们在图 12 中 展示了一个例子：在 celebA-HQ 数据集中，不管人脸正对哪里，眼睛和头发都正对着相机方向。 当旋转物体的时候，眼睛



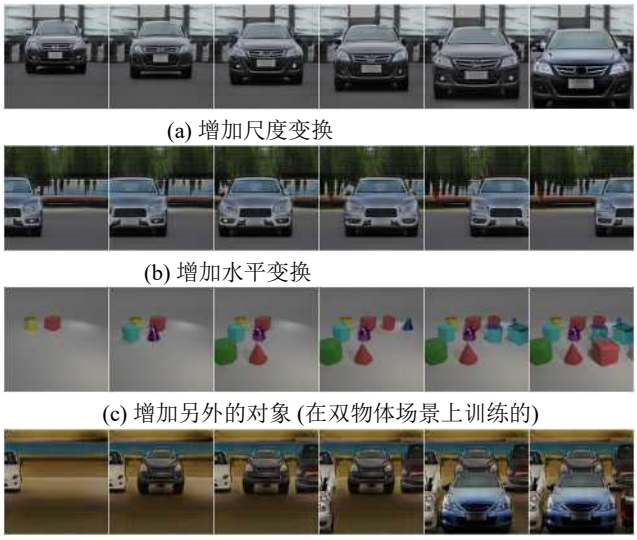
(a) 旋转　(b) 相机高程

(c)物体外观

(d)深度变换　(e)平移变换

(f)一个对象绕着另一个对象旋转

图 7: **在 256*256 分辨率下的可控场景图像生成。**在图像生成时控制场景的生成：在这里我们对物体进行旋转或者其他变换，改变它们的外观，或者对它们进行像绕轴旋转之类的复杂操作。



(a) 增加尺度变换

(b) 增加水平变换

(c) 增加另外的对象 (在双物体场景上训练的)

(d) 增加另外的对象 (在单物体场景上训练的)

图 8: **生成训练集中所没有的图像。**因为单个物体被准确结耦出来了，我们的模型就能在测试时生成一定的分布样本。例如，我们可以让物体呈现出更多的变换，或者在场景中生成比训练时更多的物体。

(a) HoloGAN [63]实现物体的 360 度旋转

(b)GRAF [77]实现物体的 360 度旋转

(c) 我们的模型实现物体的 360 度旋转

图 9: **定性比较.** 相比于基准模型, 我们的模型能够在背景杂乱的复杂场景图像中（64*64 分辨率在上，256*256 分辨率在下）达到更一致的效果。这是因为我们将物体从背景中结耦出来，并能固定背景，只旋转物体。

Full	-Skip	-Act.	+NN. RGB Up.	+Bi. Feat. Up.
16.16	16.66	21.61	17.28	20.68

表 4: **消融实验。**我们分别展示了在 *CompCars* 数据集上，我们的全模型、以及无 RGB 残差连接(-Skip)、无最终激活函数(-Act.)、用最近邻上采样法(+ NN. RGB Up.)、用双线性上采样法(+ Bi. Feat. Up.)的方法分别获得的 FID 分值 。

图 10: **神经渲染。**我们用我们的模型来对 256*256 分辨率的图像上在基于前景不变的条件下做背景改变。观察神经渲染器是怎样适应背景的改变。

(a) 基于旋转轴的位置编码 [61]的 0 度旋转

(b) 基于随机傅立叶特征 [82]的 0 度旋转

图 11: **典型姿势。** 相比于随机傅立叶特征 s [82], 基于旋转轴的位置编码(1) 能够让模型学到物体的常见姿势。

图 12: **数据集偏差。** 对于眼睛与头发的旋转是说明数据集偏差的例子。他们的脸朝向镜头，但我们的模型想要将他们的脸进行旋转。

和头发并不是固定在某一位置上的，它们同样也要适应数据集产生的偏差。

对象转化分布图：有时候我们会观察到结耦失败，例如：对于背景中包含教堂的 Churches 数据集图像，或是前景与背景相似的 CompCars 的数据集图像(详见 Sup. Mat.)。我们将结耦失败的原因归咎于图像集中相机位姿呈均匀分布，但这和对象级的转化及相机真实位姿分布不匹配。

5. 结论

我们提出了 *GIRAFFE，一种生成可控图像的新方法*。我们的核心思想是将叠加后的三维场景同生成模型相结合。通过将场景表示成组合的神经特质场。我们的模型能在无监督的情况下将单个物体与背景、不同形状与外观的物体结耦出来。将其与神经渲染场相结合，能快速生成可控图像。未来我们计划从数据中学得对象级别的变换和相机位姿的分布。此外，若将我们的方法与一些易监督的任务相结合（如：预测对象掩码），就可以成为一种应对复杂的多对象场景的好方法。

致谢

这项研究工作得到了 NVIDIA 的特别赞助。同时，我们也感谢 IMPRS-IS 对 Michael Niemeyer 的支持和 ERC 及 DFG EXC 编号为 2064/1（项目编号 390727645）对 Andreas Geiger 的支持。

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv.org,* 2008.02401, 2020. 2

[2] Hassan Alhaija, Siva Mustikovela, Andreas Geiger, and Carsten Rother. Geometric image synthesis. In *Proc. o^fthe Asian Conf, on Computer Vision (.ACCV),* 2018. 2

[3] Titas Anciukevicius, Christoph H. Lampert, and Paul Henderson. Object-centric image generation with factored depths, locations, and appearances. *arXiv.org,* 2004.00642, 2020. 2

[4] Relja Arandjelovic and Andrew Zisserman. Object discovery with a copy-pasting GAN. *arXiv.org,* 1905.11369, 2019. 2

[5] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*,35(8):1798-1828,2013. 1

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. of the International Conf. on Learning Representations(ICLR),2019.* 1,2

[7] Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew M Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv.org*, 1901.11390, 2019. 2

[8] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard A. Newcombe. Deep local shapes: Learning local SDF priors for detailed 3d reconstruction. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[9] Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 1, 2

[10] Xuelin Chen, Daniel Cohen-Or, Baoquan Chen, and Niloy J. Mitra. Neural graphics pipeline for controllable image generation. *arXiv.org*, 2006.10569, 2020. 2

[11] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. BAE-NET: branched autoencoder for shape co-segmentation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[13] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[14] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[15] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains.

[16] Edo Collins, Raja Bala, Bob Price, and Sabine Sui sstrunk. Editing in style: Uncovering the local semantics of gans. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[17] Robert A. Drebin, Loren C. Carpenter, and Pat Hanrahan. Volume rendering. In *ACM Trans. on Graphics,* 1988. 4

[18] Sebastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy J. Mitra, and Andrea Vedaldi. RELATE: physically plausible multi-object scene synthesis using structured latent spaces. *arXiv.org*, 2007.01272, 2020. 2

[19] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: generative scene inference and sampling with object-centric latent representations. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020. 2

[20] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ru- derman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil C. Rabinowitz, Helen King, Chloe Hillier, Matt M. Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360:1204-1210, 2018. 3

[21] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *Proc. of the International Conf, on 3D Vision (3DV),* 2017. 2

[22] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[23] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS),* 2014. 1, 2, 5

[25] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schoi lkopf. Recurrent independent mechanisms. *arXiv.org*, 1909.10893, 2019. 1

In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[26] Klaus Greff, Raphaeil Lopez Kaufmann, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proc. of the International Conf. on Machine learning (ICML)*, 2019. 2

[27] Erik Hai rkoi nen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable GAN controls. *arXiv.org*, 2004.02546, 2020. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE*

[29] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision (IJCV),* 2019. 2

[30] Paul Henderson and Christoph H. Lampert. Unsupervised object-centric video generation and decomposition in 3d. *arXiv.org,* 2007.06705, 2020. 2

[31] Paul Henderson, Vagia Tsiminaki, and Christoph H. Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* 2020. 2

[32] Philipp Henzler, Niloy J Mitra, , and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2, 5, 6

[33] 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. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 6

[34] Ali Jahanian, Lucy Chai, and Phillip Isola. On the ‘‘steerability” of generative adversarial networks. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020. 2

[35] Chiyu Max Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias NieBner, and Thomas A. Funkhouser. Local implicit grid representations for 3d scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[36] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[37] James T. Kajiya and Brian Von Herzen. Ray tracing volume densities. In *ACM Trans. on Graphics,* 1984. 4

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018. 5

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition(CVPR),2019.* 1,2,5

[40] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando,

Tony Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey. *arXiv.org*, 2006.12057, 2020. 3

[41] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[42] Hanock Kwak and Byoung-Tak Zhang. Generating images part by part with composite generative adversarial networks. *arXiv.org,* 1607.05387, 2016. 1

[43] Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. *arXiv.org*, 2001.04296, 2020. 1

[44] Nanbo Li, Robert Fisher, et al. Learning object-centric representations of multi-object scenes from multiple views. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[45] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[46] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[47] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *arXiv.org*, 2008.02793, 2020. 1

[48] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: Differentiable rendering for unsupervised singleview mesh reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 3

[49] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[50] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. DIST: rendering deep implicit signed distance function with differentiable sphere tracing. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[51] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2015. 5

[52] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Ratsch, Sylvain Gelly, Bernhard Scholkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. of the International Conf. on Machine learning (ICML)*, 2019. 1

[53] Francesco Locatello, Dirk Weissenborn, Thomas Un- terthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Objectcentric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[54] Sebastian Lunz, Yingzhen Li, Andrew W. Fitzgibbon, and Nate Kushman. Inverse graphics GAN: learning to generate 3d shapes from unstructured 2d data. arXiv.org, 2020. 2

[55] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. of the International Conf. on Machine learning (ICML) Workshops*, 2013. 4

[56] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *arXiv.org,* 2008.02268, 2020. 4

[57] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *Proc. of the International Coif on Machine learning (ICML),* 2018. 5,6,7

[58] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[59] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[60] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2, 3, 4, 8

[61] Thu Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yong-Liang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 3

[62] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2, 5, 6, 7

[63] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong- Liang Yang, and Niloy Mitra. Blockgan: Learning 3d object- aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5, 6

[64] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[65] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[66] Michael Oechsle, Michael Niemeyer, Christian Reiser, Lars Mescheder, Thilo Strauss, and Andreas Geiger. Learning implicit surface light fields. In *Proc. of the International Conf. on 3D Vision (3DV),* 2020. 5

[67] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation.

[68] William S. Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 1, 2

[69] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[70] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2016. 5

[71] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *Proc. of the International Conf. on Machine learning (ICML)*, 2014. 1

[72] Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2

[73] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[74] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3, 4, 5, 6, 7, 8

[75] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[76] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[77] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias NieBner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[78] Vincent Sitzmann, Michael Zollhofer, and Gordon Wetzstein. Scene representation networks: Continuous 3dstructure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 2, 3

[79] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ra-mamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurlPS),* 2020. 3, 8

[80] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin- Brualla, Tomas Simon, Jason M. Saragih, Matthias NieBner, Rohit Pandey, Sean Ryan Fanello, Gordon Wetzstein, Jun- Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhofer. State of the art on neural rendering. *Computer Graphics Forum*, 2020. 3

[81] Justus Thies, Michael Zollhofer, and Matthias NieBner. Deferred neural rendering: image synthesis using neural textures. *ACM Trans. on Graphics,* 2019. 3

[82] T. Tieleman and G. Hinton. Lecture 6.5 — RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012. 5

[83] Sjoerd van Steenkiste, Karol Kurach, Jurgen Schmidhuber, and Sylvain Gelly. Investigating object compositionality in generative adversarial networks. *Neural Networks*, 2020. 2

[84] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *Proc. ofthe European Conf. on Computer Vision (ECCV),* 2016. 2

[85] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems (NIPS),* 2016. 2

[86] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv.org*, 1505.00853, 2015. 4

[87] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. LR-GAN: layered recursive generative adversarial networks for image generation. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2017. 2

[88] Jiaolong Yang and Hongdong Li. Dense, accurate optical flow estimation with piecewise parametric model. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[89] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[90] Yasin Yazici, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018. 5

[91] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianx- iong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv.org*, 1506.03365, 2015. 5

[92] Li Zhang, Brian Curless, Aaron Hertzmann, and Steven M. Seitz. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multiview stereo. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2003. 5

[93] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yi- nan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. *arXiv.org*, 2010.09125, 2020. 2

[94] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal. Modular generative adversarial networks. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018. 1

[95] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2015. 1

[96] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.