# 1 引言

NeRF 是 2020 年 ECCV 论文。仅仅过去不到 2 年，关于 NeRF 的论文数量已经十分可观。相比于计算机视觉，尤其是相比于基于深度学习的计算机视觉，计算机图形学是比较困难、比较晦涩的。被深度学习席卷的计算机视觉任务数不胜数，但被深度学习席卷的计算机图形学任务仍然尚少。

由于 NeRF 及其众多 follow-up 工作在图形学中非常重要的渲染任务上给出了优秀的结果，可以预见未来用深度学习完成图形学任务的工作会快速增长。今年的 GIRAFFE 是 NeRF 的后续工作之一，它摘下 2021CVPR 的最佳论文奖对整个方向的繁荣都起到积极的推动作用。

本文希望讨论以下问题：

- NeRF 被提出的基础（2 前 NeRF 时代）；
- NeRF 是什么（3 NeRF！）；
- NeRF 的代表性 follow-up 工作（4 后 NeRF 时代）；
- 包含 NeRF 的更宽泛的研究方向 Neural Rendering 的简介（5 不止是 NeRF）。

# 2 前 NeRF 时代

## 2.1 传统图形学的渲染

本质上，NeRF 做的事情就是用深度学习完成了图形学中的 3D 渲染任务。那么我们提两个问题。

- 问题 1：3D 渲染是要干什么？

看 2 个比较官方的定义。

MIT 计算机图形学课程 EECS 6.837 对渲染（Rendering）的定义：

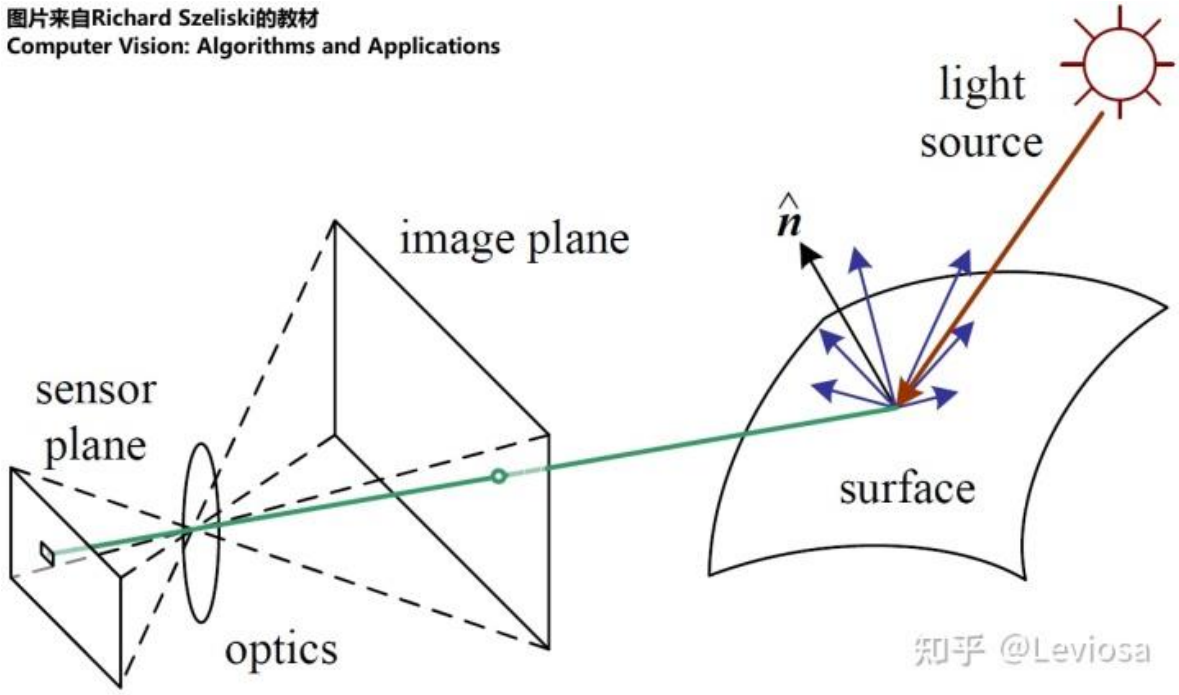
“Rendering” refers to the entire process that produces color values for pixels, given a 3D representation of the scene.

综述 State of the Art on Neural Rendering 对渲染（Rendering）的定义：

The process of transforming a scene definition including cameras, lights, surface geometry and material into a simulated camera image is known as rendering.

也就是说，渲染就是用计算机模拟照相机拍照，它们的结果都是生成一张照片。

用照相机拍照是一个现实世界的物理过程，主要是光学过程，拍照对象是现实世界中真实的万事万物，形成照片的机制主要就是：光经过镜头，到达传感器，被记录下来。



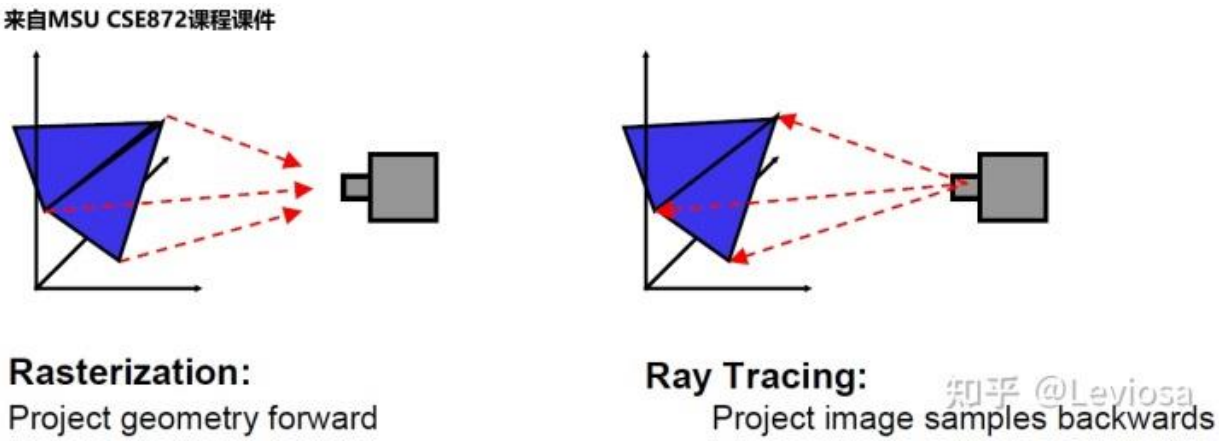

拍照的物理过程

而渲染就是**用计算机模拟这一过程**，模拟“拍照”的对象是已存在的某种三维场景表示（3D representation of the scene），模拟生成照片的机制是图形学研究人员精心设计的算法。

关键前提：渲染的前提是某种三维场景表示已经存在。渲染一词本身不包办生成三维场景表示。不过，渲染的确与三维场景表示的形式息息相关；因此研究渲染的工作通常包含对三维场景表示的探讨。

- 问题 2：3D 渲染是图形学问题，那么原先大家是用什么传统图形学方法实现 3D 渲染的呢？

主要有两种算法：光栅化（rasterization），光线追踪（ray tracing）；都是对照相机拍照的光学过程进行数学物理建模来实现的。

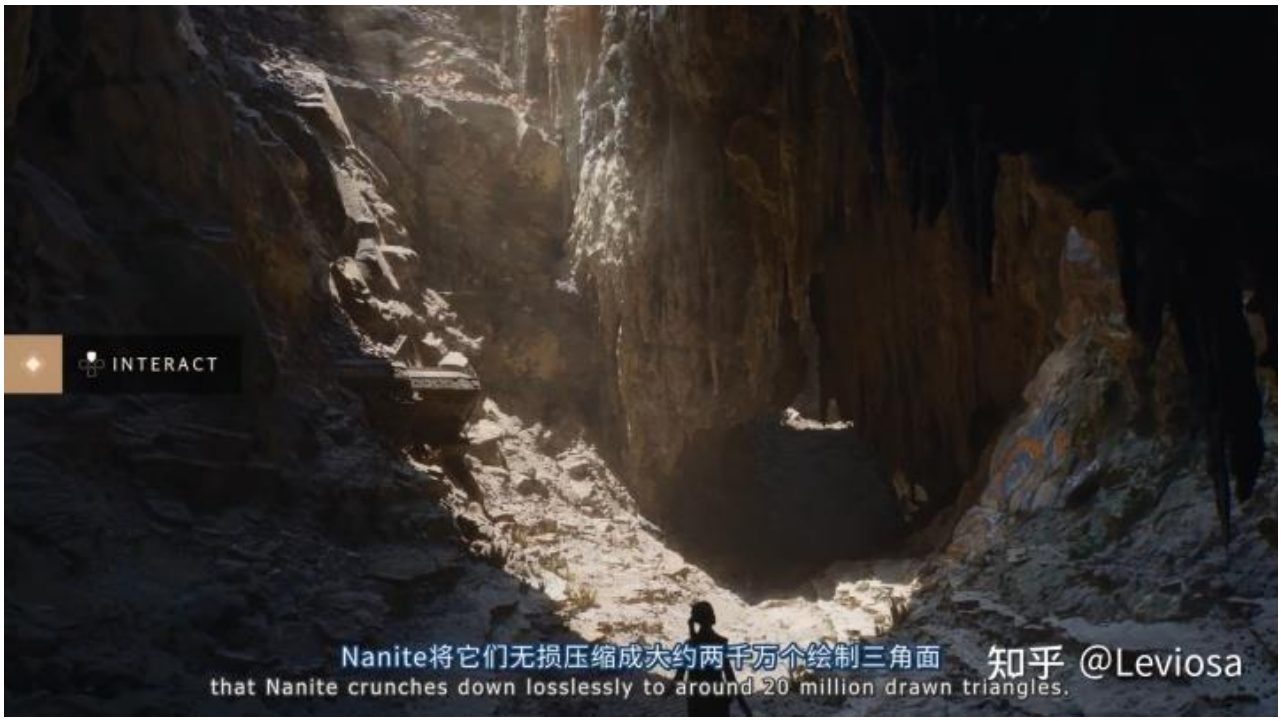



Rrasterization，Ray Tracing

传统渲染的详细原理参阅此教材。

光栅化是一种前馈过程，几何体被转换为图像域，是上世纪比较早的算法。光线追踪则是将光线从图像像素向后投射到虚拟三维场景中，并通过从与几何体的交点递归投射新光线来模拟反射和折射，有全局光照的优势（能模拟光线的多次反射或折射）。

当下，在学术界，还在研究传统图形学的渲染算法的人应该大部分在搞优化加速，怎么用 GPU 实时渲染更复杂的场景之类的事儿。在工业界，不少游戏重度依赖渲染技术，所以应该也有不少游戏公司在研究更逼真、更快速、更省算力的渲染算法。去年虚幻引擎出的新款“虚幻引擎 5”效果很是震撼，光照、纹理、流体的实时渲染模拟都逼真到了前所未有的新高度，可以看下虚幻引擎官方的宣传视频，真的很不错。



虚幻引擎 5

## 2.2 神经网络侵略 3D 渲染任务：NeRF 呼之欲出

### 隐式场景表示（implicit scene representation）

基于深度学习的渲染的先驱是使用神经网络隐式表示三维场景。 许多 3D-aware 的图像生成方法使用体素、网格、点云等形式表示三维场景，通常基于卷积架构。 而在 CVPR 2019 上，开始出现 使用神经网络拟合标量函数 来表示三维场景的工作。

### DeepSDF

2019 年 CVPR 的 DeepSDF 或许是最接近 NeRF 的先驱工作。

SDF 是 Signed Distance Function 的缩写。DeepSDF 通过回归（regress）一个分布来表达三维表面的。如下图所示，SDF>0 的地方，表示该点在三维表面外面；SDF<0 的地方，表示该点在三维表面里面。回归这一分布的神经网络是多层感知机（Multi-Layer Perceptron，MLP），非常简单原始的神经网络结构。

来自DeepSDF

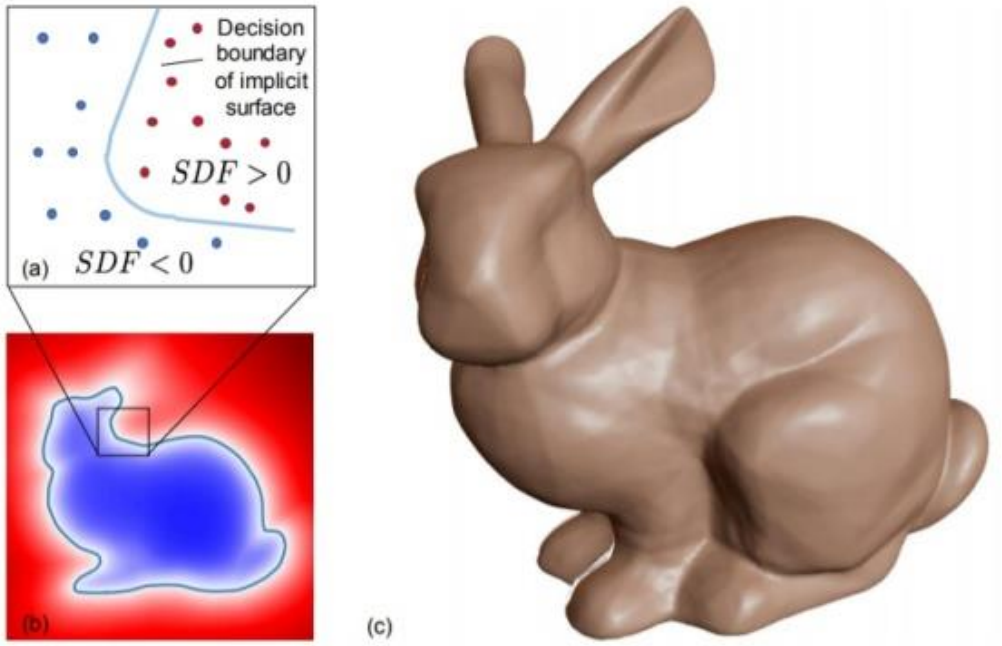



**Figure 2:** Our DeepSDF representation applied to the Stanford Bunny: (a) depiction of the underlying implicit surface $SDF = 0$ trained on sampled points inside $SDF < 0$ and outside $SDF > 0$ the surface, (b) 2D cross-section of the signed distance field, (c) rendered 3D surface recovered from $SDF = 0$. Note that (b) and (c) are recovered via DeepSDF.



DeepSDF

NeRF 比 DeepSDF 进步的地方就在于，NeRF 用 RGBσ 代替了 SDF，所以除了能推理一个点离物体表面的距离，还能推理 RGB 颜色和透明度，且颜色是 view-dependent 的（观察视角不同，同一物点的颜色不同），从而实现功能更强大的渲染。

# 3 NeRF!

建议前往 NeRF 项目网站查看视频效果图。

NeRF 效果图

## 3.1 Radiance Fields (RF)

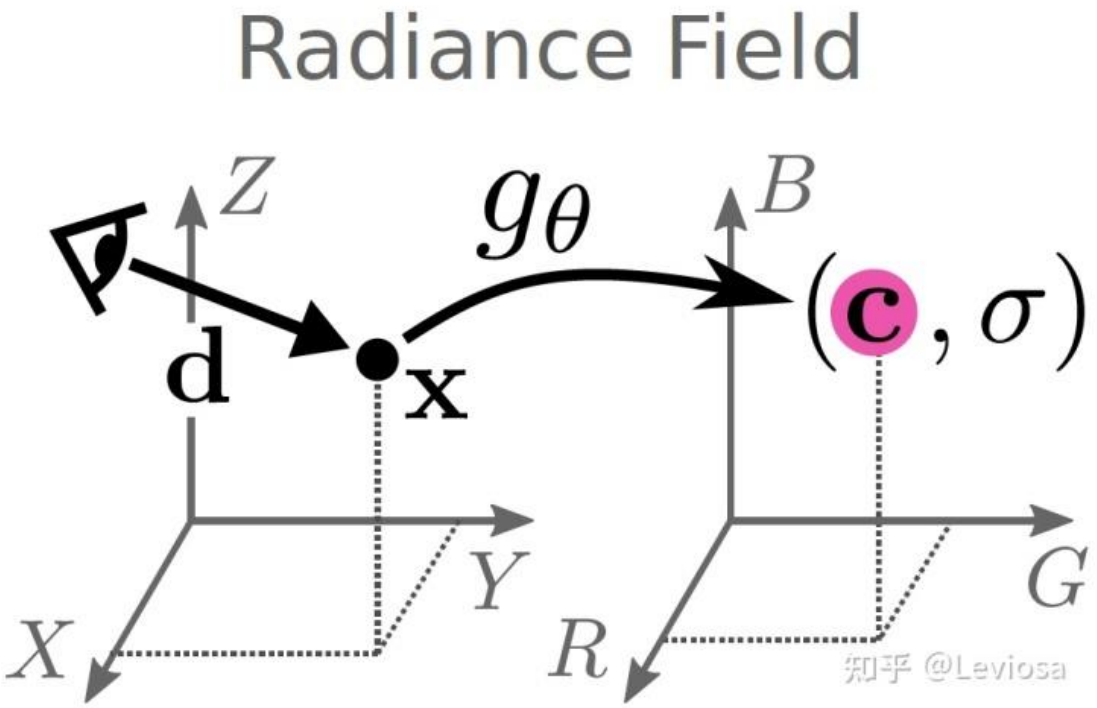
NeRF 是 Neural Radiance Fields 的缩写。其中的 Radiance Fields 是指一个函数、或者说映射 g θ g_\theta 。

(σ,c)=g θ (x,d)(σ,\bold{c})=g_ θ (\bold{x},\bold{d})

映射的输入是 x\bold{x} 和 d\bold{d} 。 x∈R3\bold{x}∈\mathbb{R}^3 是三维空间点的坐标， d∈S2\bold{d}∈\mathbb{S}^2 是观察角度。

映射的输出是 σ σ 和 c\bold{c} 。 σ ∈R+σ ∈\mathbb{R}^+ 是 volume density（可以简单理解为不透明度）， c∈R3\bold{c}∈\mathbb{R}^3 是 color，即 RGB 颜色值。

**来自GRAF论文Fig.1**




Radiance Fields（RF）

Radiance Fields，或者说映射 g θ g_\theta ，能对三维场景进行隐式表示（implicit scene representation）。在上一节，我们说过某种三维场景表示正是渲染的前提。实现渲染也是 作者提出 Radiance Fields 这一新型三维场景表示方法 的目的所在。

## 3.2 Neural Radiance Fields (NeRF)

Radiance Fields 是映射 g θ g_\theta 。那么 Neural Radiance Fields 则是指用神经网络拟合 Radiance Fields g θ g_\theta 。论文中，该神经网络具体是多层感知机（与 DeepSDF 一样）。

来自NeRF论文Fig.1

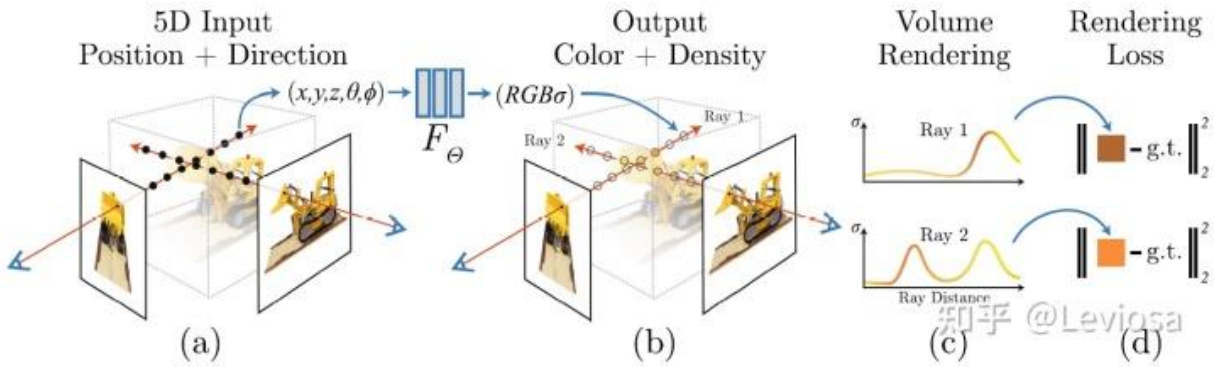

Neural Radiance Fields（NeRF）

## 3.3 NeRF 的体积渲染

NeRF（Neural Radiance Fields）其实是一种三维场景表示（scene representation），而且是一种隐式的场景表示（implicit scene representation），因为不能像 point cloud、mesh、voxel 一样直接看见一个三维模型。

NeRF 将场景表示为空间中任何点的 volume density σ σ 和颜色值 c\bold{c} 。 有了以 NeRF 形式存在的场景表示后，可以对该场景进行渲染，生成新视角的模拟图片。论文使用经典体积渲染（volume rendering）的原理，求解穿过场景的任何光线的颜色，从而渲染合成新的图像。

来自NeRF论文Fig.2




NeRF volume rendering

Volume density σ(x) σ(\bold{x}) 的严谨解释是：光射线在位置 x\bold{x} 处的无穷小粒子处终止的微分概率。 于是，具有近边界 tnt_n 、远边界 tft_f 的相机光线 r(t)=o+tdr(t) = \bold{o} + t\bold{d} 的颜色 C(r)C(\bold{r}) 是

C(r)=∫tntfT(t)σ(r(t))c(r(t),d)dtC(\bold{r})= \int^{t_f}_{t_n}{T(t)σ(\bold{r}(t))\bold{c}(\bold{r}(t),\bold{d})}{\rm d}t

*其中 T(t)T(t) 表示沿光线从 tnt_n 到 tt 累积透射率，也就是光线从 tnt_n 传播到 tt 而没有碰到任何其他粒子（仍存活）的概率。*

T(t)=exp(−∫tntσ(r(s))ds)T(t) = \rm{exp} ( -\int^{t}_{t_n} {σ(\bold{r}(s))}{\rm d}s )

那么，从 NeRF 渲染合成一张完整的图片，就需要为通过虚拟相机的每个像素的光线计算这个积分 C(r)C(\bold{r}) ，得到该像素的颜色值。

使用计算机求积分，必然是离散的采样，作者采用分层采样（stratified sampling）对这个连续积分进行数值估计。 算积分的具体细节见文尾附录。

### 3.4 NeRF 的训练

训练 NeRF 的输入数据是：从不同位置拍摄同一场景的图片，拍摄这些图片的相机位姿、相机内参，以及场景的范围。若图像数据集缺少相机参数真值，作者便使用经典 SfM 重建解决方案 COLMAP 估计了需要的参数，当作真值使用。

在训练使用 NeRF 渲染新图片的过程中，

- 先将这些位置输入 MLP 以产生 volume density 和 RGB 颜色值；
- 取不同的位置，使用体积渲染技术将这些值合成为一张完整的图像；
- 因为体积渲染函数是可微的，所以可以通过最小化上一步渲染合成的、真实图像之间的差来训练优化 NeRF 场景表示。

这样的一个 NeRF 训练完成后，就得到一个 以多层感知机的权重表示的 模型。一个模型只含有该场景的信息，不具有生成别的场景的图片的能力。

除此之外，NeRF 还有两个优化的 trick：

- 位置编码（positional encoding），类似于傅里叶变换，将低维输入映射到高维空间，提升网络捕捉高频信息的能力；
- 体积渲染的分层采样（hierarchical volume sampling），通过更高效的采样策略减小估算积分式的计算开销，加快训练速度。

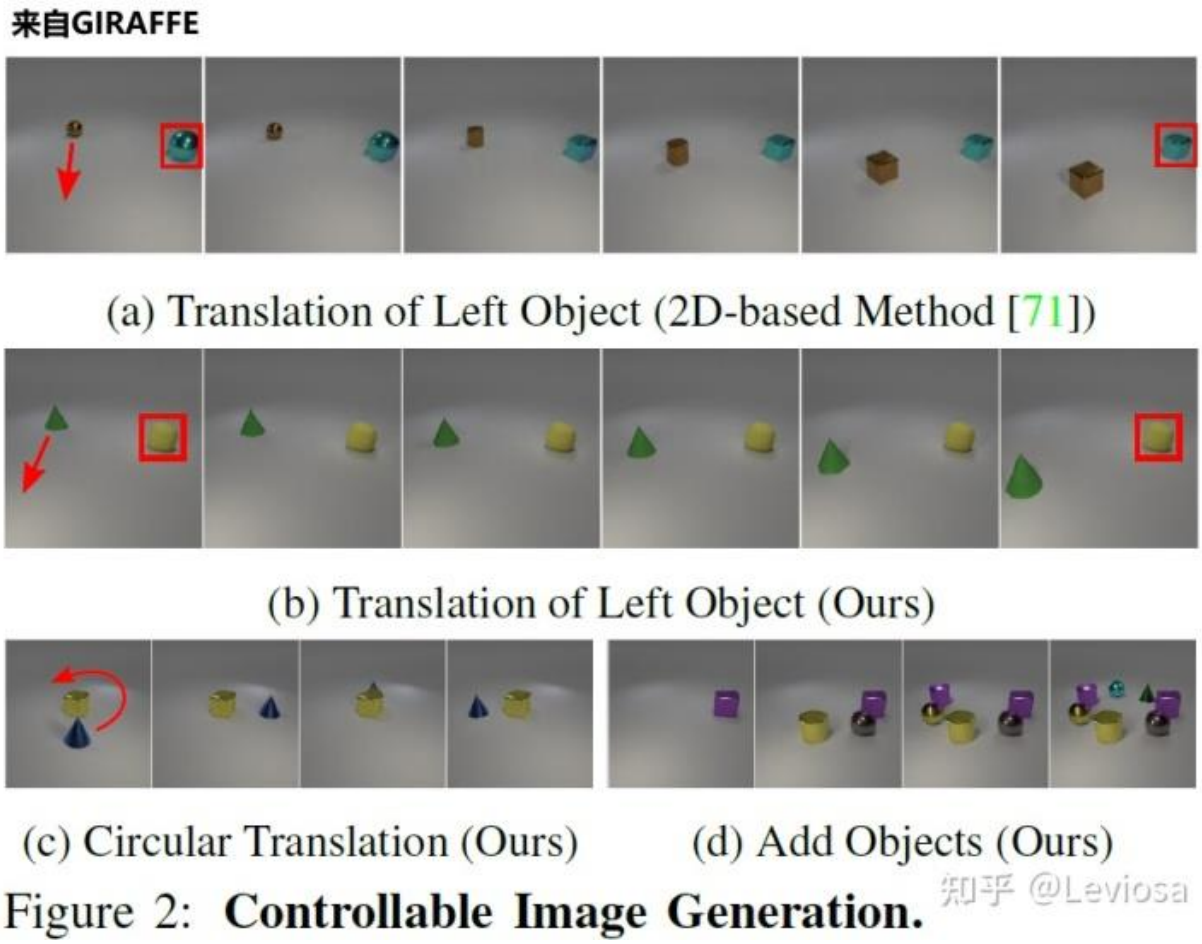
## 4 后 NeRF 时代

### GIRAFFE：composition 方向的代表作

2021CVPR 的最佳论文奖得主 GIRAFFE 是 NeRF、GRAF 工作的延申。

在 NeRF 之后，有人提出了 GRAF（Generative Radiance Fields），关键点在于引入了 GAN 来实现 Neural Radiance Fields；并使用 conditional GAN 实现对渲染内容的可控性。

在 GRAF 之后，GIRAFFE 实现了 composition。在 NeRF、GRAF 中，一个 Neural Radiance Fields 表示一个场景，one model per scene。而在 GIRAFFE 中，一个 Neural Radiance Fields 只表示一个物体，one object per scene（背景也算一个物体）。这样做的妙处在于可以随意组合不同场景的物体，可以改变同一场景中不同物体间的相对位置，渲染生成更多训练数据中没有的全新图像。



GIRAFFE 实现 composition

如图所示，GIRAFFE 可以平移、旋转场景中的物体，还可以在场景中增添原本没有的新物体。

另外，GIRAFFE 还可以改变物体的形状和外观，因为网络中加入了形状编码、外观编码变量（shape codes zsi\bold{z}^i_s ， appearance codes zai\bold{z}^i_a ）。

## 其他最新相关工作

2021 年 CVPR 还有许多相关的精彩工作发表。例如，提升网络的泛化性：

- pixelNeRF：将每个像素的特征向量而非像素本身作为输入，允许网络在不同场景的多视图图像上进行训练，学习场景先验，然后测试时直接接收一个或几个视图为输入合成新视图。
- IBRNet：学习一个适用于多种场景的通用视图插值函数，从而不用为每个新的场景都新学习一个模型才能渲染；且网络结构上用了另一个时髦的东西 Transformer。
- MVSNeRF：训练一个具有泛化性能的先验网络，在推理的时候只用 3 张输入图片就重建一个新的场景。

针对动态场景的 NeRF:

- Nerfies：多使用了一个多层感知机来拟合形变的 SE(3) field，从而建模帧间场景形变。
- D-NeRF：多使用了一个多层感知机来拟合场景形变的 displacement。
- Neural Scene Flow Fields：多提出了一个 scene flow fields 来描述时序的场景形变。

其他创新点：

- PhySG：用球状高斯函数模拟 BRDF（高级着色的上古神器）和环境光照，针对更复杂的光照环境，能处理非朗伯表面的反射。
- NeX：用 MPI（Multi-Plane Image ）代替 NeRF 的 RGBσ 作为网络的输出。

## 5 不止是 NeRF：Neural Rendering

Neural Radiance Fields 的外面是 Neural Rendering；换句话说，NeRF（Neural Radiance Fields）是 Neural Rendering 方向的子集。

在针对这个更宽泛的概念的综述 State of the Art on Neural Rendering 中，Neural Rendering 的主要研究方向被分为 5 类，NeRF 在其中应属于第 2 类“Novel View Synthesis”（不过这篇综述早于 NeRF 发表,表中没有 NeRF 条目）。

Neural Rendering 的 5 类主要研究方向

表中彩色字母缩写的含义：

- ***Network Inputs.*** The data that is directly fed into the learned part of the system, i.e., the part of the system through which the gradients flow during backpropagation.
- ***Network Outputs.*** Everything produced by the learned parts of the system. This is the last part of the pipeline in which supervision is provided.

  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.*** The types of objects and environments that the system is designed to handle as input and output. Possible values: Head, Person, Room, outdoor Environment, Single object (of any category).
- ***Controllable Parameters.*** The parameters of the scene that can be modified. Possible values: Camera, Pose, Lighting, coloR, Texture, Semantics, Expression, speecH.
- ***Explicit control.*** Refers to systems in which the user is given interpretable parameters that, when changed, influence the generated output in a predictable way. Possible values: ✗ uninterpretable or uncontrollable, ✓ interpretable controllable parameters.
- ***CG module.*** The level of "classical" graphics knowledge embedded in the system. Possible values: ✗ no CG module, Non-differentiable CG module, Differentiable CG module.



在这篇综述中，Neural Rendering 被定义为：

Deep image or video generation approaches that enable explicit or implicit control of scene properties such as illumination, camera parameters, pose, geometry, appearance, and semantic structure.

Neural Rendering 包含所有**使用神经网络生成可控（且 photo-realistic）的新图片的方法**。“可控”指人可以显式或隐式地控制生成新图片的属性，常见的属性包括：光照，相机内参，相机位姿（外参），几何关系，外观，语义分割结构。在这个大框架下，NeRF 是一种比较受欢迎的可控相机位姿的 Neural Rendering 算法。但 Neural Rendering 这个方向不止于此。

在目前的 Neural Rendering 方向，最火的子方向就是“Novel View Synthesis”，这与 NeRF 的强势蹿红密不可分；第二火的子方向是“Semantic Photo Synthesis”，这主要归功于语义分割以及相关的 GAN 领域的成熟度。“Semantic Photo Synthesis”方向也是成果颇丰，例如 2019 年 CVPR 的

Semantic Image Synthesis with Spatially-Adaptive Normalization，其效果图如下。

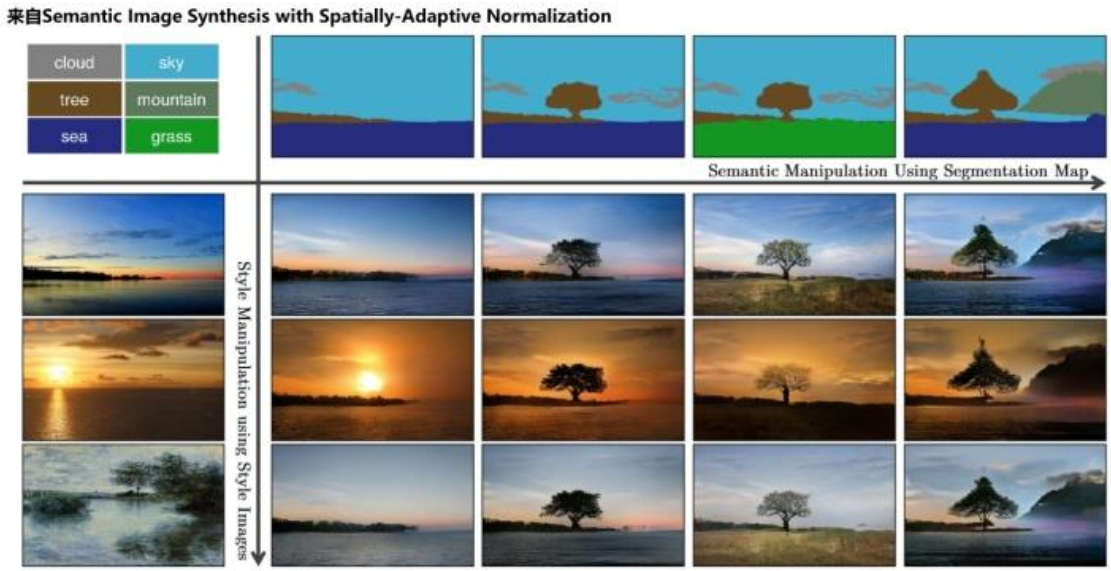



Figure 1: Our model allows user control over both semantic and style as synthesizing an image. The semantic (e.g., existence of a tree) is controlled via a label map (visualized in the top row), while the style is controlled via the reference style image (visualized in the leftmost column). Please visit our website for interactive image synthesis demos.

Semantic Image Synthesis