State of the Art on Neural Rendering

Dezeming Family

2023 年 4 月 30 日

DezemingFamily 系列文章和电子书**全部都有免费公开的电子版**，可以很方便地进行修改和重新发布。如果您获得了 DezemingFamily 的系列电子书，可以从我们的网站 [https://dezeming.top/] 找到最新的版本。对文章的内容建议和出现的错误也欢迎在网站留言。

目录

一 摘要和引文 **1**

二 Related Surveys and Course Notes **2**

三 Scope of this STAR **2**

四 Theoretical Fundamentals **3**

4 1 Physical Image Formation . . . 3
4 2 Physical Image Formation . . . 3
4 3 Deep Generative Models . . . 3
4 4 学习一个生成器 . . . 4
4 5 使用感知距离学习 . . . 4
4 6 使用条件 GAN 学习 . . . 4
4 7 在没有成对数据条件下学习 . . . 6

五 神经渲染 **7**

5 1 控制 . . . 7
5 2 Computer Graphics Modules . . . 7
5 3 Explicit vs. Implicit Control . . . 8
5 4 Multi-modal Synthesis . . . 8
5 5 Generality . . . 8

六 神经渲染的应用 **9**

6 1 Semantic Photo Synthesis and Manipulation . . . 9
6 1.1 Semantic Photo Synthesis . . . 9
6 1.2 Semantic Image Manipulation . . . 10
6 1.3 Improving the Realism of Synthetic Renderings . . . 10
6 2 Novel View Synthesis for Objects and Scenes . . . 11
6 2.1 Neural Image-based Rendering . . . 11
6 2.2 Neural Rerendering . . . 12
6 2.3 Novel View Synthesis with Multiplane Images . . . 12
6 2.4 Neural Scene Representation and Rendering . . . 12
6 2.5 Voxel-based Novel View Synthesis Methods . . . 13
6 2.6 Implicit-function based Approaches . . . 13

6 3 Free Viewpoint Videos . . . 14
6 3.1 LookinGood with Neural Rerendering . . . 14
6 3.2 Neural Volumes . . . 14
6 3.3 Free Viewpoint Videos from a Single Sensor . . . 15
6 4 Learning to Relight . . . 15
6 4.1 Deep Image-based Relighting from Sparse Samples . . . 15
6 4.2 Multi-view Scene Relighting . . . 15
6 4.3 Deep Reflectance Fields . . . 16
6 4.4 Single Image Portrait Relighting . . . 16
6 5 Facial Reenactment . . . 16
6 5.1 Deep Video Portraits . . . 17
6 5.2 Editing Video by Editing Text . . . 17
6 5.3 Image Synthesis using Neural Textures . . . 17
6 5.4 Neural Talking Head Models . . . 17
6 5.5 Deep Appearance Models . . . 17
6 6 Body Reenactment . . . 17

七 Open Challenges **18**
7 1 泛化 . . . 18
7 2 Scalability . . . 18
7 3 Editability . . . 18
7 4 Multimodal Neural Scene Representations . . . 18

八 Social Implications **20**

参考文献 **20**

一　摘要和引文

本文 [1] 的时间相对较早，是 2020 年的产物，那时很多技术还没有成熟。

高效渲染照片级真实感的虚拟世界是计算机图形学长期以来的努力。现代图形技术已经成功地从手工制作的场景表示中合成了照片级真实感的图像。然而，场景的形状、材质、照明和其他方面的自动生成仍然是一个具有挑战性的问题，如果解决了这个问题，将使照片级真实感的计算机图形更容易获得。

与此同时，计算机视觉和机器学习的进步催生了一种新的图像合成和编辑方法，即深度生成模型。神经渲染是一个新兴领域，它将生成性机器学习技术与计算机图形学中的物理知识相结合，例如通过将可微渲染集成到网络训练中。

虚拟世界照片级真实感图像的创建一直是复杂计算机图形技术发展的主要驱动力之一。计算机图形学的方法涵盖了从实时渲染到复杂的全局光照模拟的范围，实时渲染使最新一代的计算机游戏成为可能，复杂的全局光照模拟用于在故事片中创建照片级真实感的数字人。在这两种情况下，主要瓶颈之一是内容创建，即，在表面几何形状、外观/材质、光源和动画方面，需要熟练的艺术家进行大量乏味且昂贵的手工工作来创建底层场景表示。与此同时，强大的生成模型已经出现在计算机视觉和机器学习领域中。Goodfellow 等人在生成对抗性神经网络（GAN）方面的开创性工作近年来已发展成为用于创建高分辨率图像和视频的深度生成模型。可以通过在控制参数或来自其他域的图像上调节网络来实现对合成内容的控制。最近，这两个领域走到了一起，并被探索为“神经渲染”。随着计算机图形学和视觉领域的大量应用，神经渲染有望成为图形学界的一个新领域，但目前还没有对这一新兴领域的调查。

最早使用神经渲染一词的出版物之一是生成查询网络（GQN）。它使机器能够基于表示和生成网络来学习感知周围环境。作者认为该网络具有 3D 的隐含概念，因为它可以将不同数量的场景图像作为输入，并输出具有正确遮挡的任意视图。取而代之的是隐含的 3D 概念 (implicit notion)，随后出现了各种其他方法，其中更明确地包括 3D 概念，利用图形管线的组件。

这篇最新的报告总结了神经渲染的最新趋势和应用。我们专注于将经典的计算机图形技术与深度生成模型相结合的方法，以获得可控和逼真的输出。从基本的计算机图形学和机器学习概念的概述开始，我们讨论了神经渲染方法的关键方面。具体而言，我们的重点是【控制类型 (type of control)，即如何提供控制，学习管线的哪些部分，显式与隐式控制，泛化，以及随机与确定性综合】。本最新报告的后半部分重点介绍了所描述算法的许多重要用例，如新颖的【视图合成 (novel view synthesis)、语义照片处理 (semantic photo manipulation)、面部和身体再现 (facial and body reenactment)、重新照明 (relighting)、自由视角视频 (free-viewpoint video)，以及为虚拟和增强现实远程呈现创建照片级真实感的数字化身 (photo-realistic avatars)】。最后，我们讨论了此类技术的社会影响，并调查了开放的研究问题。

虽然经典的计算机图形学是从物理学的角度出发的，例如通过建模几何、表面特性和相机参数；但机器学习是从统计学的角度出发，即从现实世界的例子中学习以生成新的图像。为此，计算机图形生成图像的质量取决于所用模型的物理正确性，而机器学习方法的质量主要取决于精心设计的机器学习模型和所用训练数据的质量。

场景属性的显式重建是困难的并且容易出错，并且会导致渲染内容中的伪影 (artifacts)。为此，基于图像的渲染方法 (image-based rendering methods) 试图通过使用简单的启发式方法 (heuristics) 来组合捕获的图像来克服这些问题。但在复杂的场景中，这些方法会显示接缝或重影 (seams or ghosting) 等伪影。

神经渲染有望通过使用深度网络学习从捕获图像到新图像的复杂映射来解决重建和渲染问题。神经渲染将物理知识（例如投影的数学模型）与学习的组件相结合，为【可控的图像生成过程 (controllable image generation)】产生新的强大算法。神经渲染在文献中还没有一个明确的定义。在这里，我们将神经渲染定义为：

- 基于深度学习的图像或视频生成方法，可以显式或隐式控制场景属性，如照明、相机参数、姿势、几何结构、外观和语义结构（注意这是神经渲染的关键所在）。

这份最先进的报告定义并分类了不同类型的神经渲染方法。我们的讨论重点是将【计算机图形学】和【基于学习】的基元相结合，为【可控图像生成】产生新的强大的算法的方法，因为图像生成过程中的【可控性】对许多计算机图形学应用至关重要。

我们构建本报告的一个核心方案是每种方法所提供的控制。我们首先讨论计算机图形学、视觉和机器学习的基本概念，这些概念是神经渲染的先决条件。然后，我们讨论了神经渲染方法的关键方面，例如:【控制类型、如何提供控制、学习管线的哪些部分、显式与隐式控制 (explicit vs. implicit control)、泛化 (generalization) 以及随机与确定性合成 (stochastic vs. deterministic synthesis)】。接下来，我们将讨论由神经渲染启用的应用前景。神经渲染的应用范围从【新颖的视图合成、语义照片处理、面部和身体再现、重新照明、自由视点视频，到为虚拟和增强现实远程呈现创建照片级真实感的化身 (avatars)】。由于创建和处理与真实照片无法区分的图像具有许多社会意义，特别是当人类被拍摄时，我们还讨论了这些含义 (implications) 和合成内容的可检测性 (detectability)。由于神经渲染领域仍在快速发展，我们总结了当前的开放研究问题。

二 Related Surveys and Course Notes

深度生成模型在学术界得到了广泛的研究，有几项 survey 和课程笔记对其进行了描述。一些报告侧重于特定的生成模型，如生成对抗性网络（GAN）和变分自动编码器（VAE）。

使用经典计算机图形学和视觉技术的可控图像合成也得到了广泛的研究。在几份调查报告中已经讨论了基于图像的渲染 (image-based rendering)。Szeliski 的书出色地介绍了 3D 重建和基于图像的渲染技术。一些最近的 Survey 讨论了各种应用中【人脸的 3D 重建和可控渲染】方法。

最近的计算机视觉会议的教程和研讨会已经涵盖了神经渲染的一些方面，其中包括【自由视点渲染】和【全身表演重新照明】的方法 (relighting of full body performances)，面部合成的神经渲染教程和使用神经网络的 3D 场景生成。

然而，上述 Survey 和课程都没有提供对神经渲染及其所有各种应用的结构化和全面的研究。

三 Scope of this STAR

在这份最新的报告中，我们重点关注将【经典计算机图形管线】和【可学习组件】相结合的新方法。具体来说，我们正在讨论在哪里以及如何通过机器学习改进经典渲染管线，以及训练需要哪些数据。

为了全面概述，我们还简要介绍了计算机图形学和机器学习这两个领域的相关基础知识。展示了当前混合方法的优点及其局限性。本报告还讨论了这些技术所赋予的新应用。我们专注于通过机器学习生成可控照片级真实感图像的技术。

我们不涉及几何和 3D 深度学习的工作，后者更侧重于 3D 重建和场景理解——这一分支工作推动了许多神经渲染方法，尤其是那些基于 3D 结构化场景 (3D-structured scene) 表示的方法，但超出了本 survey 的范围。

我们也没有关注使用机器学习对光线追踪图像进行去噪的技术。

四 Theoretical Fundamentals

第一节讨论图像生成（计算机图形学）和图像合成；第二节介绍基于图像的渲染；第三节讨论深度生成模型。

4 1 Physical Image Formation

模拟物理光传输的要素：light sources, scene geometry, material properties, light transport, optics, and sensor behavior。

场景几何有两种，一种是显式表示，比如三角 mesh；另一种是隐式表示，比如 signed distance functions mapping（$\mathbb{R}^3 \to \mathbb{R}$），表面被定义为函数的 zero-crossing（水平集）。在实践中，大多数硬件和软件渲染器被调整为在三角形网格上工作得最好，并且将其他表示转换为三角形进行渲染。

跨几何的空间变化行为可以通过将离散材质绑定到不同的几何图元来表示（比如一个基元都是该材质），或者通过使用纹理映射来表示（比如，纹理值作为粗糙度）。纹理图定义了一组连续的材料参数值，如从二维或三维域到表面的漫反照率。三维纹理表示整个有界空间区域的值，可以应用于显式或隐式几何体。二维纹理从二维域映射到参数化表面上；因此，它们通常仅适用于显式几何。

为了生成新视图、编辑材质或照明或创建新动画，从真实世界的数据中估计不同模型参数（摄影机、几何体、材质、灯光参数）的过程称为反向渲染 (inverse rendering)。反向渲染与神经渲染密切相关，在计算机视觉和计算机图形学的背景下进行了探索。反向渲染的一个缺点是，由于数学复杂性或计算费用，经典渲染中使用的预定义物理模型或数据结构并不总是准确地再现真实世界物理过程的所有特征。相反，神经渲染将学习组件引入到渲染管线中，以代替此类模型。深度神经网络可以在统计上近似这些物理过程，从而产生与训练数据更接近的输出，比反向渲染更准确地再现一些真实世界的效果。

注意，有一些方法在反向渲染和神经渲染的交集上。例如，使用近似全局光照效果的神经渲染来有效地训练预测深度、法线、反照率和粗糙度图的反向渲染方法。还有一些方法使用神经网络来增强经典渲染管线的特定构建块，例如着色器。再比如学习双向纹理函数 (Bidirectional Texture Functions)，或者学习外观贴图 (Appearance Maps)。

4 2 Physical Image Formation

与将 3D 内容投影到 2D 平面的经典渲染不同，基于图像的渲染技术通过变换现有的图像集来生成新颖的图像，通常是通过将它们 warping 并合成在一起。

基于图像的渲染可以处理动画，但最常见的用例是静态对象的新视图合成，其中基于代理几何结构和估计的相机姿态将捕获视图中的图像内容 warp 为新视图。为了生成一个完整的新图像，必须将多个捕捉到的视图 warp 到目标视图中，这需要一个混合阶段。生成的图像质量取决于几何体的质量、输入视图的数量和排列以及场景的材质属性，因为某些材质会在不同视点之间显著更改外观。尽管混合和校正视效的启发式方法显示出良好的结果，但最近的研究已经用学习组件取代了这些基于图像的渲染管线的一部分。深度神经网络已成功用于减少混合伪影和源于视图相关效应的伪影（论文第 6.2.1 节）。

4 3 Deep Generative Models

虽然传统的计算机图形学方法侧重于对场景进行物理建模和模拟光传输以生成图像，但机器学习可以通过学习真实世界图像的分布，从统计的角度来解决这个问题。与传统的基于图像的渲染相比，深度生成模型可以从大规模的图像集合中学习图像先验，传统的基于基于图像的渲染上使用了小的图像集（例如，数百个）。

深度生成模型研讨会学习生成简单 digits 和 frontal faces 的随机样本。在这些早期的结果中，无论是质量还是分辨率都远远不如使用基于物理的渲染技术实现。然而，最近，照片逼真的图像合成已经使用生成对抗性网络及其扩展进行了演示。最近的工作可以合成随机的高分辨率肖像，这些肖像通常与真实人脸无法区分。

深度生成模型擅长生成具有类似于训练集的统计数据的随机逼真图像。然而，用户控制和交互在图像合成和操作中起着关键作用。例如，概念艺术家希望创建反映其设计思想的特定场景，而不是随机场景。因此，对于计算机图形学应用，生成模型需要扩展到【条件设置】，以获得对图像合成过程的【明确控制】。

早期结果是对前馈神经网络使用每像素 l_p 距离来作为条件控制，假设每个像素都是独立的，而没有考虑到视觉结构的复杂性；此外，它倾向于对多个可能的输出进行平均。

为了解决上述问题，最近的工作提出了感知相似性距离，以在由预训练网络构建的高级深度特征嵌入空间中测量合成结果和 ground truth 输出之间的差异。应用包括艺术风格化、图像生成和合成和超分辨率。将输出与其 ground truth 图像相匹配并不能保证输出看起来自然。

条件 GANs（cGANs）的目标不是最小化输出和目标之间的距离，而是匹配给定输入的输出的条件分布。结果可能看起来与 ground truth 图像不一样，但它们看起来很自然。条件 GANs 已被用于弥合粗略的计算机图形渲染和相应的真实世界图像之间的差距，或在给定用户指定的语义布局的情况下生成逼真的图像。

下面我们为网络架构和学习目标提供更多技术细节。

4 4 学习一个生成器

图像输入 $x \in \mathcal{X}$，输出 $y \in \mathcal{Y}$。输入是用户的条件输入，比如草图、相机参数、光照条件、场景属性、纹理描述等。输出可以是一个图像、一个视频、或者 3D 数据比如 mesh 或体素。

FCN(Fully Convolutional Networks) 早期用于图像分类识别或者语义分割，但现在也经常用于图像合成。U-Net 是基于 FCN 的结构，由于跳跃连接，它增强了局部化能力。这些跳跃连接有助于产生更详细的输出，因为来自输入的高频信息可以直接传递到输出。基于 ResNet 的生成器把高频信息从输入传递到输出，已经被用于风格迁移或图像超分辨。

4 5 使用感知距离学习

最直接的损失函数就是输出和 Ground Truth 之间的距离 y：

$$\mathcal{L}_{recon}(G) = \mathbb{E}_{x,y}||G(x) - y||_p \tag{四.1}$$

不幸的是，学习的生成器倾向于合成模糊的图像或在多个看似合理的输出上的平均结果。例如，在图像着色中，由于平均效应，学习的生成器有时会产生去饱和的结果 (desaturated results)。在图像超分辨率中，生成器无法合成结构和细节，因为 p 范数独立地作用于每个像素，没有邻域信息。

为了设计更好地与人类对图像相似性的感知相一致的学习目标，最近的工作提出测量由预训练的图像分类器 F（例如，VGG 网络）提取的深度特征表示之间的距离。由于深度表示整体地概括了整个图像，这种损失会更好。即数学上，生成器尝试去最小化如下特征损失：

$$\mathcal{L}_{perc}(G) = \mathbb{E}_{x,y} \sum_{t=1}^{T} \lambda_t \frac{1}{N_t} ||F^{(t)}(G(x)) - F^{(t)}(y)||_1 \tag{四.2}$$

其中 $F^{(t)}$ 就是里面的中间层。

尽管上述距离通常被称为“感知距离”，但有趣的是，为什么多级深度特征空间中的匹配统计数据可以匹配人类的感知，并有助于合成更高质量的结果，因为网络最初是为图像分类任务而非图像合成任务训练的。最近的一项研究表明，强分类器学习的丰富特征也为人类感知任务提供了有用的表示，优于经典的手工感知度量。

4 6 使用条件 GAN 学习

Blau 和 Michaeli 证明过，最小化输入和输出的距离并不能保证产生真实感外观，他们还证明了小距离和写实主义是不一致的。因此，深度生成模型关注的不是距离最小化，而是分布匹配，即将生成结果的分布与训练数据的分布相匹配。在许多类型的生成模型中，生成对抗性网络（GANs）在许多计算机图形学任务中显示出了有希望的结果。

GAN 学习从低维随机向量到一个图像输出的映射关系，生成器 G 被训练以产生无法通过对抗训练的鉴别器 D 与“真实”图像区分的输出。鉴别器被训练以检测生成器生成的合成图像。虽然为人脸或车辆等对象类别训练的 GANs 学习合成对象的高质量实例，但合成的背景通常质量较低。最近的论文试图通过学习完整场景的生成模型来缓解这个问题。

通过增加条件信息，conditional GANs(cGAN) 学习从可观测输入 x 以及随机采样的 z 来输出图像 y。cGAN 学习映射关系 $G:\{x,z\}\to y$。观测输入 x 也会被传递到鉴别器，用于检测是否图像对 x,y 是真的还是假的。输入 x 和输出 y 都根据目标应用而变化。在类条件 GAN(class-conditional GANs) 中，输入 x 是一个分类标签，用于控制模型应该生成哪个对象类别。在诸如 pix2pix 的 image-conditional GANs 的情况下，生成器 G 旨在将输入图像 x（例如语义标签图）转换为逼真的输出图像，而鉴别器 D 旨在将真实图像与生成的图像区分开来。这种模型需要成对数据来训练，$\{x_i,y_i\}_{i=1}^N$ 就是 N 个图像对。

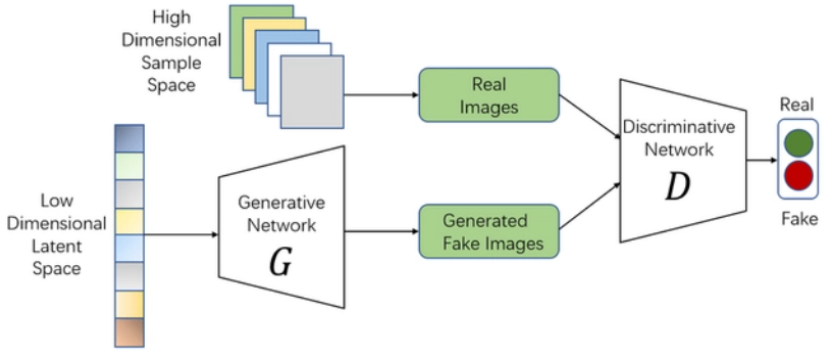
cGAN 通过以下极小极大 (minimax) 博弈匹配给定输入的输出的条件分布：

$$\min_G \max_D \mathcal{L}_{cGAN}(G,D) \tag{四.3}$$

其中，目标函数 $\mathcal{L}_{cGAN}(G,N)$ 一般是：

$$\mathcal{L}_{cGAN}(G,D)=\mathbb{E}_{x,y}[\log D(x,y)]+\mathbb{E}_{x,z}[\log(1-D(x,G(x,z)))] \tag{四.4}$$

注意最原始的 GAN



其优化目标：

$$\min_G \max_D \mathcal{L}_{GAN}(G,D)=\min_G \max_D \Big[\mathbb{E}_x[\log D(x)]+\mathbb{E}_z[\log(1-D(G(z)))]\Big] \tag{四.5}$$

这样优化可以达到两个目的，第一个目的是让生成器 G 能够生成真实的样本，第二个目的是让判别器 D 能更好地区分开真实样本和生成样本。

在早期的 cGAN 实现中，没有注入噪声向量，并且映射是确定性的，因为它在训练期间往往被网络忽略。最近的工作使用潜在矢量 z 来实现多模式图像合成。为了使训练更稳定，cGAN 方法也会采用每像素 l_1 损失和感知度量损失。

cGAN 增加了额外的条件信息，生成器生成的图片只有足够真实且与条件相符才能够通过判别器。

在训练期间，鉴别器 D 试图提高其区分真实图像和合成图像的能力，而生成器 G 同时试图提高其欺骗鉴别器的能力。pix2pix 方法采用 U-Net 作为生成器的架构，采用基于 patch 的全卷积网络作为鉴别器。

从概念上讲，感知距离和条件 GANs 是相关的，因为它们都使用辅助网络（F 或 D）来定义学习更好生成器 G 的有效学习目标。在高级抽象中，用于评估合成结果 $G(x)$ 质量的精确计算机视觉模型（F 或 D）可以显著帮助解决神经渲染问题。然而，存在两个显著差异。首先，感知距离旨在测量输出实例与其 ground truth 之间的差异，而条件 GANs 则测量真实图像和伪图像的条件分布的接近度。其次，对于感知距离，特征提取器 F 是预先训练和固定的，而条件 GANs 根据生成器动态调整其鉴别器 D。在实践中，这两种方法是互补的（使用 D 的模型以及使用 F 的模型），许多神经渲染应用程序同时使用这两种损失。除了 GANs，最近还出现了许多有前途的研究方向，包括变分自动编码器（VAE）、自回归网络（例如，PixelCNN、PixelRNN）、可逆密度模型等。StarGAN 能够基于具有不同域的多个数据集训练用于图像到图像翻译的单个模型。为了使讨论简明扼要，我们在这里重点讨论 GANs。

4 7　在没有成对数据条件下学习

学习具有上述目标的生成器需要数亿个成对的训练数据。在许多实际应用中，成对训练数据的收集既困难又昂贵。与为分类任务标记图像不同，注释器必须为图像合成任务标记每个像素。例如，对于像语义分割这样的任务，只有几个小的数据集。获得图形任务（如艺术风格化）的输入输出对可能更具挑战性，因为所需的输出通常需要艺术创作，有时甚至没有明确定义。

在这种设置下，我们可能只知道原域和目标域范围，但是没有它们的一一对应关系。因为从 $\mathcal{X}$ 到 $\mathcal{Y}$ 的映射不是一一对应的。因此，我们需要额外的约束。研究者们已经提出了几个约束条件，包括用于强制双射映射的循环一致性损失，用于鼓励输出在像素空间或特征嵌入空间中接近输入图像的距离保持损失，用于跨域学习共享表示的权重共享策略等。上述方法拓宽了条件 GAN 的应用范围，并实现了许多图形应用，如 object transfiguration、domain transfer 和 CG2real。

五　神经渲染

给定高质量的场景规范，经典渲染方法可以为各种复杂的现实世界现象渲染照片级真实感图像。此外，渲染使我们能够对场景摄影机视点、照明、几何体和材质的所有元素进行明确的编辑控制。然而，构建高质量的场景模型，特别是直接从图像中构建场景模型，需要大量的手动工作，而从图像中进行自动场景建模是一个开放的研究问题。另一方面，深度生成网络现在开始从随机噪声中产生视觉上引人注目的图像和视频，或者以场景分割和布局等特定用户规范为条件。然而，它们还不允许对场景外观进行细粒度控制，并且不能始终处理场景属性之间复杂的、非局部的 3D 交互。

相比之下，神经渲染方法有望将这些方法相结合，以实现从输入图像/视频中可控、高质量地合成新图像。神经渲染技术是多种多样的，它们对【场景外观的控制】、【所需的输入】、【产生的输出】以及【使用的网络结构】各不相同。一种典型的神经渲染方法将与特定场景条件（例如，视点、照明、布局等）相对应的图像作为输入，从中构建“神经”场景表示，并在新的场景属性下“渲染”该表示，以合成新的图像。所学习的场景表示不受简单的场景建模近似的限制，并且可以针对高质量的新颖图像进行优化。同时，神经渲染方法以【输入特征、场景表示和网络架构】的形式结合了经典图形的思想，使学习任务更容易，输出更可控。

我们沿着我们认为最重要的轴提出了神经渲染方法的分类：

- 控制：我们想要控制什么？我们如何在控制信号上调节渲染？
- CG 模块：使用了哪些计算机图形模块，它们是如何集成到神经渲染管线中的？
- 显式或隐式控制：该方法是对参数进行显式控制，还是通过显示我们期望得到的输出的示例来隐式控制？
- 多模态综合：该方法是否经过训练，在给定特定输入的情况下输出多个可选输出？
- 通用性：渲染方法是否适用于多个场景/对象？

在下文中将讨论这些内容，这些项用于对当前最先进的方法进行分类。

5 1　控制

神经渲染旨在在用户指定的场景条件下渲染高质量的图像。在一般情况下，这是一个开放的研究问题。相反，当前的方法解决了特定的子问题，如新的视图合成，在新的照明下重照明，以及在新的表情和姿势下设置人脸动画和身体动画。这些方法的不同之处在于如何将控制信号提供给网络。

一种策略是将场景参数作为输入直接传递到第一或中间网络层。相关策略是在输入图像的所有像素上平铺场景参数，或将它们连接到内部网络层的激活。另一种方法是依赖图像的空间结构，并采用图像到图像的翻译网络将“引导图像”或“调节图像”映射到输出。例如，这种方法可以学习从语义掩码映射到输出图像。我们在下面描述的另一个选项是使用控制参数作为图形层的输入。

5 2　Computer Graphics Modules

神经绘制的一个新兴趋势是将计算机图形学知识集成到网络设计中。因此，方法可能在嵌入系统的“经典”图形知识的水平上有所不同。例如，从场景参数到输出图像的直接映射不使用任何图形知识。

集成图形知识的一种简单方法是不可微的计算机图形模块。例如，这样的模块可以用于渲染场景的图像，并将其作为 dense conditioning 输入传递到网络。可以提供许多不同的通道作为网络输入，例如深度图、法线图、相机/世界空间位置图、反照率图、场景的漫反射渲染等等。这将问题转化为图像到图像的翻译任务，这是一个经过充分研究的环境，例如可以通过具有跳跃连接的深度条件生成模型来解决。

基于可微分的图形模块，可以将图形知识更深地集成到网络中。例如，这种可微分模块可以实现完整的计算机图形渲染器、3D 旋转或照明模型。这样的组件为网络添加了受物理启发的感应偏置，同时仍然允许通过反向传播进行端到端训练。这可以用于在网络结构中分析地强化关于世界的 truth，释放网络容量，并导致更好的泛化，特别是在只有有限的训练数据可用的情况下。

5 3 Explicit vs. Implicit Control

另一种对神经渲染方法进行分类的方法是通过控制的类型。

一些方法允许显式控制，即用户可以以语义上有意义的方式手动编辑场景参数。例如，当前的神经渲染方法允许对相机视点、场景照明、面部姿势和表情进行显式控制。

其他方法仅允许通过代表性样本的方式进行隐式控制。虽然它们可以从参考图像/视频中复制场景参数，但不能明确地操纵这些参数。这包括将人的头部运动从参考视频转移到目标人的方法，或重定全身运动的方法。

允许显式控制的方法需要带有图像/视频及其相应场景参数的训练数据集。另一方面，隐式控制通常需要较少的监督。这些方法可以在没有显式 3D 场景参数的情况下进行训练，只需要较弱的注释。例如，虽然需要密集的面部表现捕捉来训练具有面部再现显式控制的网络，但隐式控制可以通过仅在具有相应稀疏 2D 关键点的视频上进行训练来实现。

5 4 Multi-modal Synthesis

通常情况下，有几个不同的输出选项可供选择是有益的。例如，当仅控制场景参数的子集时，相对于其他场景参数，潜在地存在大的多模态输出空间。用户可以看到一个由几个项组成的库（视觉上看起来各不相同），而不是一个单一的输出。这样的库可以帮助用户更好地了解输出情况，并根据自己的喜好选择结果。

为了实现彼此显著不同的各种输出，网络或控制信号必须具有一些随机性或结构化方差。例如，变分自动编码器对具有内置可变性的过程进行建模，并可用于实现多模态合成。最新的例子是 Park 等人的，它展示了一种结合可变性并通过用户界面显示：给定相同的语义图，只需按下按钮即可生成截然不同的图像。

5 5 Generality

神经渲染方法的对象特异性不同，一些方法旨在训练一次通用模型，并将其应用于手头任务的所有实例（例如，如果该方法适用于人类头部，则其目标是适用于所有人类）；其他方法是特定实例的，例如人头示例中，这些网络将在一个人身上运行（在特定的位置穿着一套特定的衣服），并且必须为每个新的主题重新培训一个新的网络。对于许多任务，特定于对象的方法目前正在产生更高质量的结果，代价是每个对象实例的训练时间过长。对于实际应用来说，这样的训练时间令人望而却步。改进通用模型是一个悬而未决的问题，也是一个令人兴奋的研究方向。

六　神经渲染的应用

神经渲染有许多重要的用例，如【语义照片处理】、【新视图合成】、【重照明】、【自由视点视频】以及【面部和身体再现】。论文的表 1 概述了本综述中讨论的各种应用。对于每种应用都报告以下属性：

- Required Data: 系统所需的所有数据。这不包括导出的数据，例如自动计算的面部标志，而是可以被认为是一个人为了能够再现系统而必须获取的最小数据量。
- Network Inputs: 直接馈送到系统的学习部分的数据，即在反向传播期间梯度流过的系统部分。
- Network Outputs: 由系统的学习部分产生的一切。这是提供监督的管线的最后一部分。
- Possible values for Required Data, Network Inputs and Network Outputs: Images, Videos, Meshes, Noise, Text, Camera, Lighting, 2D Joint positions, Renders, Semantic labels, 2D Keypoints, volumE, teXtures, Depth (for images or video).
- Contents: 系统设计用于作为输入和输出处理的对象和环境的类型。可能的值：Head, Person, Room, outdoor Environment, Single object (of any category)。
- Controllable Parameters: 场景可以被修改的参数。可能值：Camera, Pose, Lighting, coloR, Texture, Semantics, Expression, speecH。
- Explicit control: 指的是向用户提供可解释参数的系统，这些参数在更改时会以可预测的方式影响生成的输出。可能值：uninterpretable or uncontrollable，interpretable controllable parameters。
- CG module: 是否需要经典图形学模块嵌入到系统。可能值：no CG module, Nondifferentiable CG module, Differentiable CG module。
- Generality: 是否训练一次以后能用于多个不同的输入实例，比如对于一个人的合成，是否需要对新的人进行重新训练。可能值：instance specific, general。
- Multi-modal synthesis: 如上所述，允许基于相同输入按需生成彼此显著不同的多个输出的系统。可能值：single output, on-demand multiple outputs。
- Temporal coherence: 指定在方法训练期间是否明确强制执行时序一致性。可能值：not enforced, enforced (e.g. in loss function)。

6 1　Semantic Photo Synthesis and Manipulation

语义照片合成和操作使得交互式图像编辑工具能够以语义上有意义的方式控制和修改照片的外观。开创性的作品《Image Analogies》使用基于 patch 的纹理合成，在给定语义布局和参考图像的情况下创建了新的纹理。这种基于单图像 patch 的方法能够进行图像重组、重定目标和修复，但它们不允许进行高级操作，如添加新对象或从头开始合成图像。

数据驱动的图形系统通过从大型照片集中检索的图像合成多个图像区域来创建新的图像。这些方法允许用户使用诸如手绘草图或语义标签图之类的输入来指定所需的场景布局。最新的开发是 OpenShapes，它通过匹配场景 context、shapes 和 parts 来组成区域。虽然获得了吸引人的结果，但这些系统在大型图像数据库中搜索时往往速度较慢。此外，由于不同图像之间的视觉不一致，有时会发现不期望的伪影。

6 1.1　Semantic Photo Synthesis

使用 GAN 直接把语义转换为图像的技术。

pix2pix 技术。

Cascaded refinement networks 学习由粗到细的生成器，用感知损失来训练，生成高分辨率结果。但是结果缺少细节。pix2pixHD 使用 cGAN 来生成高分辨率纹理，相比于 pix2pix，它包含类似于 CRN 的粗到

细生成器、多尺度 discriminators、多尺度特征匹配对象（类似于感知距离，但是使用自适应 discriminator 来抽取任务明确特征）。

值得注意的是，多尺度管线是视觉和图形领域已有数十年历史的方案，对于深度图像合成仍然非常有效。pix2pixHD 和 BicycleGAN 都可以在给定相同用户输入的情况下合成多种可能的输出，允许用户选择不同的风格。

随后的系统扩展到视频域，允许用户控制视频的语义。半参数系统结合了经典的数据驱动图像合成和前馈网络。

最近，GauGAN 使用 SPatially-Adaptive (DE)normalization layer（规范化层）(SPADE) 来更好地保存生成器中的语义信息。虽然先前的 conditional models 通过多个规范化层（例如，InstanceNorm）处理语义布局，但 channel-wise 规范化层倾向于“洗去”语义信息，尤其是对于统一和平坦的输入布局区域。相反，GauGAN 生成器采用随机潜在向量作为图像样式代码，并使用具有空间自适应归一化层（SPADE）的多个 ResNet 块来产生最终输出。这种设计不仅产生了视觉上吸引人的结果，而且使用户能够更好地控制样式和语义。自适应归一化层也被发现对程式化和超分辨率是有效的。

6 1.2 Semantic Image Manipulation

上述图像合成系统擅长于创建新的视觉内容，给定用户控件作为输入。然而，由于两个原因，使用深度生成模型对用户提供的图像进行语义图像操作仍然具有挑战性。

首先，编辑输入图像需要用生成器精确地重建它，即使在最近的 GANs 中这也是一项困难的任务。其次，一旦应用了控制，新合成的内容可能与输入的照片不兼容。为了解决这些问题，iGAN 提出使用无条件 GAN 作为图像编辑任务的自然图像先验。该方法首先优化低维潜在矢量，使得 GAN 可以忠实地再现输入照片。该重建方法将准牛顿优化与基于编码器的初始化相结合。然后，系统使用颜色、草图和扭曲工具修改生成图像的外观。为了呈现结果，他们使用引导图像滤波将编辑从生成的图像转移到原始照片。

神经照片编辑的后续工作使用 VAE-GAN 将图像编码为潜在向量，并通过混合修改后的内容和原始像素来生成输出。该系统允许对人脸进行语义编辑，例如添加胡须。几个作品将编码器与 generator 一起训练。他们部署第二编码器来预测额外的图像属性（例如，语义、3D 信息、面部属性），并允许用户修改这些属性。这种使用 GANs 作为深度图像先验的想法后来被用于图像修复和去模糊。

上述系统在具有单个对象或特定类别的低分辨率图像上工作良好，并且由于直接 GANs 的结果不够真实，因此经常需要后处理（例如，滤波和混合）。为了克服这些挑战，GANPaint 将预先训练的 GAN 模型应用于特定图像。学习到的 image-specific GAN 将从整个图像集合中学习到的先验知识与该特定图像的图像统计相结合。类似于先前的工作，该方法首先将输入图像投影到潜在向量中。从矢量重建接近输入，但缺少许多视觉细节。然后，该方法稍微改变网络的内部参数，以更精确地重建输入图像。在测试期间，GANPaint 根据用户输入修改 GAN 的中间表示。GANPaint 没有像 Deep image Prior 中那样在单个图像上训练随机初始化的 CNN，而是利用从预先训练的生成模型中学习的先验知识，并对每个输入图像进行微调。这使得能够以逼真的方式添加和删除某些对象。通过预训练学习分布先验，然后对有限的数据进行微调，对于许多一次性和少量合成场景是有用的。

6 1.3 Improving the Realism of Synthetic Renderings

上面讨论的方法使用深度生成模型来合成来自用户指定的语义布局的图像，或者以语义上有意义的方式修改给定的输入图像。

如前所述，计算机图形学中的渲染方法已经被开发用于根据场景规范生成照片级真实感图像的完全相同的目标。然而，计算机渲染图像的视觉质量取决于场景建模的保真度；使用低质量的场景模型和/或渲染方法会导致图像看起来明显是合成的。

Johnson 等人通过使用从大规模照片收藏中检索到的类似真实照片的内容来提高合成渲染的真实感，从而解决了这一问题。然而，这种方法受到数据库大小和匹配度量的简单性的限制。Bi 等人建议使用深度生成模型来完成这项任务。他们训练一个条件生成模型，将低质量的渲染图像（以及场景法线和漫反射率等辅助信息）转换为高质量的真实感图像。他们建议对反照率着色分解（而不是图像像素）进行这种转换，

以确保纹理得到保留。

Shrivastava 等人学习基于未标记的真实图像来提高人眼渲染的真实感，Mueller 等人对人手采用了类似的方法。

Hoffman 等人通过特征匹配扩展了 CycleGAN，以提高用于领域自适应的街景渲染的真实性。

类似地，Nalbach 等人建议使用深度卷积网络将着色缓冲区（如每像素位置、法线和材质参数）转换为复杂的着色效果（如环境遮挡、全局照明和景深），从而显著加快渲染过程。将粗略渲染与深度生成模型结合使用以生成高质量图像的想法也被用于面部编辑方法。

6 2 Novel View Synthesis for Objects and Scenes

新视图合成是在给定同一场景的固定图像集的情况下生成场景的新相机视角的问题。因此，新的视图合成方法处理以相机姿态为条件的图像和视频合成。新视图合成背后的关键挑战是在稀疏观测的情况下推断场景的 3D 结构，以及修复场景中被遮挡和看不见的部分。

在经典的计算机视觉中，基于图像的渲染（IBR）方法通常依赖于基于优化的多视图立体方法来重建场景几何结构，并将观测扭曲到新视图的坐标系中。然而，如果只有很少的观测可用，场景包含与视图相关的效果，或者新视角的很大一部分没有被观测覆盖，IBR 可能会失败，导致出现重影状伪影和孔洞。

已经提出了神经渲染方法来生成更高质量的结果。在基于神经图像的渲染中，IBR 管线中以前手工制作的部分被基于学习的方法取代或增强。

其他方法从观察中重建场景的学习表示，使用可微分渲染器端到端地学习。这使得能够在学习的特征空间中学习几何、外观和其他场景属性的先验。

这种基于神经场景表示的方法从在表示和渲染器上规定少量结构，到提出三维结构化表示，如 voxel grids of features，再到 explicit 3D disentanglement of voxels and texture，point clouds，multi-plane images，implicit functions，其使网络在图像形成和几何方面具有归纳偏差 (inductive biases)。

神经渲染方法在以前公开的挑战中取得了重大进展，例如生成视图相关效应或从极稀疏的观测中学习形状和外观的先验。尽管与经典方法相比，神经渲染显示出更好的结果，但它仍有局限性。也就是说，它们仅限于特定的用例，并且受训练数据的限制。尤其是与视图相关的效果（如反射）仍然具有挑战性。

6 2.1 Neural Image-based Rendering

Neural Image-based Rendering（N-IBR）是经典的基于图像的渲染和深度神经网络的混合，它用学习组件取代了手工制作的启发式算法。经典的 IBR 方法使用一组捕获的图像和代理几何体来创建新的图像，例如，从不同的角度看到的新图像。代理几何体用于将图像内容从捕获的图像重新投影到新的目标图像域。

在目标图像域中，来自源图像的投影被混合以合成最终图像。这个简化的过程只对【具有用足够数量的捕捉视图】重建的【精确几何体的漫反射对象】给出精确的结果。但是，由于视图相关效果、不完美的代理几何体或源图像过少，可能会出现重影、模糊、孔洞或接缝等伪影。

为了解决这些问题，N-IBR 方法用学习的混合函数或考虑视图相关影响的校正来取代经典 IBR 方法中常见的启发式方法。DeepBlending 提出了一种广义网络来预测投影源图像的混合权重，以便在目标图像空间中进行合成。它们在室内场景中显示出令人印象深刻的结果，与经典的 IBR 方法相比，混合伪影更少。

在图像引导的神经对象渲染中，通过名为 EffectsNet 的网络训练特定场景网络来预测视图相关效果。它用于从源图像中移除镜面反射高光，以生成仅漫反射 (diffuse-only) 的图像，这些图像可以投影到目标视图中，而无需复制源视图的高光。该 EffectsNet 以暹罗式的方式 (Siamese fashion) 同时在两个不同的视图上进行训练，从而导致多视图照片的一致性丢失。在目标视图中，重新应用新的视图相关效果，并使用类似 U-Net 的架构混合图像。因此，该方法在对象和小场景上演示了新视点合成，包括与视图相关的效果。

6 2.2 Neural Rerendering

神经重新渲染将经典的 3D 表示和渲染器与深度神经网络相结合，后者将经典渲染重新渲染为更完整、更逼真的视图。

与基于神经图像的渲染（N-IBR）相比，神经重新渲染在运行时不使用输入视角，而是依赖于深度神经网络来恢复丢失的细节。

Neural Rerendering in the Wild 使用神经重渲染来合成各种照明条件下旅游地标的逼真视图。作者将该问题描述为一个多模态图像合成问题，该问题以包含深度和颜色通道的渲染深度缓冲区以及外观代码作为输入，并输出场景的逼真视图。该系统使用“运动结构”和“多视图立体”从互联网照片中重建密集的彩色点云，并为每个输入照片将恢复的点云渲染到估计的相机中。使用真实照片对和相应的渲染深度缓冲区，多模式图像合成管线学习外观的隐式模型，该模型表示一天中的时间、天气条件和 3D 模型中不存在的其他特性。

为了防止模型合成行人或汽车等瞬态物体，作者建议用预期图像的语义标记来调节重渲染器。在推理时，这种语义标记可以被构造为省略任何这样的瞬态对象。Pitaluga 等人使用神经重渲染技术来反转运动重建中的结构，并强调运动三维重建中的组织的隐私风险，这些重建通常包含颜色和 SIFT 特征。作者展示了稀疏点云是如何反转的，并从中生成逼真的新视图。为了处理非常稀疏的输入，他们提出了一种可见性网络，该网络将点分类为可见或不可见，并使用来自 3D 重建的 ground truth 对应进行训练。

6 2.3 Novel View Synthesis with Multiplane Images

给定对象的稀疏输入视图集，Xu 等人还解决了从新视点渲染对象的问题。与以前使用在自然照明和小基线下捕获的图像的视图插值方法不同，它们旨在捕获场景的光传输，包括与视图相关的效果，如镜面反射。此外，他们试图通过在 large baselines 下捕获的稀疏图像集来实现这一点，以使捕获过程更加轻量级。他们在大约 60 度圆锥体中，在点照明下拍摄了六张场景图像 。并渲染该圆锥体内的任何新颖视点。输入图像用于构建与新视点对准的平面扫描体。该体由 3D 卷积神经网络处理，以重建场景深度和外观。为了处理由 large baselines 引起的遮挡，他们提出了预测注意力图，该图捕捉不同像素处输入视点的可见性。这些注意力图用于调整外观平面扫描体并删除不一致的内容。该网络是在综合渲染的数据上进行训练的，并在几何和外观上进行监督；在测试时，它能够合成具有高频光传输效果（如阴影和镜面反射）的真实场景的照片真实感结果。

DeepView 是一种在新视图下可视化光场的技术。视图合成基于多平面图像，这些图像是在给定稀疏输入视图集的情况下通过学习的梯度下降方法估计的。与基于图像的渲染类似，可以将图像平面扭曲为新视图，并将其前后渲染到目标图像中。

6 2.4 Neural Scene Representation and Rendering

虽然基于多平面图像和基于图像的渲染的神经渲染方法已经实现了一些令人印象深刻的结果，但它们将模型对场景的内部表示规定为点云、多平面图像或 meshes，并且不允许模型学习场景几何体和外观的最佳表示。因此，新视图合成中的一条最新路线是使用神经场景表示来构建模型：学习场景属性的基于特征的表示。

生成查询网络是一个用于学习场景的低维特征嵌入的框架，它明确地建模了由于观测不完整而导致的这种神经场景表示的随机性。场景由观察的集合表示，其中每个观察是图像及其相应相机姿势的元组。以一组上下文观测和目标相机姿态为条件，GQN 将在目标相机姿态下观测到的帧上的分布参数化，与上下文观测一致。在给定与上下文相同的场景的其他观测的情况下，通过最大化每个观测的对数似然来训练 GQN。给定单个场景的几个上下文观测，卷积编码器将它们中的每一个编码为低维潜在向量。通过求和将这些潜在向量聚合为单个表示 r。卷积长短期记忆网络（ConvLSTM）将潜在变量 z 上的自回归先验分布参数化。在每个时间步长，ConvLSTM 的隐藏状态被解码为表示采样观测的画布 u 的残差更新。为了使优化问题易于处理，GQN 在训练时使用近似后验。作者展示了 GQN 在新的视图合成、模拟机械臂的控制和迷宫环境的探索中学习场景的丰富特征表示的能力。GQN 的概率公式允许模型对不同的帧进行采样，这些帧都与上下文观测一致，例如，捕捉上下文观测中被遮挡的场景部分的不确定性。

6 2.5 Voxel-based Novel View Synthesis Methods

虽然非结构化神经场景表示是手工制作的场景表示的一种有吸引力的替代方案，但它们也有许多缺点。首先，他们忽略了场景的自然 3D 结构。因此，他们未能在有限训练数据的情况下发现多视图和透视几何。受几何深度学习最近进展的启发，出现了一系列神经渲染方法，它们建议将场景表示为体素网格，从而增强 3D 结构。

RenderNet 提出了一种卷积神经网络架构，该架构从明确表示为 3D 体素网格的场景实现可微分渲染。该模型针对每类对象进行再训练，并需要具有标记相机姿势的图像元组。RenderNet 支持新视图合成、纹理编辑、重照明和着色。使用相机姿势，首先将体素栅格变换为相机坐标。一组 3D 卷积提取 3D 特征。特征的三维体素网格通过称为“投影单元”的子网络转换为二维特征图。投影单元首先折叠三维特征体素网格的最后两个通道，然后通过 1×1 二维卷积减少通道数量。1×1 卷积可以访问沿着单个相机光线的所有特征，使它们能够执行典型经典渲染器的投影和可见性计算。最后，$2D$ 上卷积神经网络对 $2D$ 特征图进行上采样并计算最终输出。作者演示了 RenderNet 学习从低分辨率体素网格渲染高分辨率图像。RenderNet 可以进一步学习应用不同的纹理和着色器，从而启用场景重照明和操纵场景的新视图合成。他们进一步证明，RenderNet 可以用于通过迭代重建算法从单个图像中恢复场景的 $3D$ 体素网格表示，从而实现对该表示的后续操作。

DeepVoxels（19 年）能够联合重建场景的几何结构和外观，并随后进行新视图合成。DeepVoxels 是在特定场景上训练的，只给定图像及其外在和内在的相机参数——不需要明确的场景几何体。这是通过将场景表示为嵌入特征的笛卡尔 3D 网格，并结合使用多视图和投影几何算子显式实现图像形成的网络架构来实现的。首先从 2D 观测中提取特征。然后，通过沿着各自的相机射线复制 2D 特征来取消投影，并通过小型 3D U-net 将其集成到体素网格中。为了使用给定的摄影机外部和内部参数渲染场景，虚拟相机位于世界坐标中。使用固有相机参数，将体素栅格重新采样为规范视图体。为了对遮挡进行推理，作者提出了一个遮挡推理模块。遮挡模块被实现为 3D U-Net，其接收沿着相机射线的所有特征及其深度作为输入，并产生沿着射线的每个特征的可见性分数作为输出，其中沿着每个射线的分数加起来为一。然后，最终投影的特征被计算为沿着每条射线的特征的加权和。最后，使用小的 UNet 将得到的 2D 特征图转换为图像。作为遮挡推理模块的副作用，DeepVoxels 以无监督的方式生成深度图。该模型从一端到另一端是完全可微分的，并且仅通过在训练集上强制执行的 2D 重渲染损失来监督。该论文在几个具有挑战性的场景上展示了宽基线的新视图合成，包括合成场景和真实场景，并且大大优于不使用 3D 结构的基线。

Visual Object Networks (VONs)（18 年）是一种 3D 感知生成模型，用于将对象的外观与 disentangled 的 3D 表示进行合成。受经典渲染管线的启发，VON 将神经图像形成模型分解为视点、形状和纹理三个因素。该模型使用端到端对抗学习框架进行训练，该框架通过可微投影模块联合学习 2D 图像和 3D 形状的分布。在测试期间，VON 可以同时合成 3D 形状、其中间 $2.5D$ 深度表示和最终 $2D$ 图像。这种 $3D$ disentanglement 允许用户独立地操纵对象的形状、视点和纹理。

HoloGAN（19 年）建立在 RenderNet 的学习投影单元之上，以建立一个允许显式视点更改的无条件生成模型。它实现了一个显式仿射变换层，该层直接将视图操作应用于学习的 3D 特征。与 DeepVoxels 一样，网络学习 3D 特征空间，但通过用随机潜在向量 z 变换这些深层体素，引入了关于 3D 对象/场景的更多偏差。这样，可以以无监督的方式训练本地支持视点变化的无条件 GAN。值得注意的是，HoloGAN 既不需要姿态标签和固有的相机信息，也不需要对象的多个视图。

6 2.6 Implicit-function based Approaches

虽然 3D 体素网格已经证明，3D 结构化场景表示有利于多视图一致的场景建模，但它们的记忆需求随着空间分辨率的变化而呈立体比例变化，并且它们不能平滑地对表面进行参数化，这需要神经网络将形状的先验作为相邻体素的联合概率来学习。

因此，它们无法以足够的空间分辨率对大型场景进行参数化，并且迄今为止未能在场景中推广形状和外观，这将允许应用，例如仅从少量观测中重建场景几何体。在几何深度学习中，最近的工作通过将几何建模为神经网络的水平集来缓解这些问题。最近的神经渲染工作将这些方法推广到允许绘制全色图像。

除了通过隐式函数参数化表面几何体外，像素对齐隐式函数还通过隐式功能表示对象颜色。首先通过

卷积神经网络将图像编码为逐像素特征图。然后，完全连接的神经网络将特定像素位置的特征以及深度值 z 作为输入，并将深度分类为对象内部/外部。相同的体系结构用于对颜色进行编码。该模型是端到端训练的，通过图像和 3D 几何结构进行监督。作者演示了单镜头和多镜头 3D 重建以及穿着衣服的人类的新颖视图合成。

场景表示网络（SRN）在单个完全连接的神经网络 SRN 中对场景几何和外观进行编码，该网络将世界坐标映射到局部场景属性的特征表示。只在给定图像及其外在和内在相机参数的情况下，端到端地训练可微分的学习神经射线行进器，不需要地面实况形状信息。SRN 将 $(x; y; z)$ 世界坐标作为输入，并计算特征嵌入。为了渲染图像，通过可微的、学习的射线机将相机射线追踪到它们与场景几何体（如果有的话）的交点，该射线机基于 SRN 在当前交点估计处返回的特征来计算下一步的长度。然后在光线交点对 SRN 进行采样，为每个像素生成一个特征。该 2D 特征图通过每像素完全连接的网络被转换为图像。类似于 DeepSDF，SRN 通过用代码向量 z 表示每个场景来在同一类中的场景之间进行推广。代码向量 z 通过完全连接的神经网络（即所谓的超网络）映射到 SRN 的参数。超网络的参数与码向量和像素生成器的参数共同优化。作者演示了 ShapeNet 数据集中对象的几何结构和外观的单图像重建，以及多视图一致视图合成。由于它们的每像素公式，SRN 推广到完全看不见的相机姿势，如变焦或相机 roll。

6 3 Free Viewpoint Videos

Free Viewpoint Videos，也称为 Volumetric Performance Capture(体表演捕捉)，依靠多摄像头设置来获取表演者的 3D 形状和纹理。从 Tanco 和 Hilton 的早期工作开始（00 年），该主题就引起了研究界的极大兴趣，并通过 Collet 等人（15 年）的工作和 Dou 等人（16、17 年）的实时对应工作取得了令人信服的高质量结果。尽管做出了努力，但由于缺少高频细节（16 年）或纹理烘焙（15 年），这些系统缺乏真实感，无法在任意场景中对这些模型进行准确和令人信服的重照明。事实上，体表现捕捉方法缺乏与视图相关的效果（例如镜面高光）；此外，估计的几何图形中的缺陷通常会导致纹理图的模糊。最后，在许多现实世界中（例如头发、半透明材料），创建时间一致的 3D 模型（15 年）非常具有挑战性。

Guo 等人最近的一项关于人类行为捕捉的工作（19 年）通过将传统的基于图像的重照明方法（00 年）与高速精确深度传感的最新进展（18 年、18 年）相结合，克服了其中的许多局限性。特别是，该系统使用 58 台 12:4MP RGB 相机与 32 个 12:4MP 有源红外传感器相结合，以恢复非常精确的几何图形。在捕捉过程中，系统基于球面梯度照明交错两种不同的照明条件。这为体捕获管线产生了前所未有的真实感水平。尽管这些 3D 捕捉系统取得了稳步的进展和令人鼓舞的结果，但它们仍然面临着重要的挑战和局限。半透明和透明的物体无法轻易捕捉；即使使用高分辨率深度传感器，重建薄结构（例如头发）仍然是非常具有挑战性的。尽管如此，这些多视图设置为机器学习方法提供了基础，这些方法在很大程度上依赖于训练数据来合成任意视图和姿势中的高质量人类。

6 3.1 LookinGood with Neural Rerendering

Martin Brualla 等人的 LookinGood 系统（18 年）引入了神经重渲染的概念，用于捕捉人类演员的表演。该框架依赖于体性能捕获系统，该系统实时重建表演者。然后，可以使用已知的几何体从任意视点渲染这些模型。由于实时性的限制，重建质量受到许多伪影的影响，如深度缺失、低分辨率纹理和过平滑几何。Martin Brualla 等人建议添加“见证相机”，这是一种高分辨率 RGB 传感器（12MP），不用于捕捉系统，但可以为深度学习架构提供训练数据，以重新渲染几何管道的输出。作者表明，这可以实时为任意视点、姿势和主题提供高质量的重新渲染结果。手头的问题非常有趣，因为它同时和实时地处理去噪、绘画和超分辨率。为了解决这一问题，作者将任务引入到图像到图像的翻译问题（17 年）中，并引入了语义感知损失函数，该函数使用语义信息（在训练时可用）来增加属于图像显著区域的像素的权重，并在忽略背景贡献的情况下检索精确的轮廓。

6 3.2 Neural Volumes

神经体（19 年）解决了从多视图视频数据自动创建、渲染和动画化高质量对象模型的问题。该方法训练神经网络将多视图视频序列的帧编码为紧凑的潜在代码，该代码被解码为半透明体，该半透明体包含每

个 $(x; y; z)$ 位置的 RGB 和不透明度值。

该体是通过从摄影机通过体进行光线行进来渲染的，累积颜色和不透明度以形成输出图像和 alpha 蒙版。在 3D 而不是屏幕空间中表述这个问题有几个好处：视点插值得到了改进，因为对象必须表示为 3D 形状，并且该方法可以很容易地与传统的三角形网格渲染相结合。尽管使用了低分辨率体素网格（128^3），该方法还是通过引入学习的 warp field 来产生高质量的模型，该学习的 warp field 不仅有助于对场景的运动进行建模，而且通过使体素变形以更好地匹配场景的几何形状来减少块体素网格伪影，并且允许系统移动体素以更好地利用可用的体素分辨率。扭曲场被建模为仿射扭曲场的空间加权混合，其可以自然地对分段变形进行建模。凭借半透明的体表示，该方法可以仅从 2D 多视图视频中重建具有挑战性的物体，如移动的头发、模糊的玩具和烟雾，而不需要明确的跟踪。潜在空间编码通过生成新的潜在空间轨迹或通过将解码器调节为诸如头部姿势之类的一些信息来实现动画。

6 3.3 Free Viewpoint Videos from a Single Sensor

在训练和测试时多视图图像的可用性是自由视点系统成功的关键因素之一。然而，对于一个典型的消费者来说，这种捕捉技术仍然很难实现，因为他们最多只能拥有一个 RGBD 传感器，比如 Kinect。因此，一些研究（19 年）试图通过深度学习降低基础设施要求，使捕获技术能够通过消费者硬件访问。从一张图像中重建表演者与以看不见的姿势合成人类的主题非常相关（17-18 年）。与其他方法不同的是，Pandey 等人最近的工作（19 年）以看不见的姿态和任意的视角合成了表演者，模仿了体捕捉系统的行为。任务更具挑战性，因为它需要解开姿势、纹理、背景和视点的纠缠。Pandey 等人提出利用半参数模型来解决这个问题。特别是，他们假设用户的校准序列很短：例如，在系统启动之前，用户在相机前旋转。多个深度学习阶段学习将当前视点（包含正确的用户姿势和表情）与预先记录的校准图像（包含正确视点但错误的姿势和表达）相结合。鉴于所需基础设施的大幅减少，结果令人信服。

6 4 Learning to Relight

在新照明下对场景进行照片逼真渲染，这一过程被称为“重照明”，是许多图形应用程序的基本组成部分，包括合成、增强现实和视觉效果。完成这项任务的一种有效方法是使用基于图像的重照明方法，将在不同照明条件下拍摄的场景图像（也称为“反射场”）作为输入图像，并将它们组合在一起，在新的照明下渲染场景外观（00 年）。基于图像的重新照明可以产生高质量、逼真的结果，甚至已被用于好莱坞制作中的视觉效果。然而，这些方法需要使用昂贵的自定义硬件进行缓慢的数据采集，从而排除了这些方法在动态性能和“野外”捕获等设置中的适用性。最近的方法通过使用合成渲染或真实捕获的反射场数据来训练深度神经网络来解决这些局限性，该网络可以从少数图像中重新照明场景。

6 4.1 Deep Image-based Relighting from Sparse Samples

Xu（18 年）等人提出了一种基于图像的重照明方法，该方法可以从在学习的最佳光方向下捕获的五幅图像的稀疏集合中重照明场景。他们的方法使用深度卷积神经网络从这五幅图像中回归任意方向光下的重照明图像。传统的基于图像的重新照明方法依赖于照明的线性叠加特性，因此需要数十到数百张图像才能获得高质量的结果。相反，通过训练非线性神经重照明网络，该方法能够从稀疏图像中实现重照明。再照明质量取决于输入光方向，作者建议以端到端的方式将定制设计的采样网络与重照明网络相结合，以共同学习最佳输入光方向和重照明函数。整个系统在一个大型合成数据集上进行训练，该数据集由程序生成的形状组成，这些形状以复杂的、空间变化的反照率呈现。在测试时，该方法能够重新照亮真实场景，并再现复杂的高频照明效果，如镜面反射和投射阴影。

6 4.2 Multi-view Scene Relighting

给定在不受控制的自然光照下拍摄的大型户外场景的多个视图，Philip 等人（19 年）可以在新的户外照明（通过太阳位置和云量水平参数化）下渲染场景。输入视图用于重建场景的 3D 几何结构；这种几何形状是粗糙和错误的，直接重照明会产生较差的结果。作者建议使用这种几何结构来构建中间缓冲区法线、反射特征和 RGB 阴影图，作为辅助输入，以指导基于神经网络的重照明方法。该方法还使用阴影细

化网络来改进阴影的去除和添加，阴影是户外图像中的一个重要线索。虽然整个方法是在合成渲染的数据集上训练的，但它可以推广到真实场景，为应用程序产生高质量的结果，如从多个（或单个）图像创建延时效果，以及在传统的基于图像的渲染管线中重新照明场景。

6 4.3 Deep Reflectance Fields

深度反射场（19 年）提出了一种新的技术，通过从几个受试者在各种表情和视点下的 4D 反射场数据数据库中学习面部反射模型来重新照明人脸图像。使用学习的模型，可以仅使用在球形颜色梯度照明下记录的两个原始图像在任意照明环境中重新定位人脸（09 年）。该方法的高质量结果表明，颜色梯度图像包含估计全 4D 反射场所需的信息，包括镜面反射和高频细节。虽然在球形颜色梯度照明下捕捉图像仍然需要特殊的照明设置，但与之前需要数百张图像的方法相比，将捕捉要求减少到仅两种照明条件（00 年），使该技术能够应用于动态面部表现捕捉。

6 4.4 Single Image Portrait Relighting

重照明方法的一个特别有用的应用是改变在野外拍摄的肖像图像的照明，即在自然无约束照明下使用现成的（可能是手机）相机。虽然这种情况下的输入只是单个（不受控制的）图像，但最近的方法已经证明了使用深度神经网络的最先进的结果（19 年、19 年）。这些方法中的重新照明模型由深度神经网络组成，该深度神经网络已被训练为将单个 RGB 图像作为输入，并在任意用户指定的环境图下产生肖像图像的重新照明版本作为输出。此外，该模型还预测了当前照明条件的估计，在 Sun 等人（19 年）的情况下，可以在约 $160ms$ 的移动设备上运行。Sun 等人将目标照明表示为环境图，并使用捕获的反射场数据训练其网络。

另一方面，Zhou 等人对目标照明使用球面谐波表示，并使用通过使用传统的基于比例图像的方法重新照明单个人像图像创建的合成数据集来训练网络。这些方法没有明确的反向渲染步骤来估计几何形状和反射率（14 年，17 年，18 年），而是直接从输入图像和“目标”照明回归到最终的重照明图像。在这样做的过程中，他们绕过了传统面部重新照明方法中的朗伯反射率和低维形状空间等限制性假设，并能够推广到包括头发和配饰在内的全人像图像重新照明。

6 5 Facial Reenactment

面部再现旨在修改视点和照明之外的场景属性，例如通过生成新的头部姿势运动、面部表情或语音。早期的方法是基于经典的计算机图形学技术。虽然其中一些方法只允许隐式控制，即将面部表情从源序列重定向到目标序列（16 年），但也探索了显式控制（03 年）。这些方法通常涉及从输入重建 3D 人脸模型，然后编辑和渲染模型以合成编辑后的结果。神经渲染技术通过更好地处理不准确的三维重建和跟踪，以及更好的真实感外观渲染，克服了经典方法的局限性。早期的神经渲染方法，如 Kim 等人的方法（18 年），使用条件 GAN 来细化经典方法估计的输出。与经典技术相比，除了更逼真的照片效果外，神经渲染方法还允许控制头部姿势和面部表情（18 年、19 年）。大多数用于面部再现的神经渲染方法都是针对每个身份单独训练的。直到最近，才探索了在多个人上推广的方法（18 年，19 年）。

这个领域暂时没有了解（与人脸检测、身体姿态检测和重建的内容目前基本都没有涉猎过），所以暂时不整理综述内容。

6 5.1 Deep Video Portraits

6 5.2 Editing Video by Editing Text

6 5.3 Image Synthesis using Neural Textures

6 5.4 Neural Talking Head Models

6 5.5 Deep Appearance Models

6 6 Body Reenactment

这个领域暂时没有了解（与人脸检测、身体姿态检测和重建的内容目前基本都没有涉猎过），所以暂时不整理综述内容。

七 Open Challenges

正如这个综述所示，在过去几年中，神经渲染取得了重大进展，并对大量应用领域产生了重大影响。尽管如此，我们仍处于学习图像生成方法这一新范式的开端，这给我们留下了许多悬而未决的挑战，但也为进一步推进这一领域提供了令人难以置信的机会。在下文中，我们描述了开放式研究的问题，并提出了下一步的建议。

7 1 泛化

泛化：许多神经渲染方法首先都是基于对描绘单个人、对象或场景的一小组图像或特定视频的过拟合。这是基于学习的方法的最佳情况，因为必须学习的变化是有限的，但它也限制了泛化能力。在某种程度上，这些方法学会了在训练示例之间进行插值。正如任何机器学习方法一样，如果在训练样本范围之外的输入上进行测试，它们可能会失败。例如，对于看不见的姿势，习得的重演方法可能会失败。

尽管如此，神经渲染范式已经为测试和训练时的数据分布相似的许多应用提供了支持。实现更好泛化的一种解决方案是将故障案例明确地添加到训练语料库中，但这是以牺牲网络容量为代价的，并且所有故障案例可能都不是先验已知的。此外，如果必须控制许多场景参数，维度诅咒使捕捉所有潜在场景变得不可行。更糟糕的是，如果一个解决方案适用于任意的人，我们就无法实际地为所有潜在用户收集训练数据，即使我们可以，也不清楚这种训练是否会成功。因此，未来面临的重大挑战之一是对看不见的环境进行真正的概括。例如，已经采取了第一个成功的步骤来在对象类别中推广 3D 结构化神经场景表示。提高泛化能力的一种可能性是在网络中明确构建物理启发的归纳偏差。这样的感应偏置例如可以是可微分的相机模型或显式 3D 结构的潜在空间。这在分析上强化了网络结构中关于世界的真相，并释放了网络容量。总之，这可以实现更好的泛化，特别是在只有有限的训练数据可用的情况下。另一个有趣的方向是探索如何利用测试时的附加信息来提高泛化能力，例如一组校准图像或 memory bank。

7 2 Scalability

到目前为止，许多工作都集中在非常特定的应用程序上，这些应用程序在其所能处理的场景的复杂性和大小方面受到限制。例如，（重）渲染人脸的工作主要集中在处理短视频剪辑中的一个人。类似地，神经场景表示已经成功地表示了复杂度有限的单个对象或小环境。虽然网络泛化可能能够处理更大多样性的对象或简单场景，但还需要可扩展性来成功处理复杂、杂乱和大型场景，例如，使动态人群、城市或全球范围的场景能够得到有效处理。这种努力的一部分当然是软件工程和改进可用计算资源的使用，但允许神经渲染技术扩展的另一个可能方向是让网络对合成性进行推理。一个复杂的场景可以理解为其各部分的总和。为了有效地对这种直觉进行建模，网络必须能够将场景分割成多个对象，理解局部坐标系，并稳健地处理部分遮挡或缺失部分的观测结果。然而，合成性只是迈向可扩展神经渲染技术的一步，必须开发神经网络架构的其他改进和无监督学习策略的步骤。

7 3 Editability

传统的计算机图形管线不仅针对建模和渲染功能进行了优化，而且还允许手动或通过模拟编辑场景的各个方面。如今的神经渲染方法并不总是提供这种灵活性。将学习的参数与管线的传统部分（如神经纹理）相结合的那些技术当然允许编辑传统部分（即网格），但如何编辑学习的参数（即神经纹理）并不总是直观的。实现一种直观的方式来编辑基于抽象特征的表示似乎并不简单，但如何设置神经渲染架构以允许艺术家编辑尽可能多的管线部分当然值得考虑。此外，对网络输出的理解和推理也很重要。即使在某些情况下可能无法获得明确的控制，也可以对失败案例进行推理。

7 4 Multimodal Neural Scene Representations

本报告主要关注渲染应用程序，因此，我们讨论的大多数应用程序都围绕着使用图像和视频作为网络的输入和输出展开。其中一些应用程序还包含声音，例如，在编辑音频时，可以在视频剪辑中实现嘴唇同步（19 年）。然而，同时使用视觉和音频作为输入的网络可以学习处理附加输入模态的有用方法。类似地，

沉浸式虚拟和增强现实体验以及其他应用可能需要神经渲染算法的多模式输出，该算法结合了空间音频、触觉和触觉体验，或者可能是嗅觉信号。将神经渲染技术扩展到包括其他感官可能是未来研究的一个富有成果的方向。

八 Social Implications

在本文中，我们提出了多种神经渲染方法，具有不同的应用和目标领域。虽然一些应用程序大多是无可指责的，但其他应用程序虽然具有合法且极其有用的用例，但也可能以邪恶的方式使用（例如，会说话的头部合成）。图像和视频处理的方法与媒体本身一样古老，例如在电影行业中很常见。然而，神经渲染方法有可能降低进入门槛，使资源有限的非专家能够使用操纵技术。

尽管我们相信本文中讨论的所有方法都是出于善意开发的，并且确实有可能通过更好的沟通、内容创作和讲故事对世界产生积极影响，但我们决不能自满。重要的是要积极讨论并制定一项计划来限制滥用。我们认为，至关重要的是，合成的图像和视频要清楚地呈现为合成的。我们还认为，在分享由此产生的视频之前，必须获得内容所有者和/或表演者的任何修改许可。

此外，作为一个社区，我们必须继续开发取证、指纹识别和验证技术（数字和非数字），以识别被操纵的视频。这种保护措施将减少滥用的可能性，同时允许创造性地使用视频编辑技术。研究人员还必须在适当的时候采用负责任的披露方式，仔细考虑如何以及向谁发布新系统。在自然语言处理领域最近的一个例子中，作者采用了“分阶段发布”的方法，没有立即发布完整的模型，而是在一整年内发布越来越强大的实施版本。作者还与安全研究人员和政策制定者合作，允许尽早访问完整的模型。从这个例子中学习，我们认为研究人员必须将披露策略作为任何可能被滥用的系统的关键部分，而不是事后考虑。我们希望，图像和视频合成的反复演示将教会人们更批判性地思考他们消费的媒体，尤其是在没有来源证明的情况下。我们还希望，公布这些系统的细节可以传播人们对其内部工作的认识和知识，引发并促进对上述伪造检测、水印和验证系统的相关研究。

最后，我们认为，有必要进行强有力的公众对话，以制定一套适当的法规和法律，平衡滥用这些工具的风险与创造性、协商一致的用例的重要性。虽然本节中描述的大多数措施都涉及法律、政策和教育工作，但其中一项措施——媒体取证——是一项技术挑战。

如今，数字内容的完整性至关重要。可以使用主动保护方法（如数字签名和水印）或被动取证分析来验证图像的完整性。一个有趣的概念是“安全数码相机”，它不仅引入了水印，还存储了拍摄者的生物特征标识符。虽然文献中探讨了用于取证应用的水印，但相机制造商迄今未能在相机硬件中实现此类方法。因此，合成或操纵图像的自动被动检测变得越来越重要。有大量的数字媒体取证文献，分为特定操纵方法和独立操纵方法。操纵特定检测方法学习检测由特定操纵方法产生的伪影。FaceForensics++ 提供了一个由不同图像合成和操作方法组成的大规模数据集，适用于以监督的方式训练深度神经网络。它现在是检测面部操作的最大取证数据集，拥有超过 400 万张图像。此外，他们还表明，即使在不同级别的图像压缩下，他们也可以训练最先进的神经网络来实现高检测率。

类似地，Wang 等人编写了 photoshop 的脚本，以便稍后检测经过 photoshop 处理的人脸。这种特定于操作的检测方法的缺点是每个操作方法需要大规模的训练语料库。在取证转移中，作者提出了一种少数镜头学习方法。基于以前看不见的操纵方法的几个样本，即使没有大的（标记的）训练语料库，也可以实现高检测率。在没有操作方法（即“野外”操作）的知识或样本的情况下，需要独立于操作的方法。这些方法集中于图像的合理性。物理和统计信息必须在整个图像或视频中保持一致（例如，阴影或 JPEG 压缩）。这种一致性可以例如以基于 patch 的方式来确定，其中将一个 patch 与图像的另一部分进行比较。使用这样的策略，检测器可以仅使用不同图像的块作为负样本在真实数据上进行训练。

参考文献

[1] Tewari A, Fried O, Thies J, et al. State of the art on neural rendering[C]//Computer Graphics Forum. 2020, 39(2): 701-727.