



三维计算机视觉感知与生成

段岳圻

duanyueqi@tsinghua.edu.cn



绪论

- 一、三维视觉相关定义
- 二、三维视觉任务与研究意义



□ 什么是人类视觉？

- 视觉是人体最重要的感官之一
- 约有70%以上的信息来源于视觉
- 人类视觉有以下几种重要能力
 - 空间感知：根据双眼视觉观测周围环境
 - 语义感知：物体识别和分类、场景理解、情境推断、情感识别
 - 想象创作：图像再现、场景建构、视觉旋转和变形、情景模拟



□ 什么是计算机视觉？

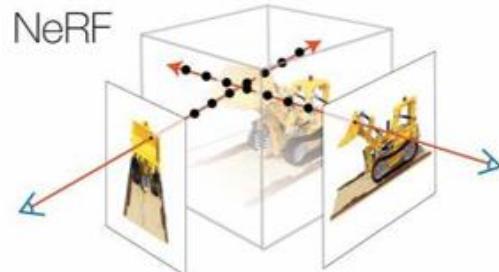
- 计算机视觉是一门用算法实现人类视觉能力，研究如何使计算机“看”和理解视觉世界的学科
- 计算机视觉主要包括图像处理、图像分析、模式识别、三维重建与立体视觉等



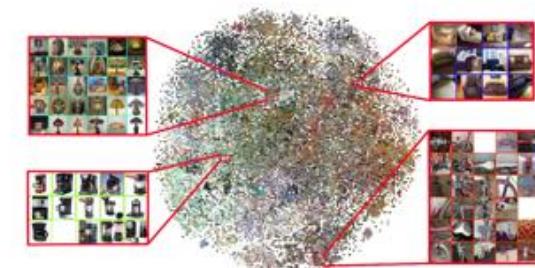
(a)



(c)



(b)



(d)

Natural Language Image Editing

IMAGE:  Prediction: IMAGE1 

Instruction: Hide Daniel Craig with 8) and Sean Connery with 1)

Program:

```
OB30=FaceDet(image=IMAGE)
OB31=Select(image=IMAGE, object=OB30, query='Daniel Craig', category=None)
IMAGE0=Emoji(image=IMAGE, object=OB31, emoji='smiling_face_with_sunglasses')
OB32=Select(image=IMAGE, object=OB30, query='Sean Connery', category=None)
IMAGE1=Emoji(image=IMAGE0, object=OB32, emoji='winking_face')
```

RESULT=IMAGE1

(e)

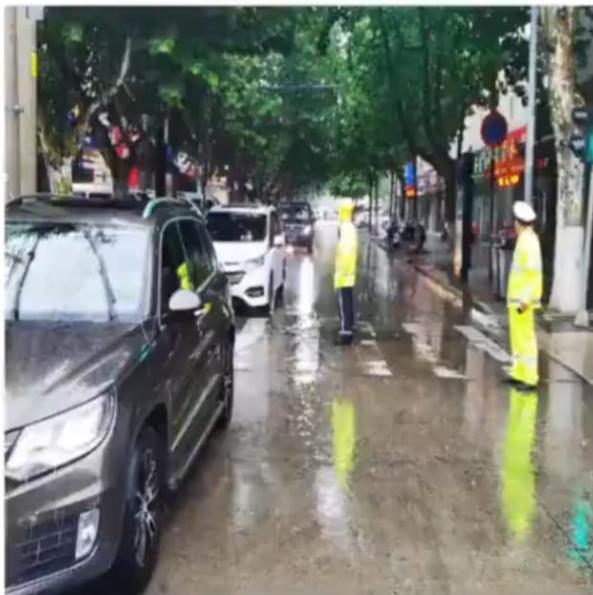


(f)



□ 什么是三维视觉？

- 三维视觉是计算机视觉的重要体现形式，是实现视觉智能不可或缺的手段
- 三维视觉基于图片、视频以及各类深度传感器信息，采用几何、统计以及优化等数学工具对现实世界进行三维测量、定位、建模及理解
- 我们将三维视觉分为：**三维重建、视觉感知和视觉生成**，分别对应人类视觉**空间感知、语义感知和想象创作**的能力



- 前面是否有车？有多少车？
- 前面是否有交警或行人？
- 前面是否有障碍物？
- 交警离我有多远？
- 两车之间的距离多少？
- 交警看到什么？想什么？
- 下一秒会发生什么？
-

二维语义理解
可以完成

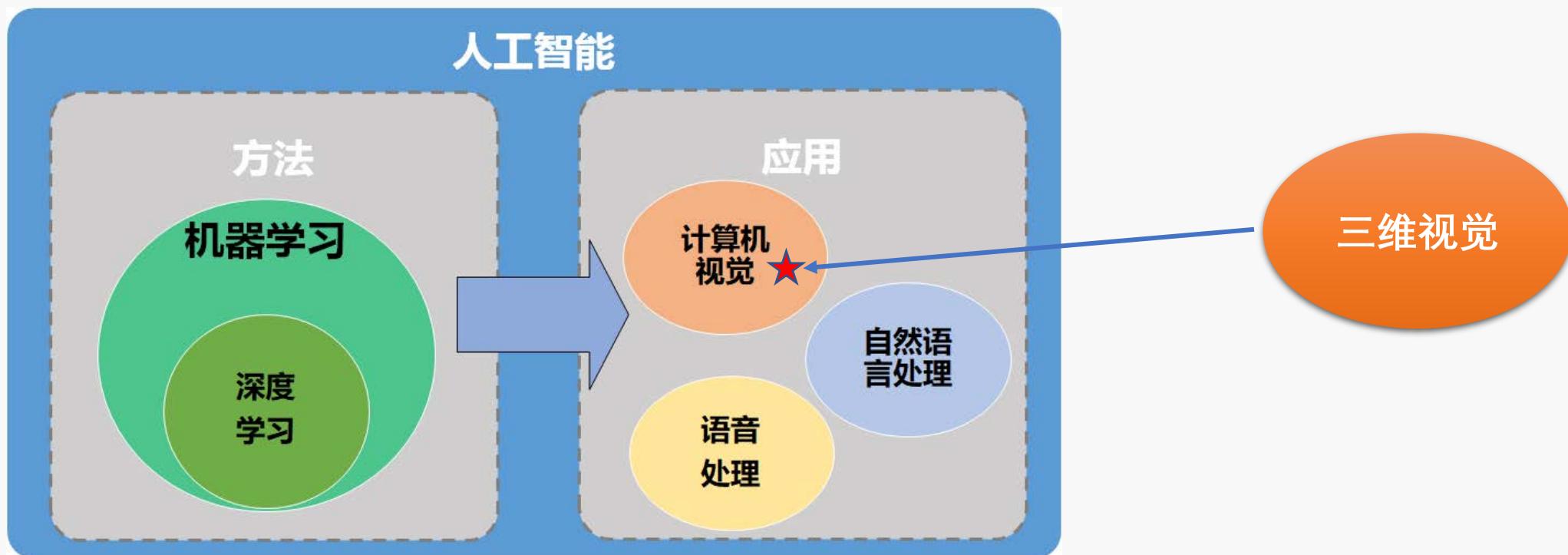
需要
三维视觉！



三维视觉相关定义

□ 计算机视觉和人工智能、机器学习、深度学习有什么区别和关系？

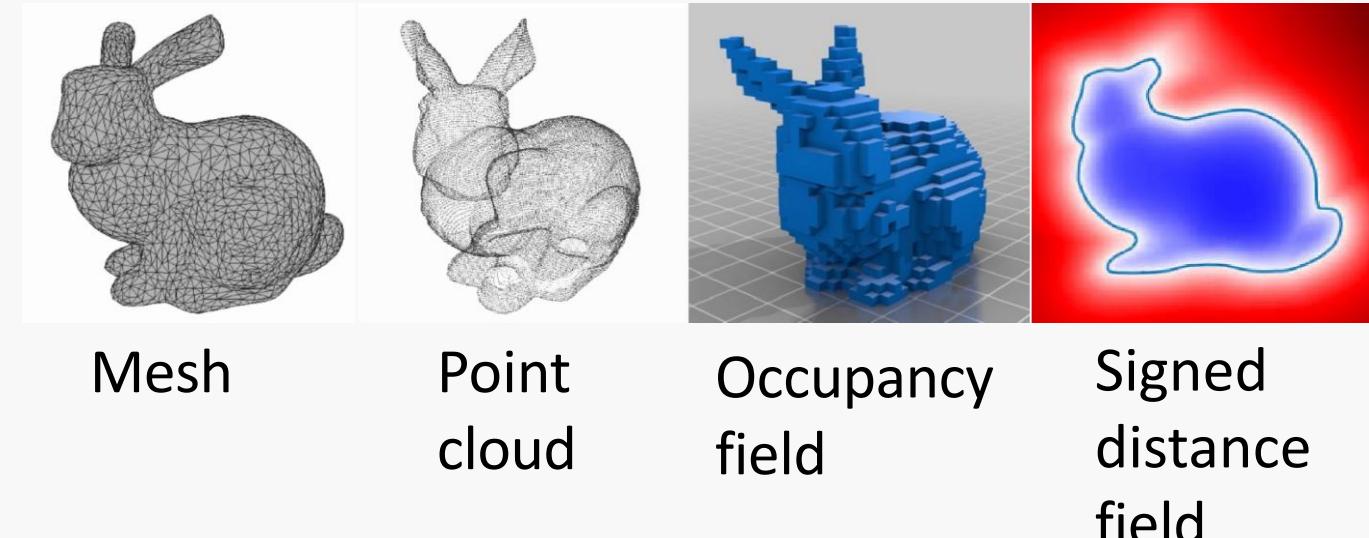
- 人工智能泛指能够达到人类智能水平的机器
- 机器学习和深度学习是实现人工智能的一种途径
- 计算机视觉是一类研究问题，目前主流方法是使用机器学习技术进行解决





□ 三维重建

- 三维重建是从二维图像或视频数据中恢复场景的三维结构和几何信息的过程
- 可以将场景表示为各种形式来进行三维重建



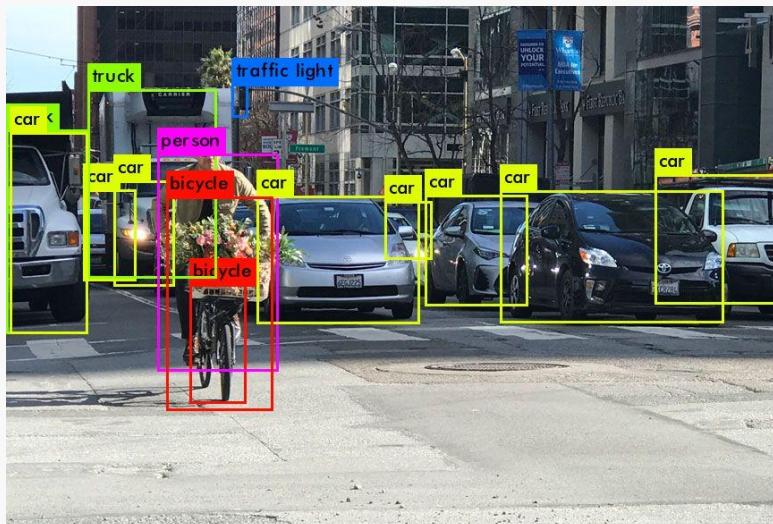
隐式辐射场



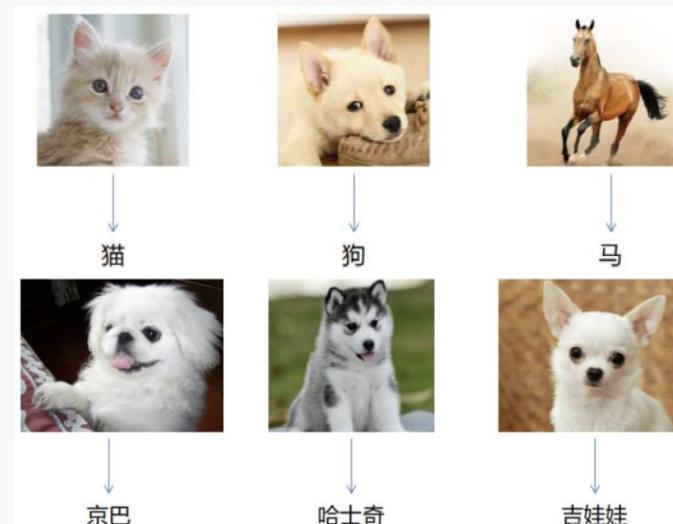
三维视觉任务

□ 二维视觉感知

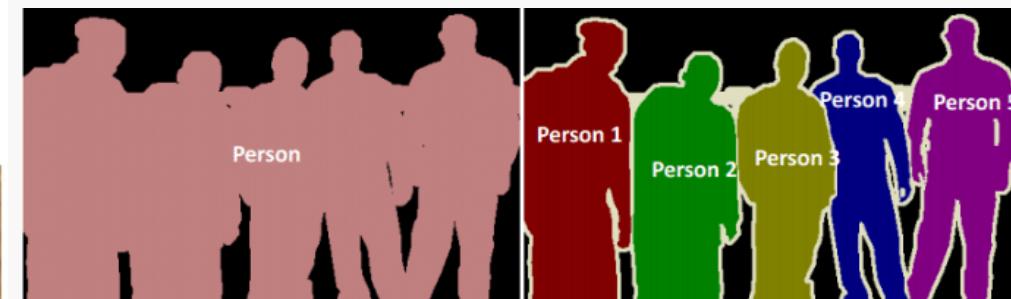
- 视觉感知指人类或计算机系统如何通过视觉信息来解释和理解周围世界
 - 在二维情境下，视觉感知任务主要集中在识别和理解图像中的对象、场景、人物等内容。主要包括目标检测、图像分类、语义分割等。



目标检测



图像分类



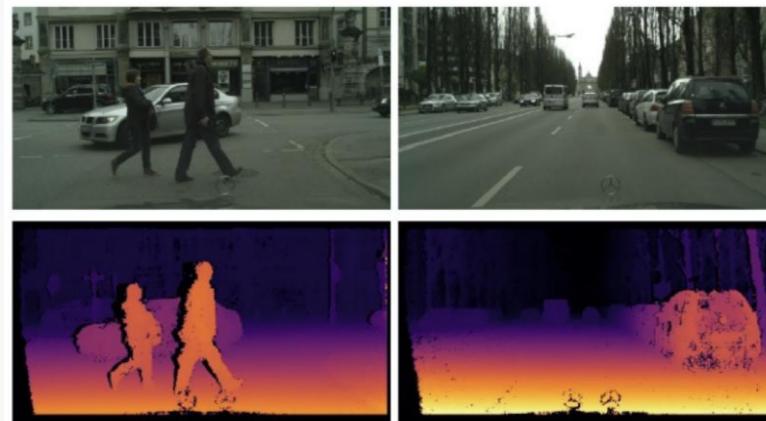
语义分割/实例分割



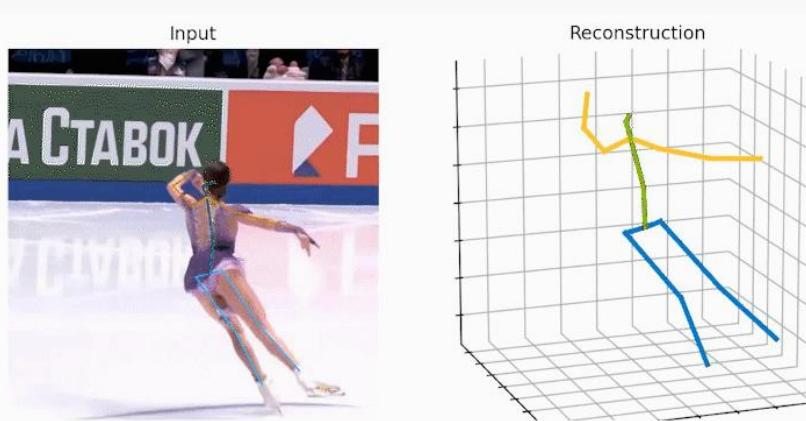
三维视觉任务

□ 三维视觉感知

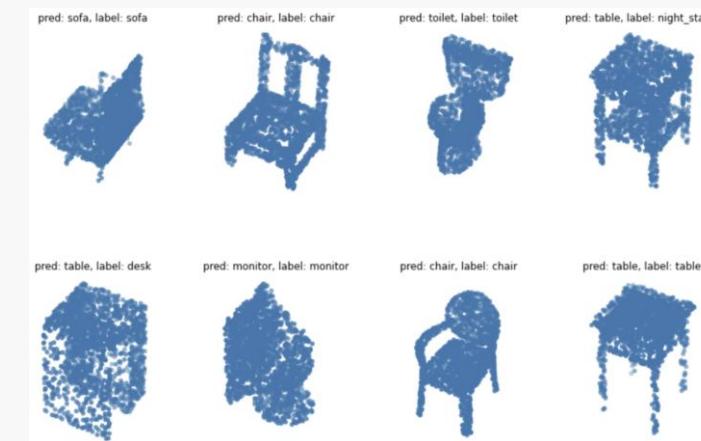
- 在三维环境中，视觉感知需要考虑物体的深度、立体感以及环境中的空间关系。例如，深度估计、点云分割、姿态估计等任务。



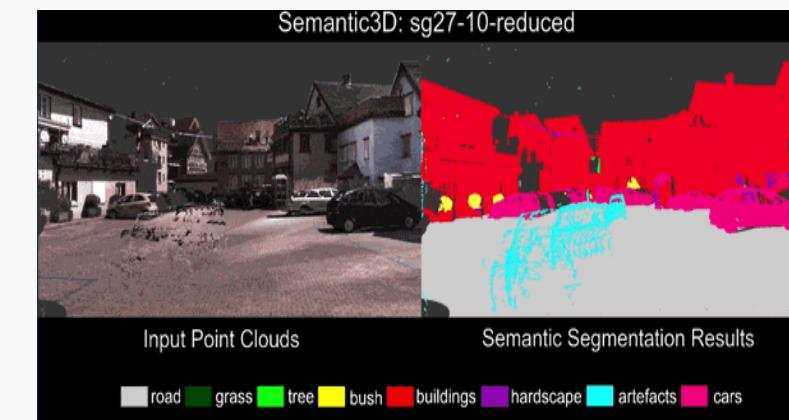
深度估计



姿态估计



点云分类



点云分割



□ 二维视觉生成

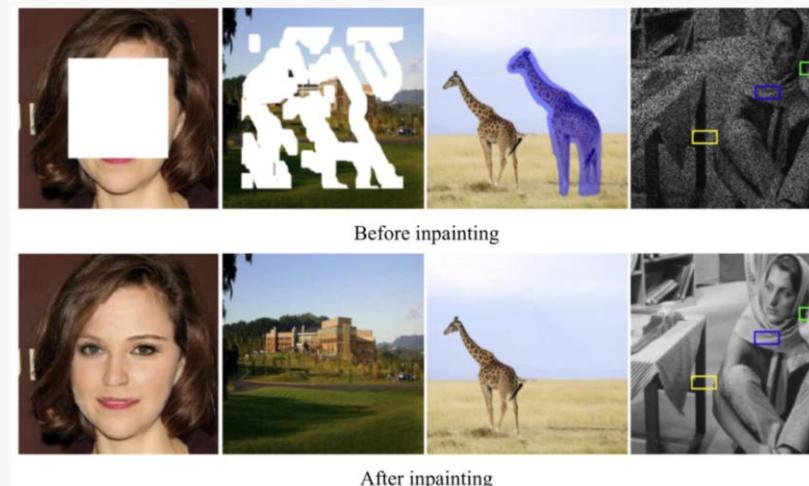
- 视觉生成是指利用计算机生成图像、视频或其他视觉内容的过程。
 - 在二维空间中，视觉生成通常指的是生成逼真的图像，如图像超分辨率、图像修复、图像风格迁移等任务。



低分辨率 (LR)

超分辨率
(SR)
→

高分辨率 (HR)



图片修复



风格迁移



三维视觉任务

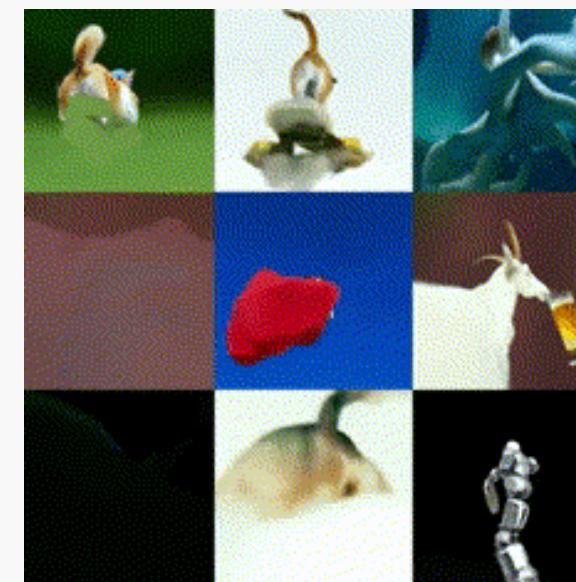
□ 三维视觉生成

- 在三维空间中，视觉生成可以涉及生成逼真的三维场景或物体模型，如三维物体生成、三维场景生成、三维点云生成等。

Input image 3D Model



三维物体生成



Text2Video

三维场景生成

GPT4Point

Does it have one body or multiple bodies?	The 3D object model is made up of multiple bodies , each with a different shape and size.	How many tails does it have?	It has two tails , one on each side.		
How many heads does the bird have?	There are two heads on the bird.	Is the frog one head?	No, the frog has two heads .		

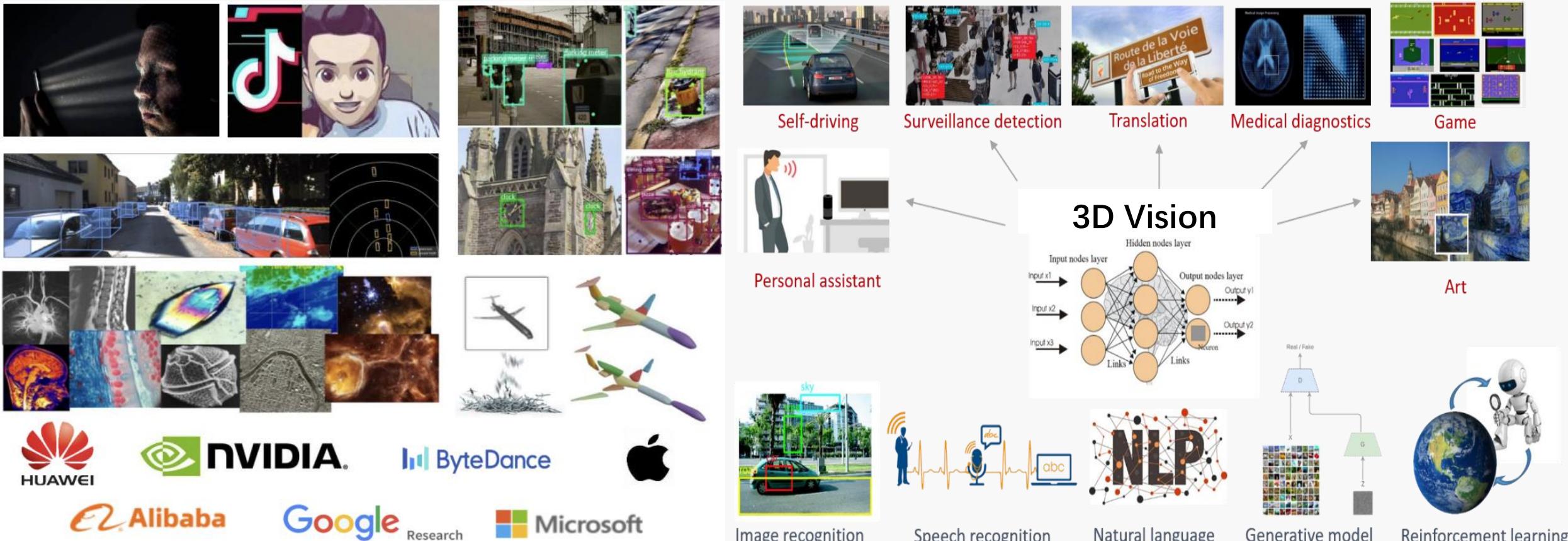
三维点云生成



三维视觉研究意义

□ 三维视觉应用广泛

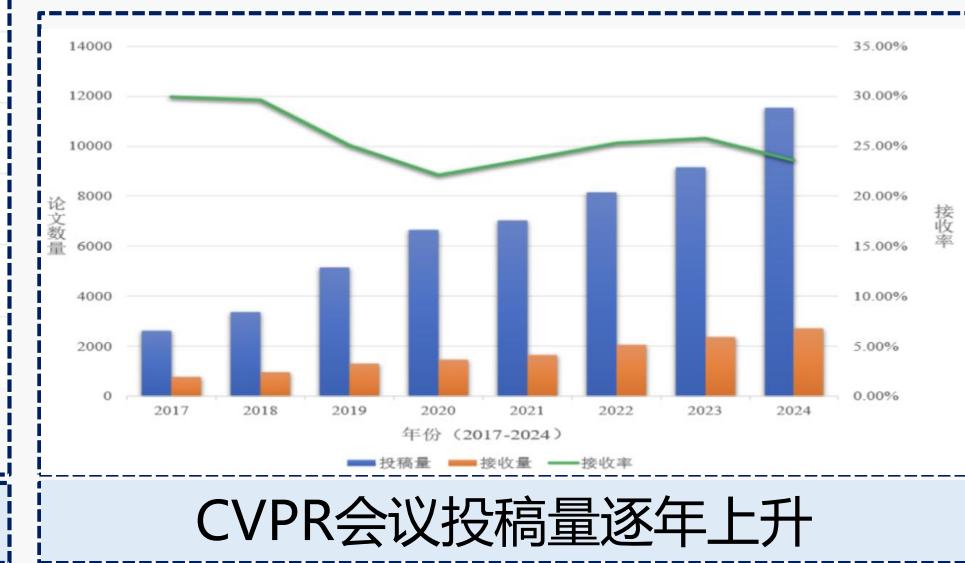
- 自动驾驶、机器人、医学影像、游戏开发、电影制作、VR/AR





三维视觉研究意义

□ 三维视觉是学术界发展最快的领域之一





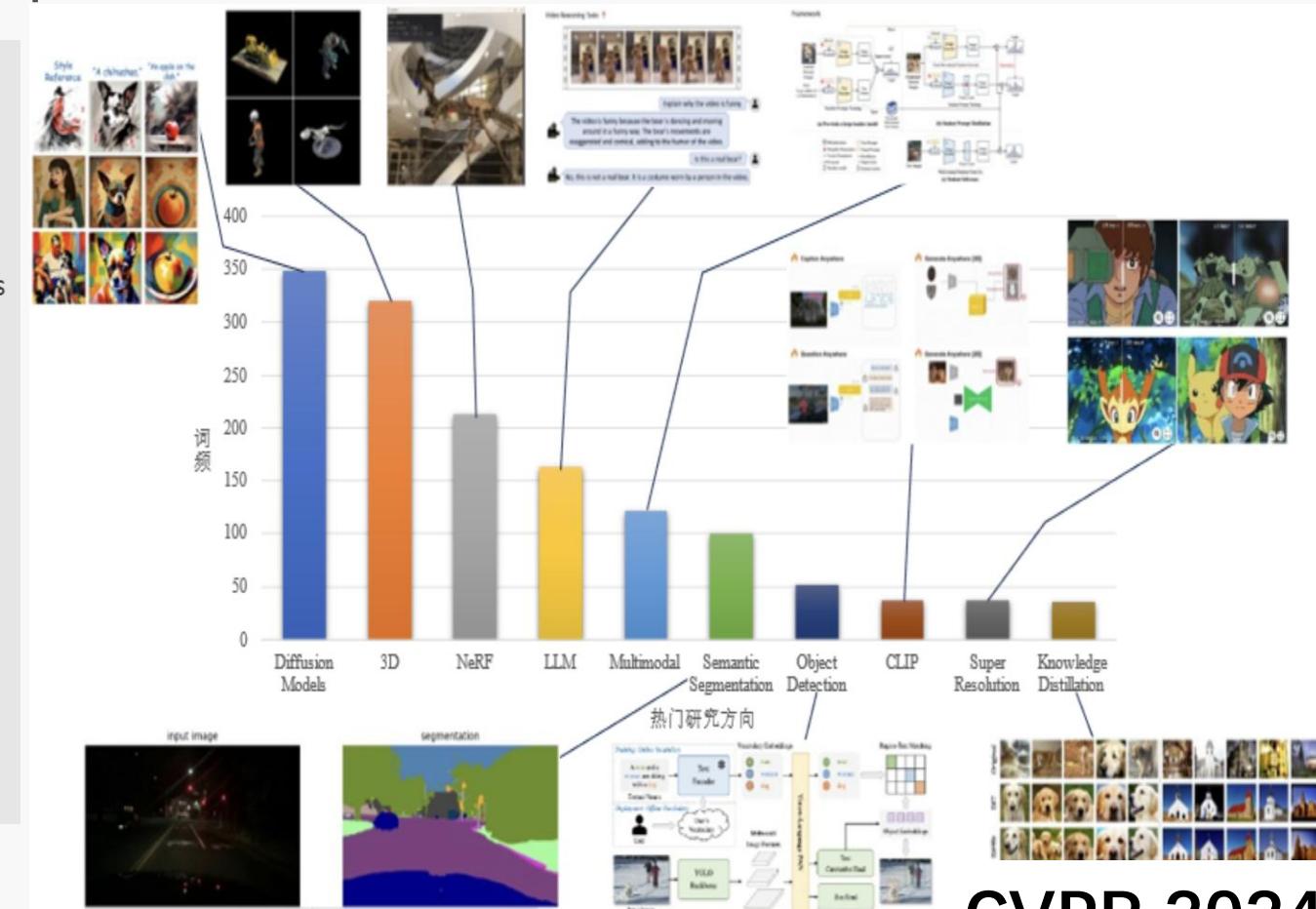
- 三维视觉是学术界发展最快的领域之一
 - 在CVPR中三维视觉占据半壁江山

CVPR 2023 by the Numbers

Selecting a category below changes the paper list on the right.

SELECT ↓ Top 10 overall by number of authors

	AUTHORS	PAPERS
1	3D from multi-view and sensors	1,090 246
2	Image and video synthesis and generation	889 185
3	Humans: Face, body, pose, gesture, movement	813 166
4	Transfer, meta, low-shot, continual, or long-tail learning	688 153
5	Recognition: Categorization, detection, retrieval	673 139
6	Vision, language, and reasoning	631 118
7	Low-level vision	553 126
8	Segmentation, grouping and shape analysis	524 113
9	Deep learning architectures and techniques	485 92
10	Multi-modal learning	450 89
11	3D from single images	431 91
12	Medical and biological vision, cell microscopy	420 53
13	Video: Action and event understanding	373 83
14	Autonomous driving	359 69
15	Self-supervised or unsupervised representation learning	349 71





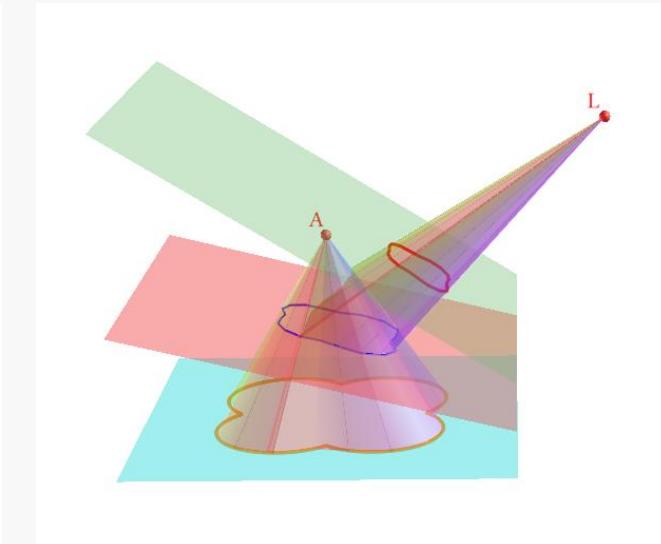
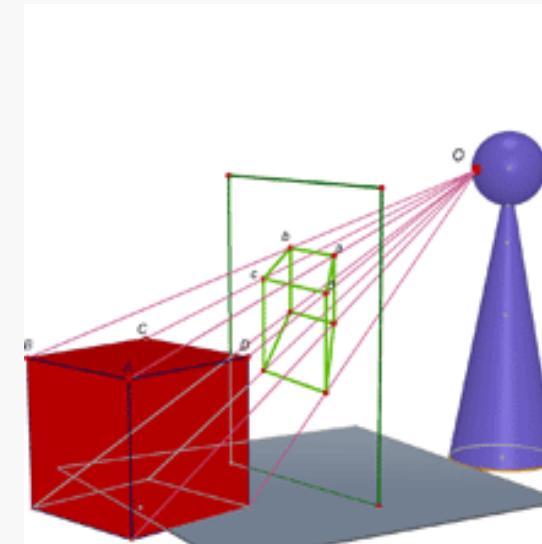
射影几何与变换

- 一、2D射影几何与变换
- 二、3D射影几何与变换



□ 引言

- 获取视觉信息依赖摄像机，摄像机是3D到2D的空间映射，成像过程包含2D、3D射影变换
- 射影变换：有限次中心射影的积定义的两条直线/平面/空间之间的一一对应变换，分别为1D/2D/3D射影变换
- 射影几何：研究图形射影性质的几何学分支学科

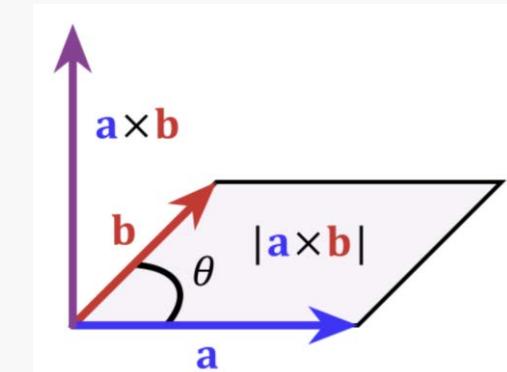




□ 点和直线的表示

- 点的齐次表示：点 $(\frac{x_1}{x_3}, \frac{x_2}{x_3})$ 表示为 $x = (x_1, x_2, x_3)^T$
- 直线的齐次表示：直线 $ax + by + c = 0$ 表示为 $I = (a, b, c)^T$
- 向量等价类： kx 和 x 表示同一个点， kI 和 I 表示同一条直线

关系	公式
点在直线上	$x^T I = x \cdot I = 0$
过直线的交点	$x = I_1 \times I_2$
过两点的直线	$I = x_1 \times x_2$



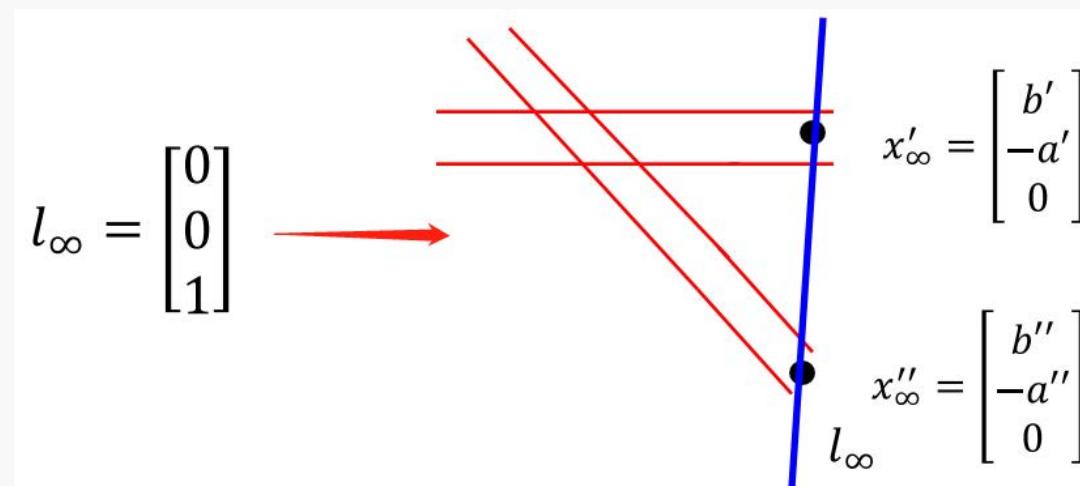
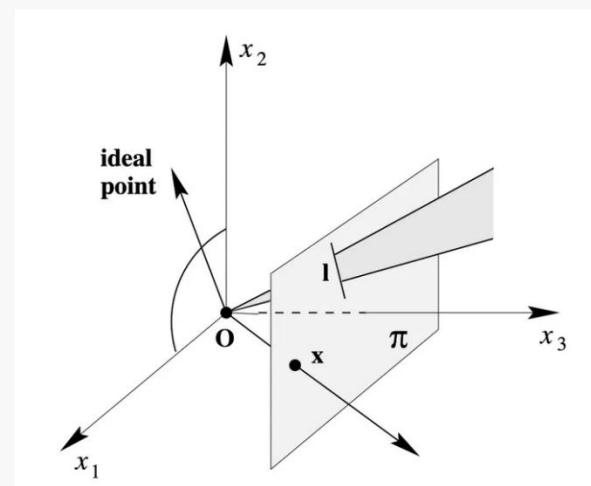
- 无穷远点（理想点）： $x = (x_1, x_2, 0)^T$ ，是平行线 $I_1 = (a, b, c_1)^T$ 和 $I = (a, b, c_2)^T$ 的交点 $(ax_1 + bx_2 = 0, c_1 \neq c_2)$
- 无穷远直线： $I_\infty = (0, 0, 1)^T$ ，是所有无穷远点组成的直线



□ 射影平面模型

■ IP^2 与 IR^3 的对应关系

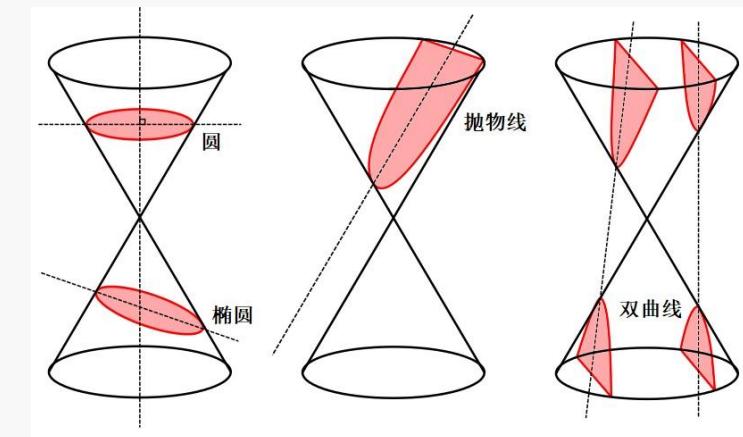
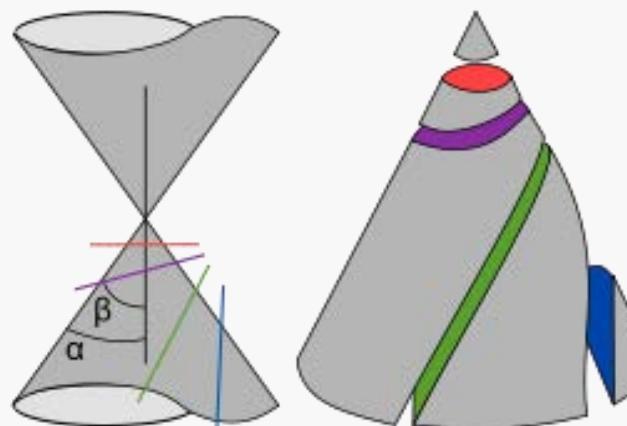
$IP^2 (x_1, x_2, x_3)$	$IR^3 (x_1, x_2, x_3)$
点	过原点的射线
直线	过原点的平面
理想点	x_1x_2 平面上的射线
无穷远直线	x_1x_2 平面





□ 二次曲线

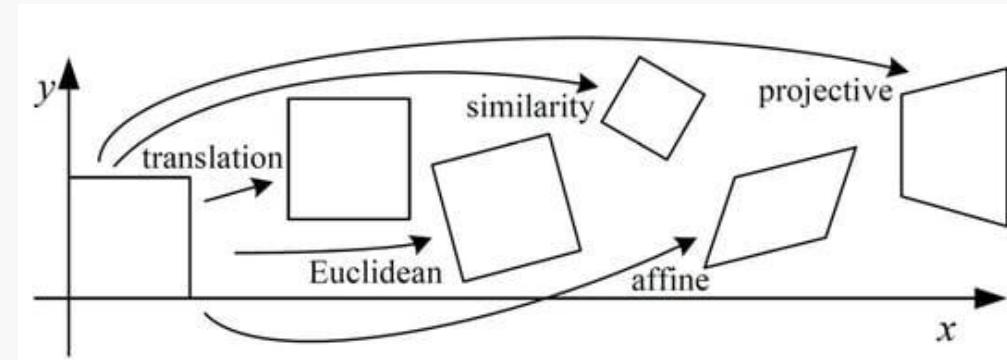
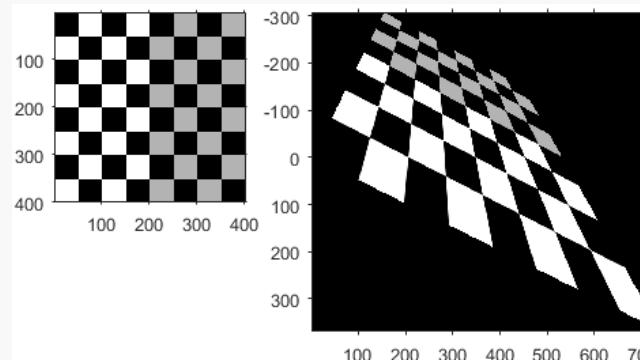
- 二次曲线非齐次方程: $ax^2 + bxy + cy^2 + dx + ey + f = 0$
- 二次曲线齐次方程: $ax_1^2 + bx_1x_2 + cx_2^2 + dx_1x_3 + ex_2x_3 + fx_3^2 = 0$
- 矩阵形式齐次方程: $x^T C x = 0$, 系数矩阵 $C = \begin{bmatrix} a & \frac{b}{2} & \frac{d}{2} \\ \frac{b}{2} & c & \frac{e}{2} \\ \frac{d}{2} & \frac{e}{2} & f \end{bmatrix}$
- $C = (a, b, c, d, e, f)^T$ 是表示二次曲线 C 的一个6维向量





□ 2D射影变换

- 定义： IP^2 到它自身的一种满足下列条件的可逆映射 h ：三点 x_1, x_2, x_3 共线 $\Leftrightarrow h(x_1), h(x_2), h(x_3)$ 共线
- 映射 h 是射影变换的充要条件：存在一个 3×3 的非奇异矩阵 H ，使得 IP^2 的任意一个用向量 x 表示的点都满足 $h(x) = Hx$
- 射影变换表示为 $\begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ ，即 $x' = Hx$ 。射影变换有几个自由度？
- 射影变换的分解： $H = H_S H_A H_P = \begin{bmatrix} sR & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ v^T & v \end{bmatrix} = \begin{bmatrix} A & vt \\ v^T & v \end{bmatrix}$





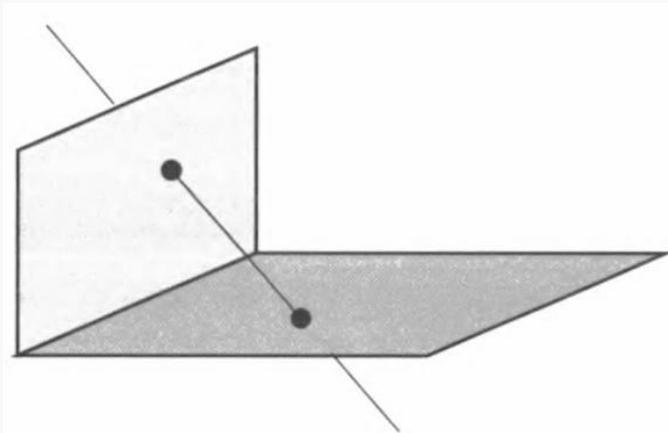
□ 变换的层次

- 等距变换: $\begin{bmatrix} \epsilon \cos \theta & -\sin \theta & t_x \\ \epsilon \sin \theta & \cos \theta & t_y \\ 0 & 0 & 1 \end{bmatrix}$, 3自由度, 不变性质: 长度、面积
- 相似变换: $\begin{bmatrix} s \cos \theta & -s \sin \theta & t_x \\ s \sin \theta & s \cos \theta & t_y \\ 0 & 0 & 1 \end{bmatrix}$, 4自由度, 不变性质: 长度比、夹角、虚圆点
- 仿射变换: $\begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix}$, 6自由度, 不变性质: 平行、面积比、共线或平行
线的长度比、向量线性组合、无穷远直线
- 射影变换: $\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$, 8自由度, 不变性质: 共点、共线、接触的阶、相
交、相切、拐点、切线不连续性和歧点、交比

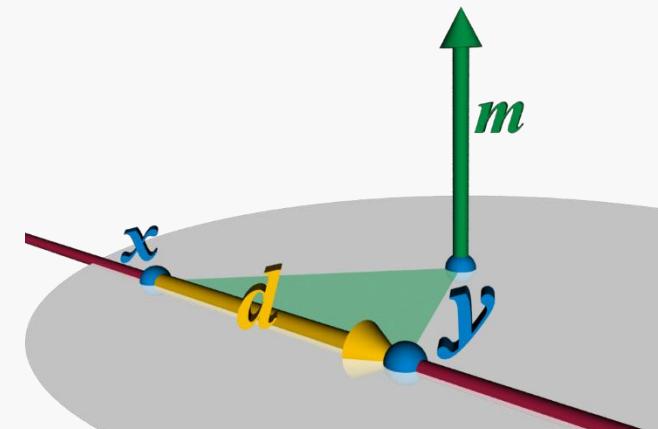


□ 点、平面和直线的表示

- 点的齐次表示：点 $(\frac{X_1}{X_4}, \frac{X_2}{X_4}, \frac{X_3}{X_4})$ 表示为4维齐次向量 $\mathbf{x} = (X_1, X_2, X_3, X_4)^T$
- 无穷远点： $\mathbf{x} = (X_1, X_2, X_3, 0)^T$
- 平面的齐次表示：3D空间的平面 $\pi_1X + \pi_2Y + \pi_3Z + \pi_4 = 0$ 表示为 $\pi_1X_1 + \pi_2X_2 + \pi_3X_3 + \pi_4X_4 = 0$, 或 $\boldsymbol{\pi}^T \mathbf{x} = 0$
- 直线的表示：零空间与生成子空间表示、Plucker矩阵、Plucker直线坐标



直线的4个自由度

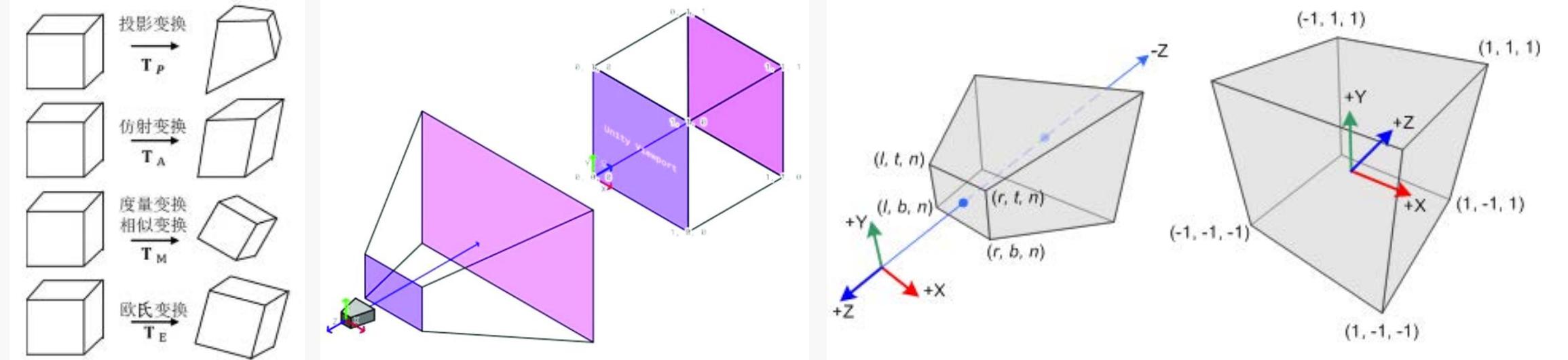


Plucker坐标



□ 3D射影变换

- 定义：非奇异 4×4 矩阵表示的线性变换 $X' = HX$
- 不变量：保线性
- 点的射影变换： $X' = HX$
- 平面的射影变换： $\pi' = H^{-T}\pi$





□ 变换的层次

- 欧式变换: $\begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}$, 6自由度, 不变性质: 体积
- 相似变换: 7自由度 $\begin{bmatrix} sR & t \\ 0^T & 1 \end{bmatrix}$, 不变性质: 绝对二次曲线, Ω_∞
- 仿射变换: 12自由度 $\begin{bmatrix} A & t \\ 0^T & 1 \end{bmatrix}$, 不变性质: 平面的平行性、体积比、形心、无穷远平面, π_∞
- 射影变换: 15自由度 $\begin{bmatrix} A & t \\ v^T & v \end{bmatrix}$, 不变性质: 接触表面相交和相切、高斯曲率的符号



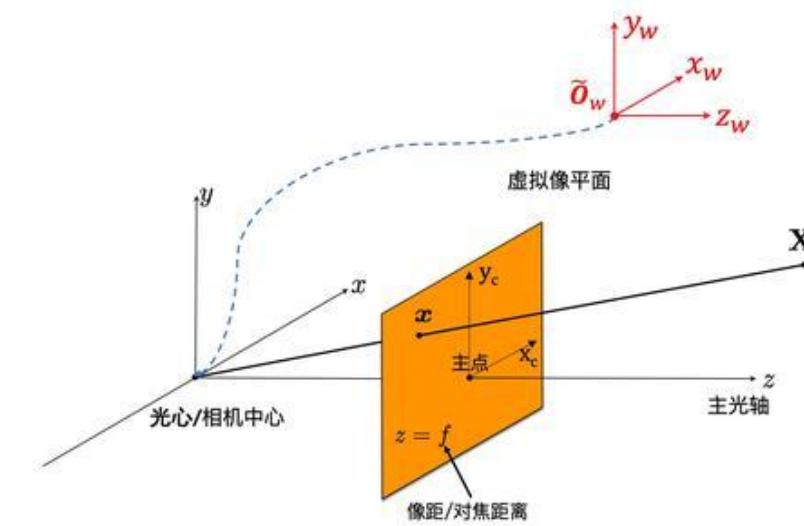
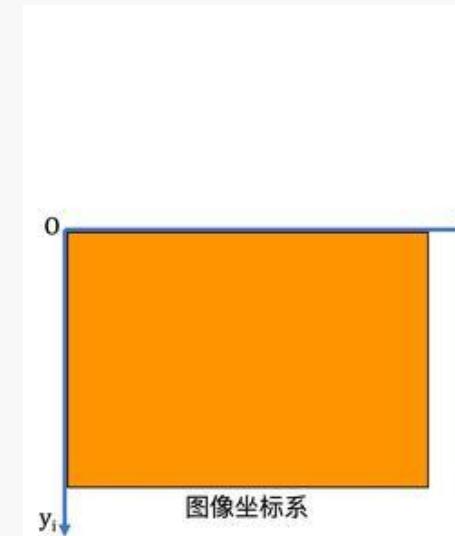
摄像机模型

- 一、有限摄像机
- 二、射影摄像机
- 三、无穷远摄像机

摄像机模型

□ 引言

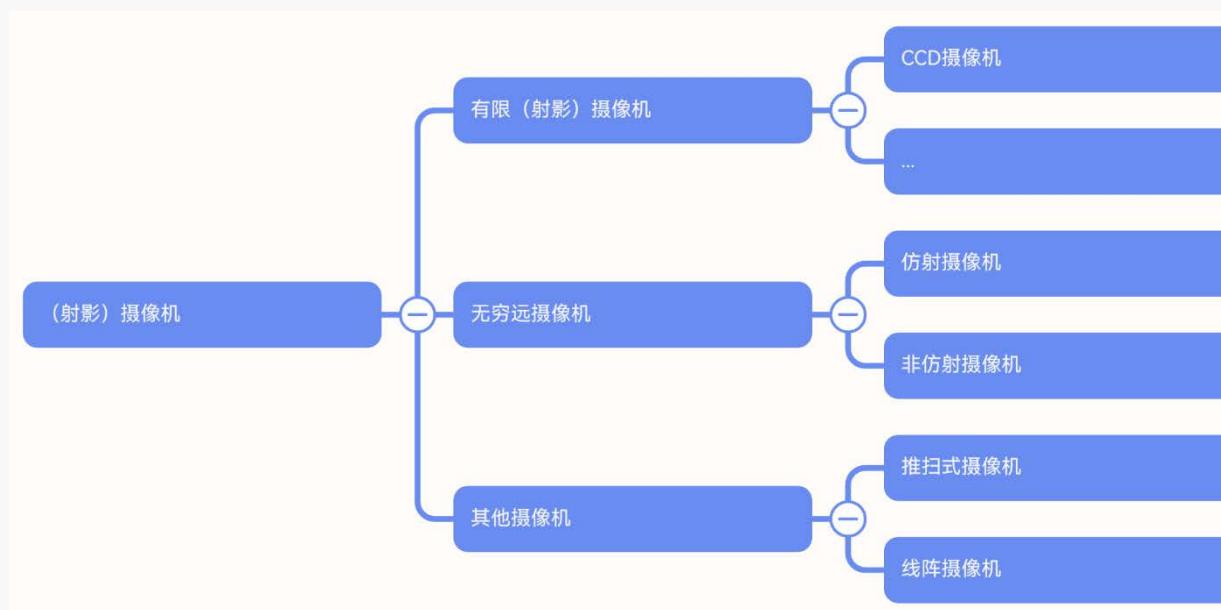
- 摄像机是3D到2D的空间映射，适合用射影几何工具研究
- 实际摄像机多为中心投影，中心投影将是讨论重点
- 摄像机投影可表示为齐次坐标下的映射矩阵，摄像机的几何元素可用映射矩阵计算得到，摄像机内在性质可用代数表达式计算



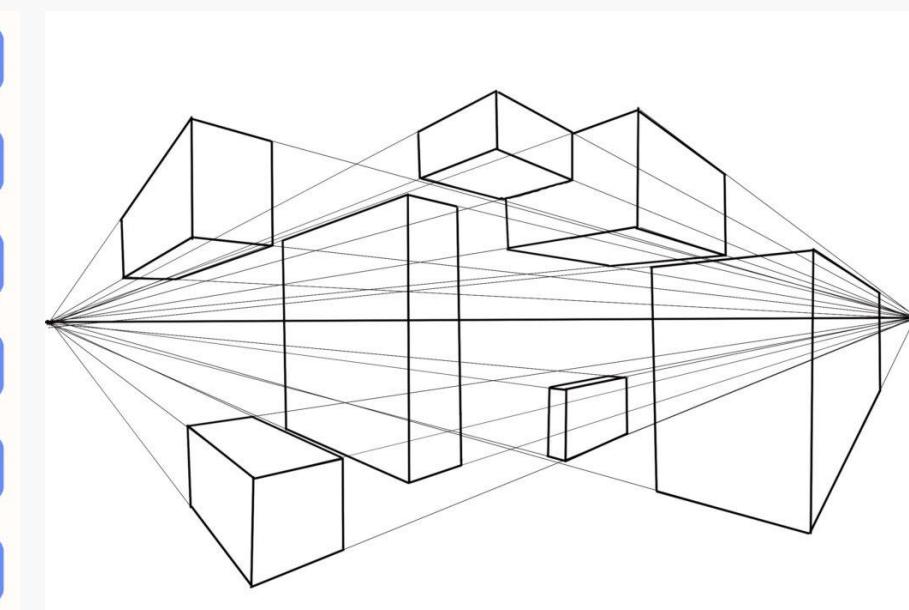


□ 摄像机的分类

- 根据透视：摄像机分为有限摄像机、无穷远摄像机和其他摄像机
- 有限摄像机相机中心在有限距离，无穷远摄像机中心在无穷远平面上
- CCD摄像机是讨论重点



摄像机分类

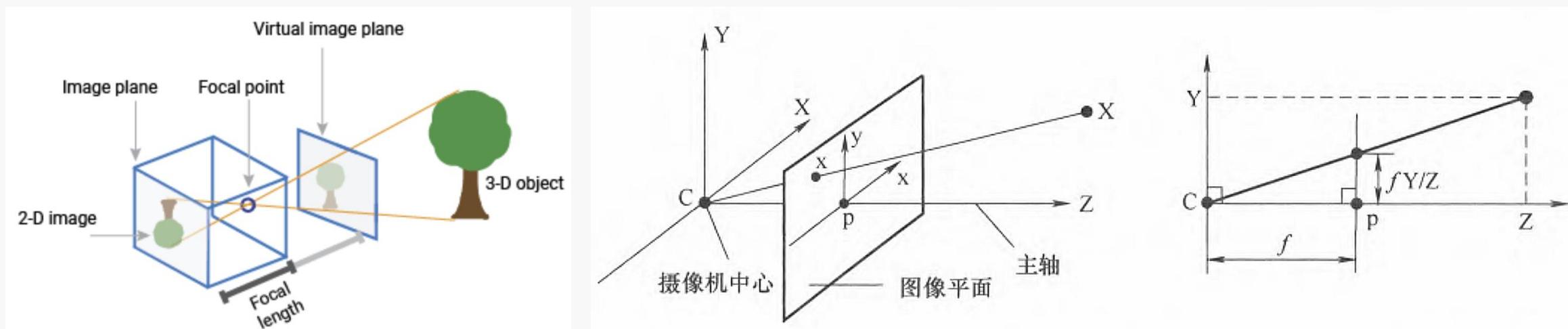


透视关系

有限摄像机

□ 基本针孔模型（摄像机与世界坐标对齐）

- 图像平面: $Z = f$, 是空间点投影到的平面
- 主平面: 过摄像机中心且平行于图像平面的平面
- 投影方式: 使用中心投影, 摄像机中心位于坐标原点
- 映射关系: 世界坐标到图像平面坐标的映射为 $(X, Y, Z)^T \rightarrow \left(\frac{fX}{Z}, \frac{fY}{Z}\right)^T$





□ 摄像机投影矩阵

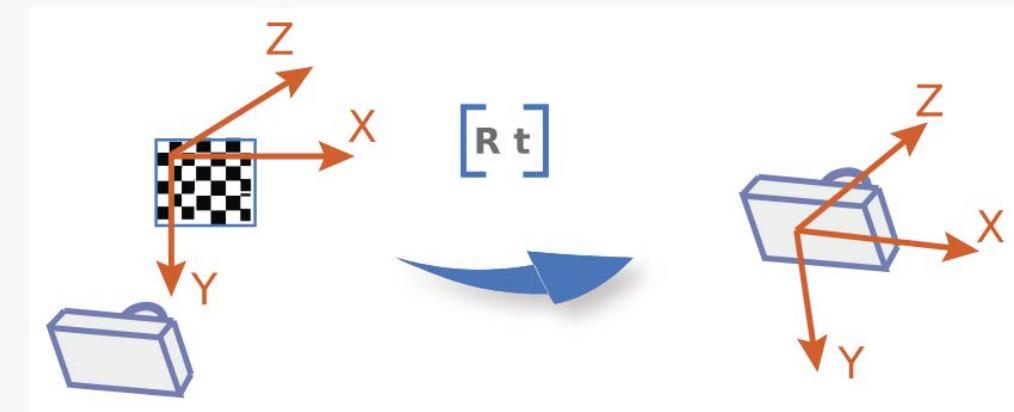
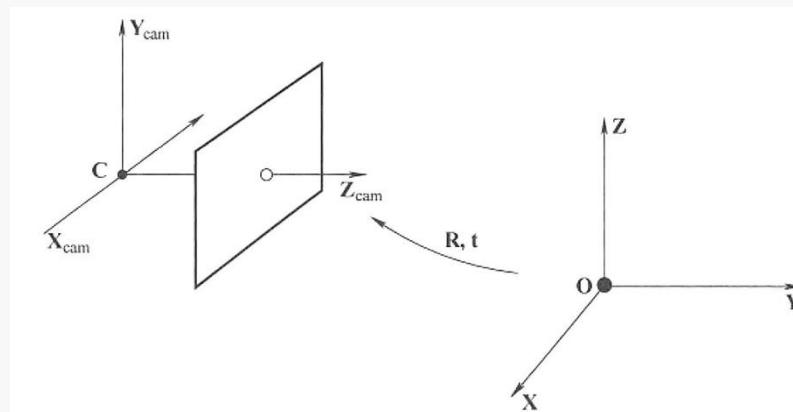
- 世界点 X 用4维齐次向量 $(X, Y, Z, 1)^T$ 表示，图像点 x 用3维齐次向量 $\left(\frac{fX}{Z}, \frac{fY}{Z}, 1\right)^T$ 表示
- P 表示 3×4 齐次摄像机投影矩阵
- 主点偏置：图像平面上，主点坐标为 $(p_x, p_y)^T$
- 中心投影： $x = PX$ 或 $x = K[I|0]X_{cam}$ 或 $\begin{bmatrix} fX + Zp_x \\ fY + Zp_y \\ Z \end{bmatrix} = \begin{bmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$
- 摄像机标定矩阵： $K = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix}$



□ 摄像机旋转与平移

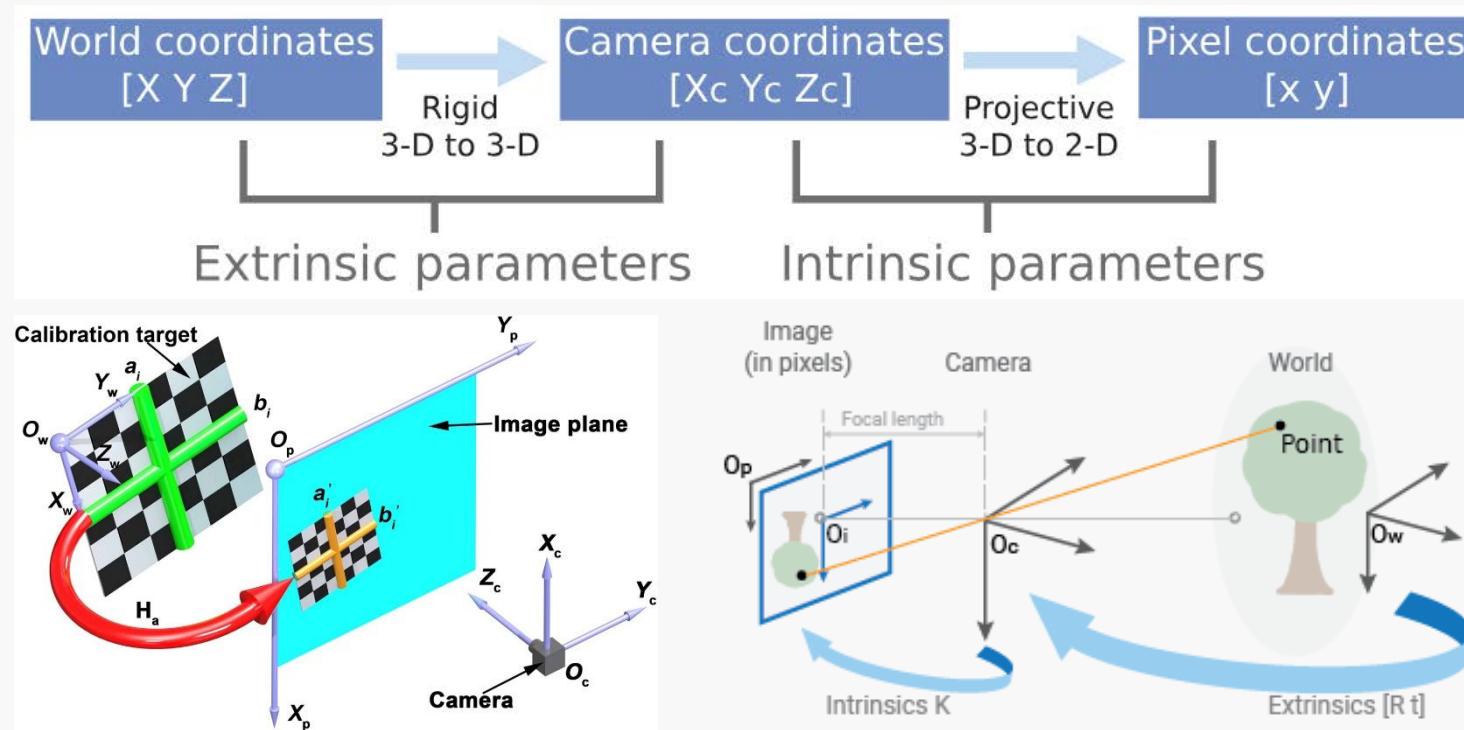
- 世界坐标系：表示空间点的不同欧式坐标系，坐标为 X
- 相机坐标系：相机位于原点，相机主轴沿 Z 轴方向，坐标为 X_{cam}
- \tilde{C} 表示相机中心在世界坐标系的坐标， R 表示相机坐标系的 3×3 旋转矩阵

$$\boxed{\text{■ } X_{cam} = \begin{bmatrix} R & -R\tilde{C} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} R & -R\tilde{C} \\ 0^T & 1 \end{bmatrix} X}$$



□ 摄像机内参数与外参数

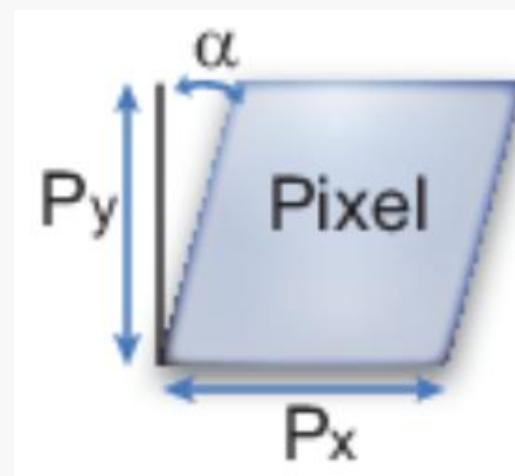
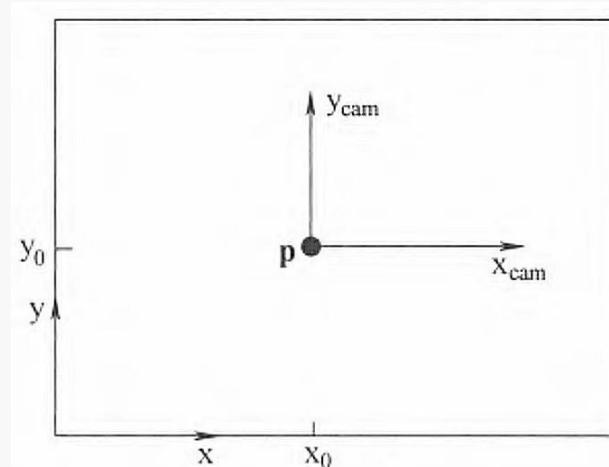
- 相机内参: K 中包含的参数
- 相机外参: R 和 \tilde{C} 中包含的参数, 可简化为 $[R|t]$
- 相机投影模型: $x = KR[I] - \tilde{C}X$, 或进一步简化为 $x = K[R|t]X$





□ 有限射影摄像机

- 图像坐标以像素为单位, x 和 y 方向上单位距离的像素数为 m_x 和 m_y
- 有限射影摄像机的内参为 $K = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}$, 其中 $\alpha_x = fm_x$, $\alpha_y = fm_y$,
 $x_0 = m_x p_x$, $y_0 = m_y p_y$
- s 为扭曲参数, 对于CCD等大多数标准摄像机 $s = 0$





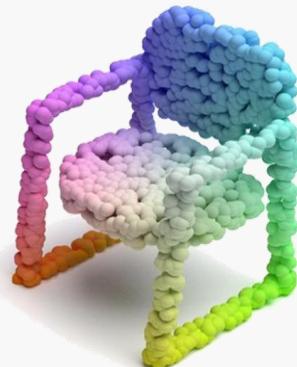
三维视觉表示

- 一、显式三维视觉表示
- 二、隐式三维视觉表示



□ 显式三维表示

- 显式表示直接给出或者通过参数映射给出空间中点线面等的信息
- 优点：数据结构直观，易于进行几何处理和渲染
- 缺点：存储开销大，分辨率受限于内存大小
- 应用：三维感知，三维建模，三维渲染



点云



面片



体素

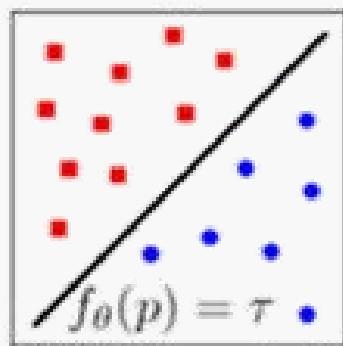


三维高斯

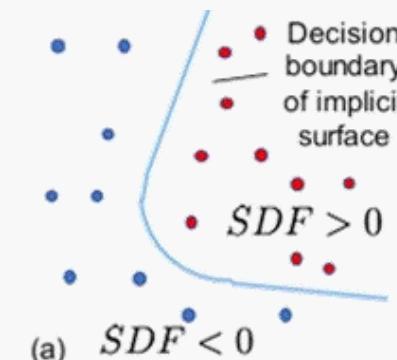


□ 隐式三维表示

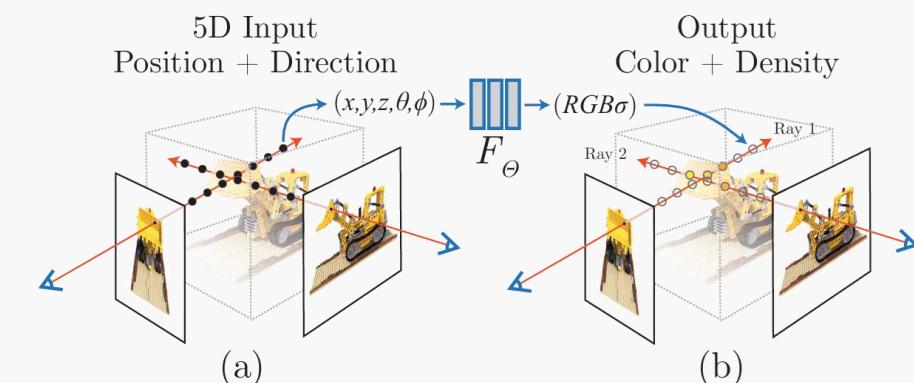
- 隐示表示通过函数 $f(x)$ 表示三维视觉内容
- 优点：拓扑结构精细，理论上具有无限的分辨率，且内存占用较少
- 缺点：推断效率较低，需要较长的训练时间，且需要后处理才能得到显式的几何结构
- 应用：三维重建，三维渲染



占用场



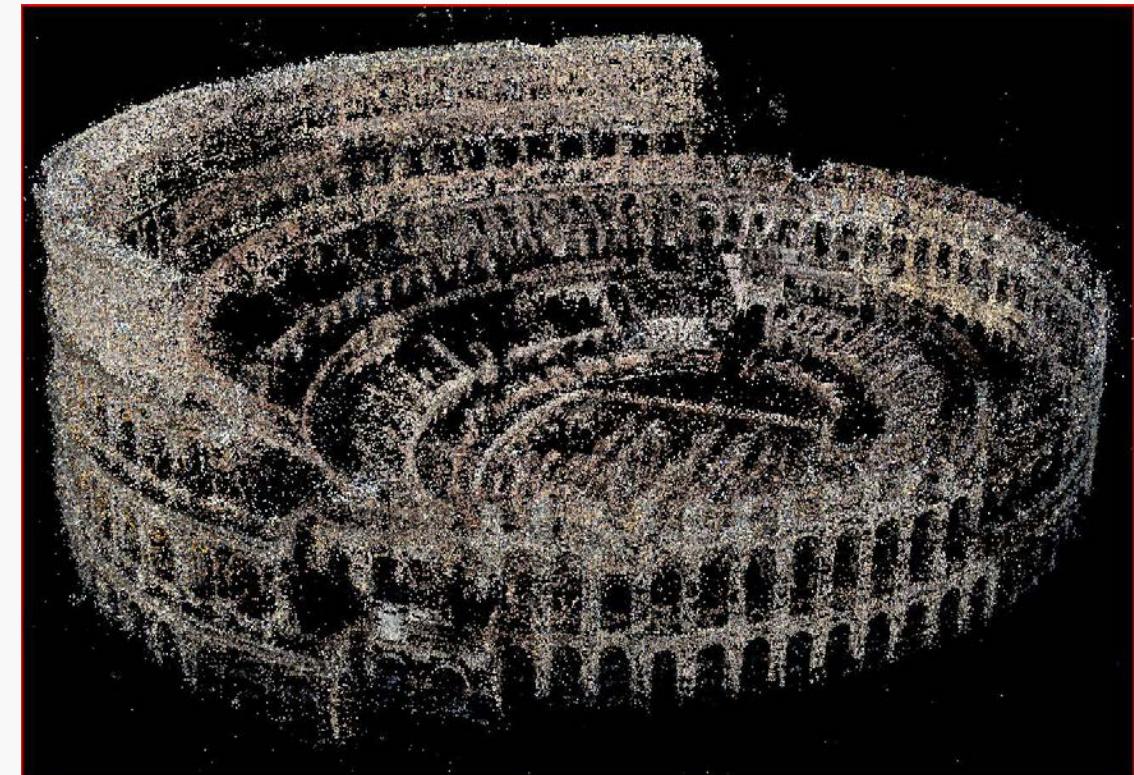
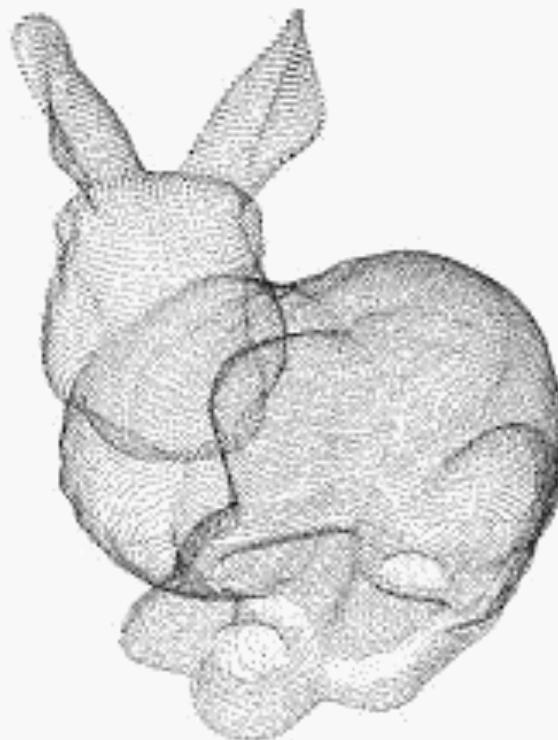
符号距离场



神经辐射场

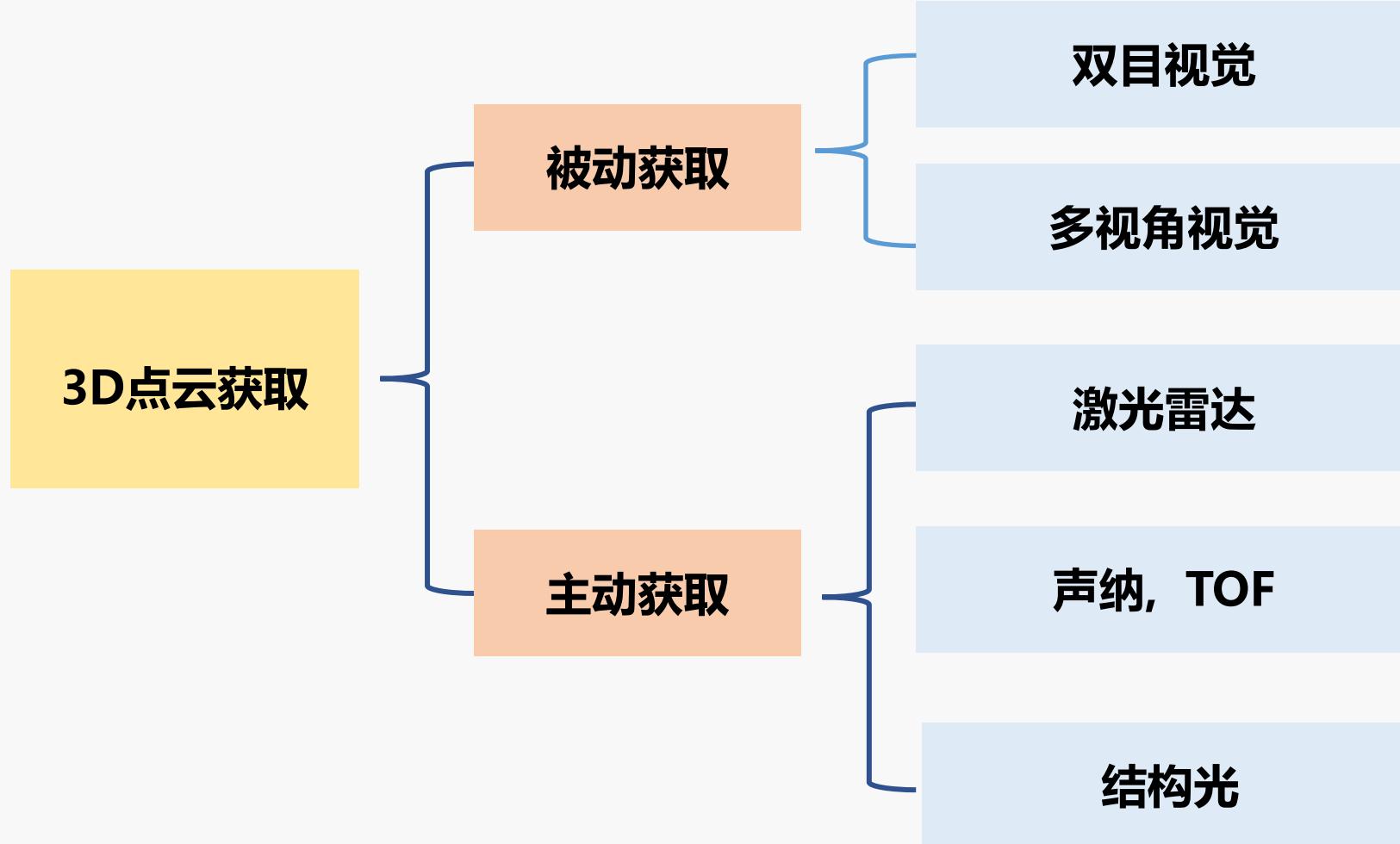


- 在三维坐标系下点的集合
- 点云是一个 $n \times 3$ 的矩阵吗？
- 可以包括三维坐标、颜色、法线方向、分类值、强度值、时间等





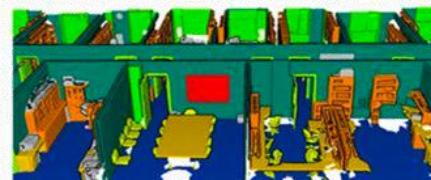
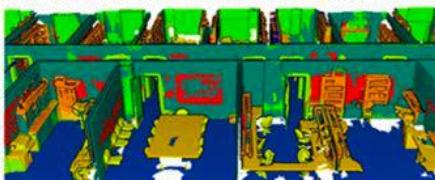
口 点云来源



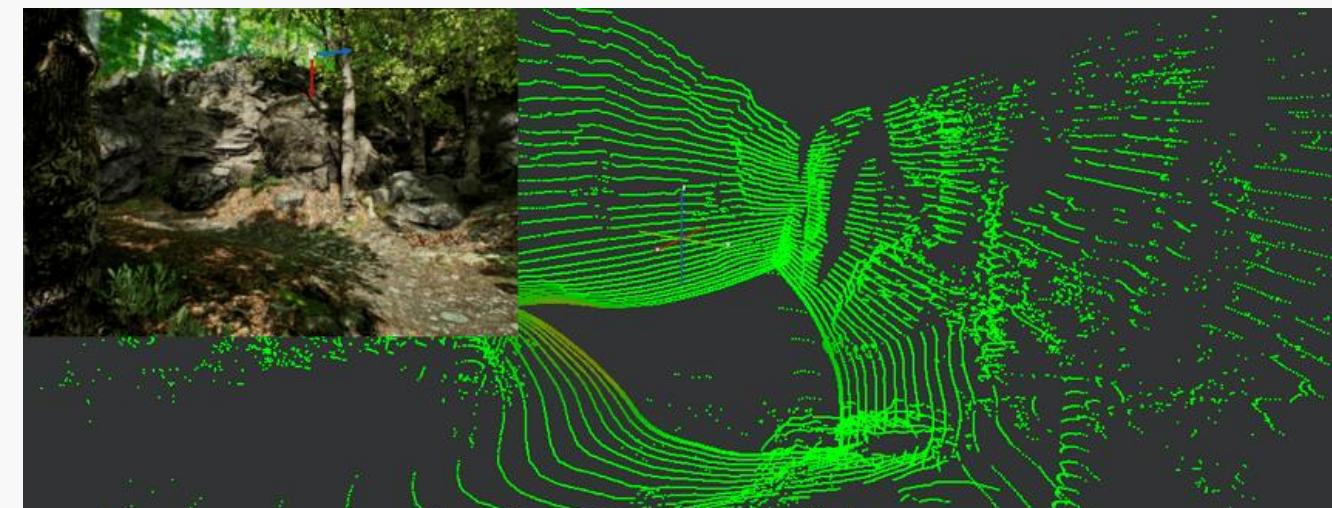


□ 两种常用的传感器

传感器	点云分布	点云属性	应用范围	成本
RGB-D相机	均匀、致密	颜色、法向	近景、室内	较低
LiDAR	不均匀、稀疏	反射率	大场景、室外	较高



RGB-D相机



LiDAR



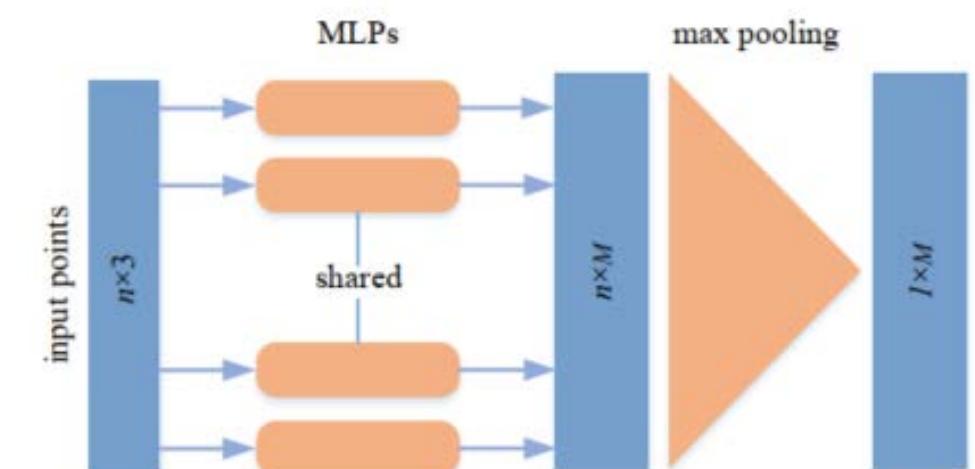
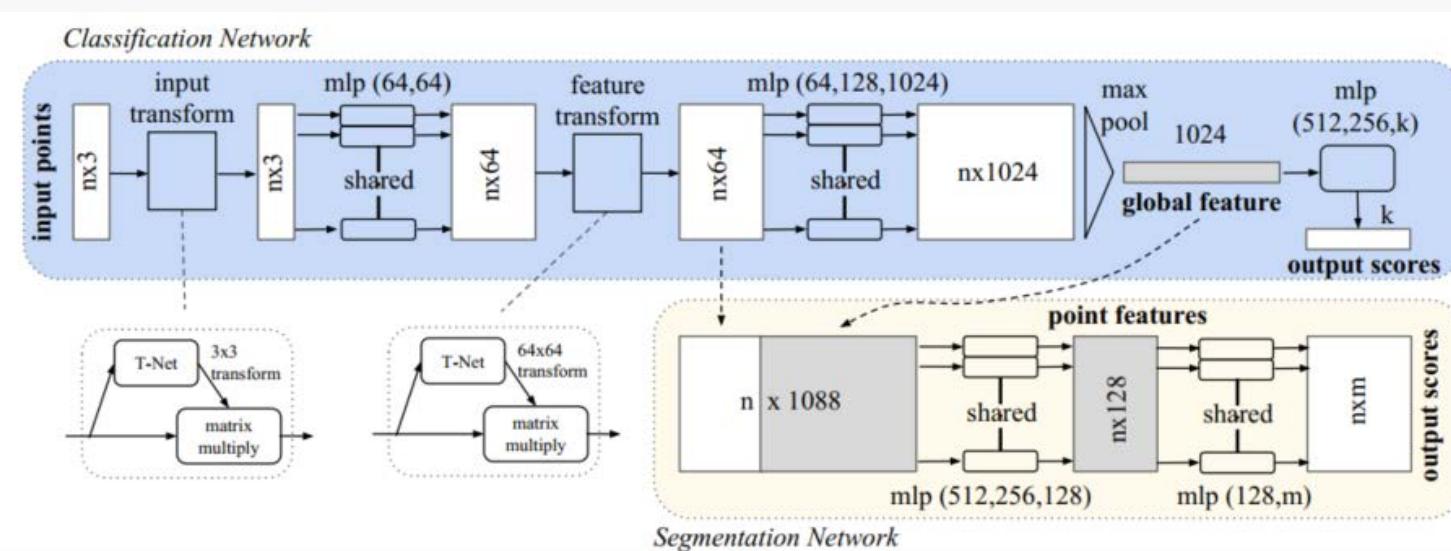
□ 常用点云数据集

数据集	内容	种类	数据	标签
ShapeNet	51,190 objects	物体	Mesh	物体类别标签
ModelNet40	12,311 objects	物体	Mesh	物体类别标签
ScanObjectNN	2,902 objects	物体	Points	物体类别标签
ScanNet	1,513 scans	室内场景	RGB-D	点云类别标签 检测框
SUN RGB-D	5K frames	室内场景	RGB-D	检测框
S3DIS	272 scans	室内场景	RGB-D	点云类别标签
KITTI	15K frames	室外场景	RGB & LiDAR	检测框
SemanticKITTI	45K frames	室外场景	LiDAR	点云类别标签
SemanticPOSS	2K frames	室外场景	LiDAR	点云类别标签
Waymo	15K frames	室外场景	LiDAR	点云类别标签 检测框
nuScene	40K frames	室外场景	RGB & LiDAR	点云类别标签 检测框

基础点云骨干网络

口 点云基础感知架构：PointNet

- 斯坦福大学
- 利用权重共享的MLP和对称函数（Max Pooling）实现点云的**保序性**
- 并未对点云的**局部特征**进行建模



- T-Net：学习 3×3 的矩阵仿射变换，将不同位姿的点云对齐

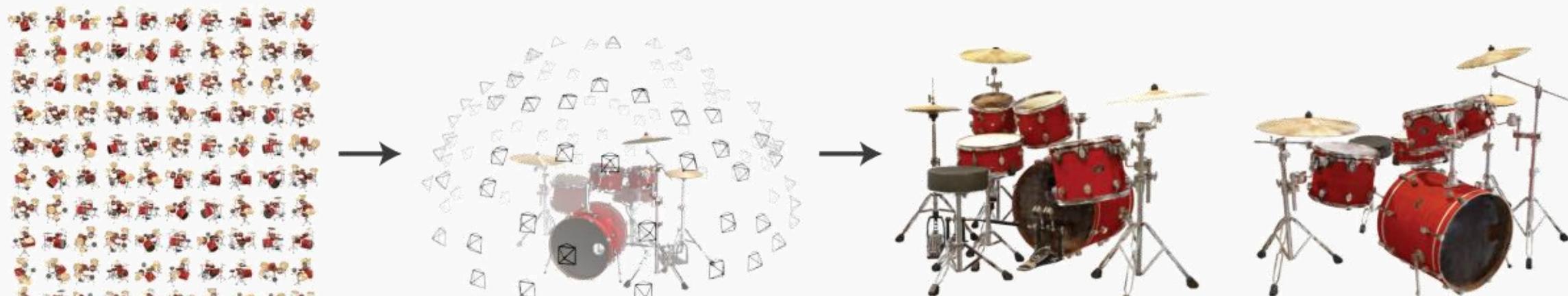
PointNet, CVPR'2017



神经辐射场

□ 神经辐射场（Neural Radiance Fields，简称NeRF）

- 从多个视角的图像中提取出对象的几何形状和纹理信息，然后使用这些信息生成一个连续的三维辐射场，从而可以在任意角度和距离下呈现出高度逼真的三维模型。



多视角图像输入

优化NeRF

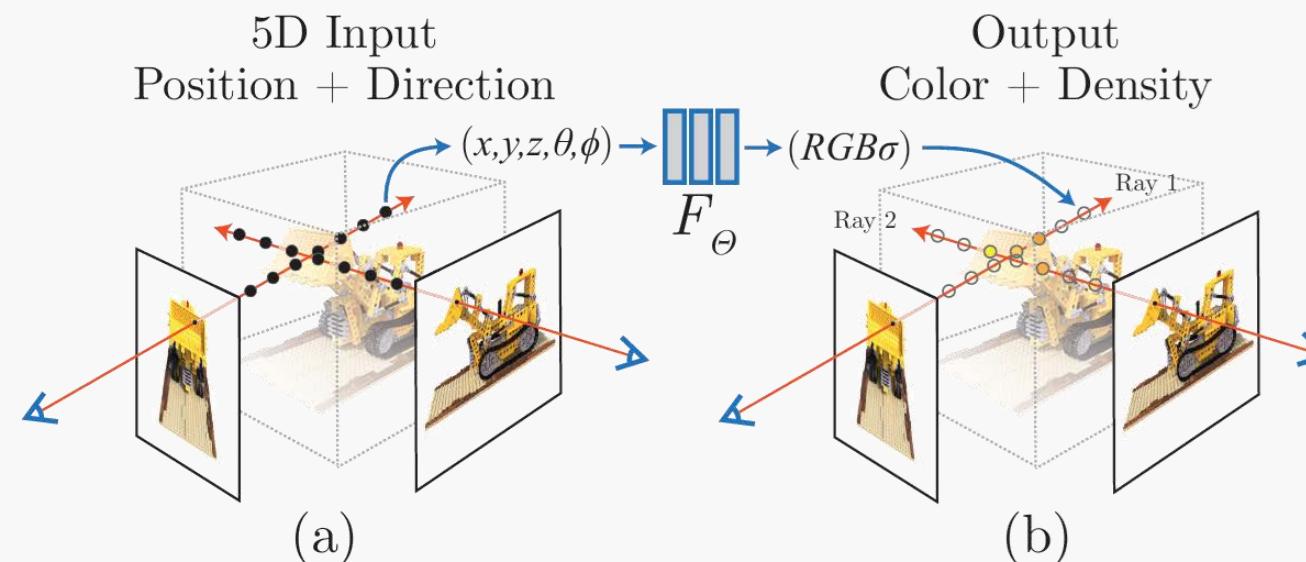
渲染新视角

神经辐射场

□ 利用神经网络将场景表示为连续的体素密度场与颜色场

- 输入：3D位置 $\mathbf{x} = (x, y, z)$ 和2D视角方向 (θ, ϕ)
- 输出：颜色 $c = (r, g, b)$ 和体密度 σ
- 使用神经网络 F_Θ 估计连续的场景表示

$$F_\Theta: (x, y, z, \theta, \phi) \rightarrow (r, g, b, \sigma)$$





□ 渲染流程: Volume Rendering

- 给定相机射线 $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$

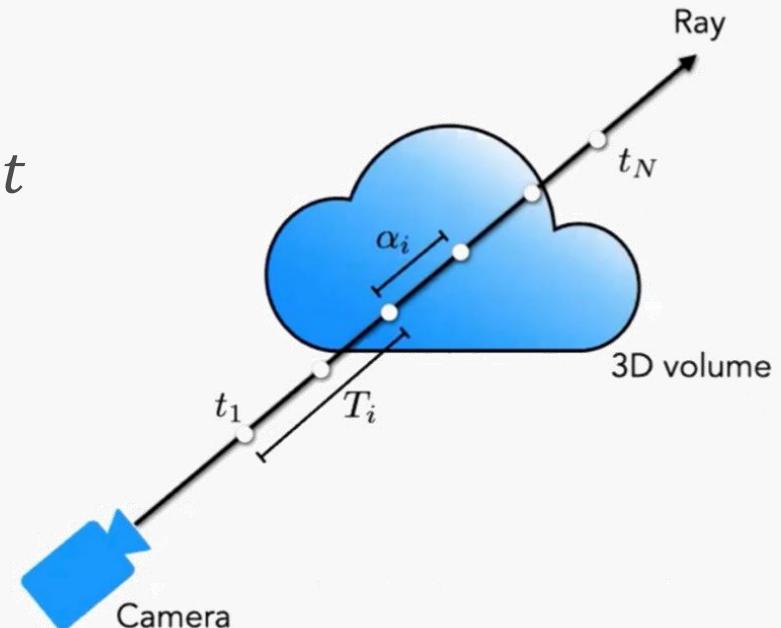
$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt$$

其中 $T(t)$ 表示沿着光线的累积透射率

$$T(t) = \exp(- \int_{t_n}^t \sigma(\mathbf{r}(s)) ds)$$

$\sigma(\cdot)$ 表示在该点的体密度, $\mathbf{c}(\cdot)$ 表示在该点的颜色

t_n 和 t_f 表示最近和最远的边界, 只计算边界内的体渲染



□ 渲染流程: Volume Rendering

■ 离散化表达

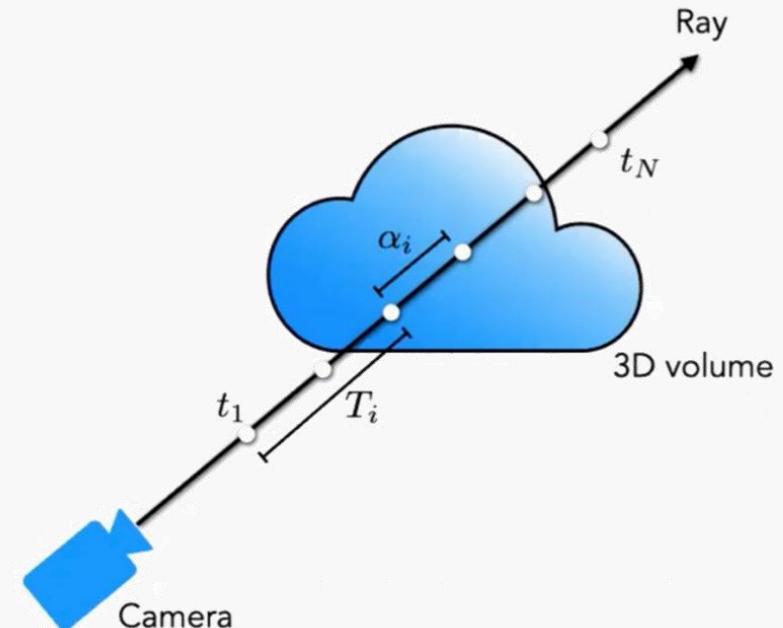
$$C(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i$$

其中 T_i 表示沿着光线的累积透射率

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$$

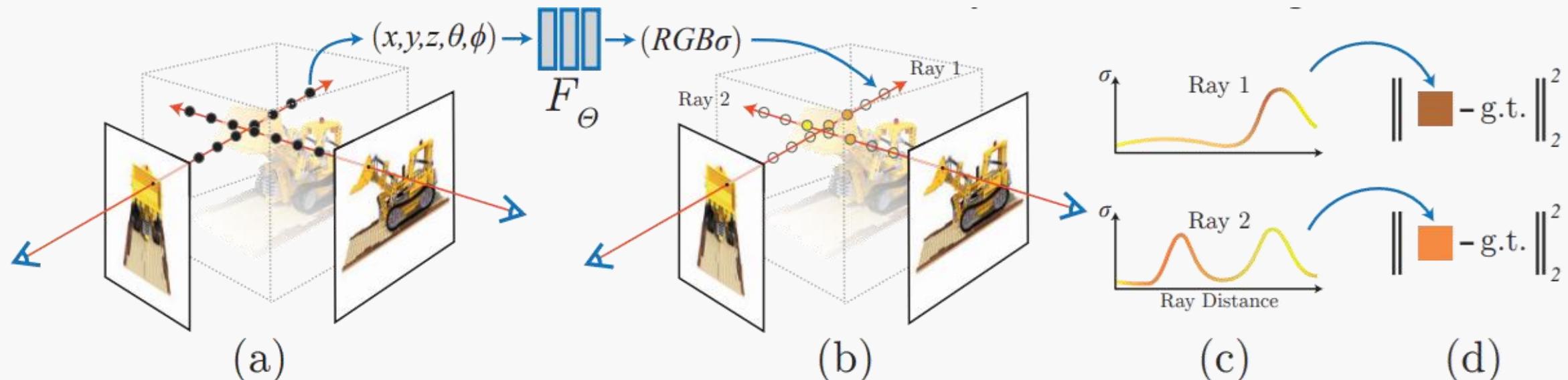
其中 $\delta_i = t_{i+1} - t_i$ 表示离散采样点之间的距离

σ_i 表示离散采样点的体密度， \mathbf{c}_i 表示离散采样点的颜色



□ 神经辐射场的训练

- 输入：多视角图像和相机位姿
- 优化：隐式神经辐射场 F_Θ
- 损失函数：MSE均方误差损失函数





□ 神经辐射场的效果

- MLP network可以表示连续高分辨率场景





□ 3D高斯 (3D Gaussian Splatting) : SIGGRAPH 2023 best paper award

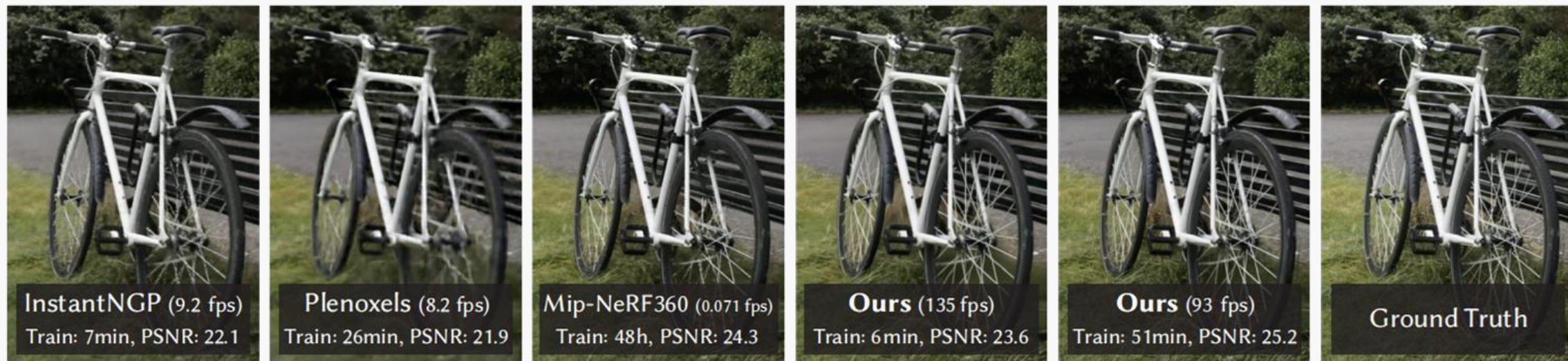
3D Gaussian Splatting for Real-Time Radiance Field Rendering

BERNHARD KERBL*, Inria, Université Côte d'Azur, France

GEORGIOS KOPANAS*, Inria, Université Côte d'Azur, France

THOMAS LEIMKÜHLER, Max-Planck-Institut für Informatik, Germany

GEORGE DRETTAKIS, Inria, Université Côte d'Azur, France



gaussian-splatting

Public

Watch 92

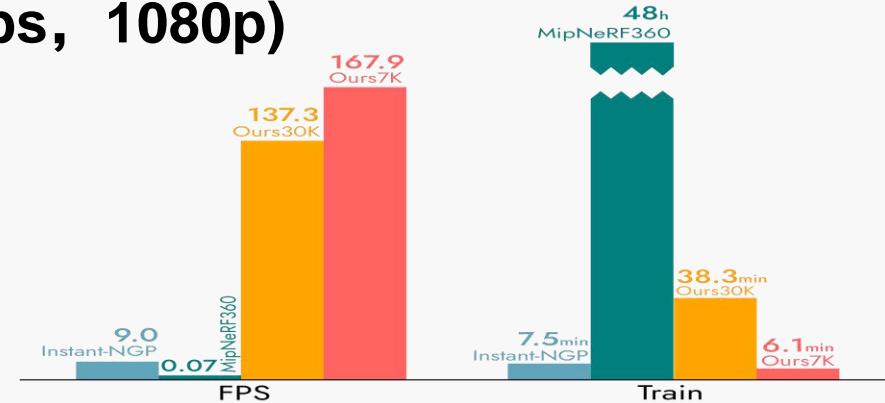
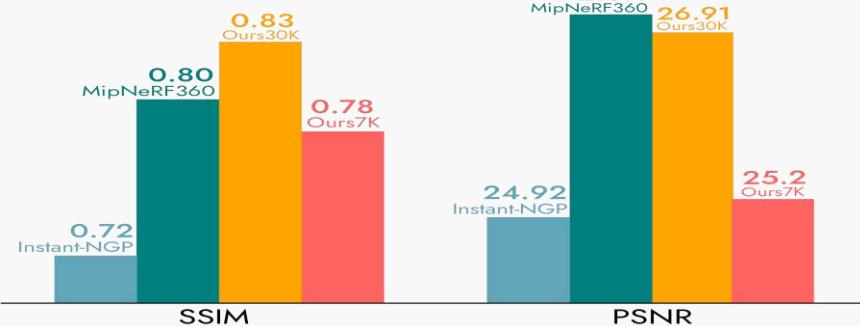
Fork 738

Starred 7.5k



□ 3D高斯的性能与样例

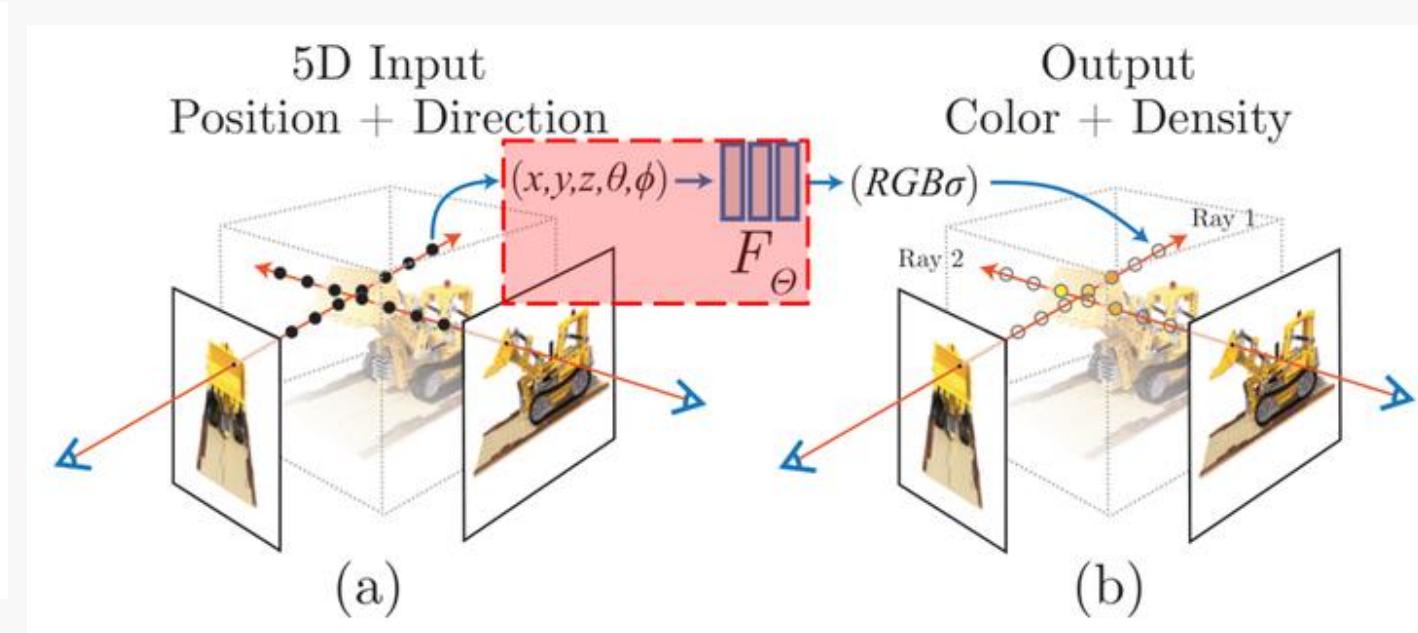
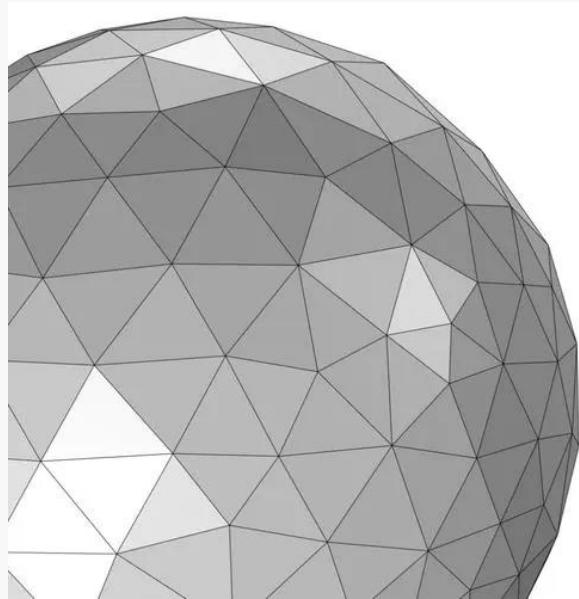
- 首个实时渲染的高质量重建方法 (>100 fps, 1080p)





□ 3D高斯表示的出发点：多种表示方式在重建任务上的不足

- 体素 / 点云: 成熟的渲染机制 (投影, 渲染管道等)
但**并非连续表示**, 难以拥有**高质量的渲染重建细节**
- NeRF: 可微的连续体素表示, 提供高质量的渲染细节
需要随机采样并经过MLP, **不适合当前显卡与渲染管道进行渲染**





□ 文章如何定义三维高斯？

$$p(\mathbf{x}) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{3/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \rightarrow G(x) = e^{-\frac{1}{2} (x)^T \Sigma^{-1} (x)}$$



单个三维高斯包含

• 位置信息	3	参数	中心位置
• 协方差矩阵	3+4	参数	形状姿态 $\Sigma = RSS^T R^T$
• 不透明度 α	1	参数	不透明度
• 球谐函数系数	n^*3	参数	颜色贡献

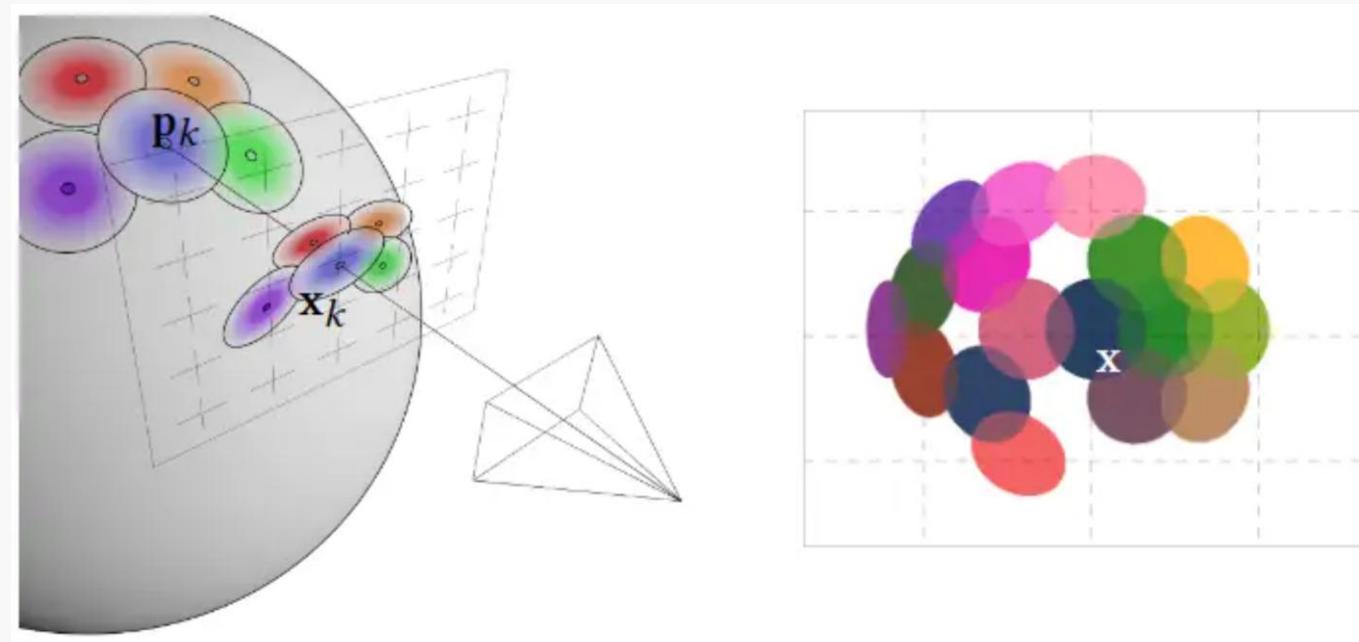
- 全局而言：离散的高斯球
- 局部而言：连续可微的球信息



高斯泼溅

□ 什么是泼溅 (splatting)

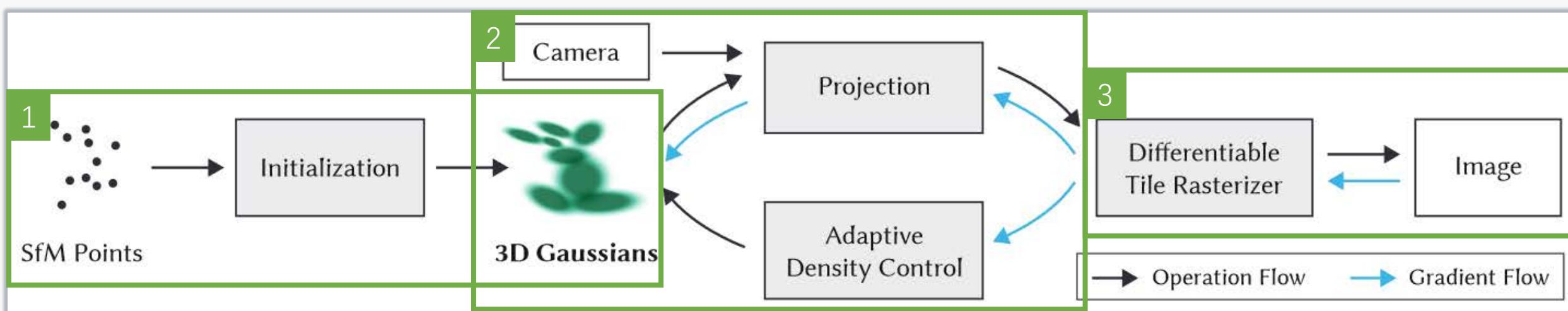
- 计算每一体素投影的影响范围，用高斯函数定义点或者小区域像素的强度分布，从而计算出其对图像的总体贡献，并加以合成，形成最后的图像。
- 速度快：无需稠密采样，渲染速度快



高斯泼溅

方法简介

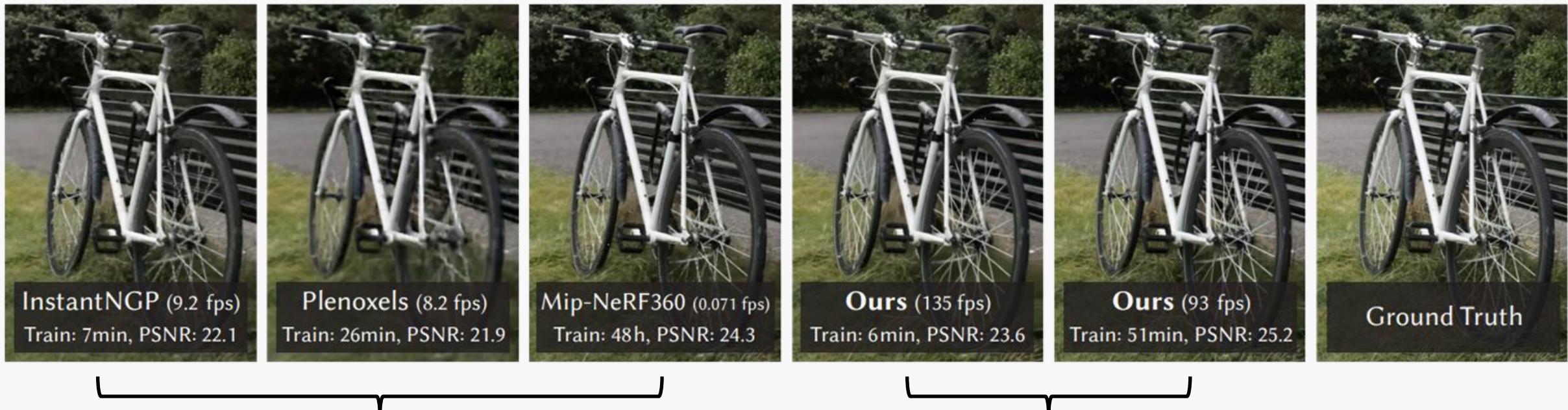
- 从初始的SFM点云出发，以每个点为中心生成三维高斯
- 用相机参数把点投影到图像平面上，使用三维高斯泼溅方法得到渲染图像，计算渲染图像和真实图像求损失函数，反向传播
- 自适应的密度控制模块根据传递到点上的梯度，来决定是否需要对3DGS做分割或者克隆。梯度传递到三维高斯，对其参数进行更新





□ 3DGS的效果

- 训练快
- 渲染快
- 重建质量好



基于神经辐射场的方法

3DGS

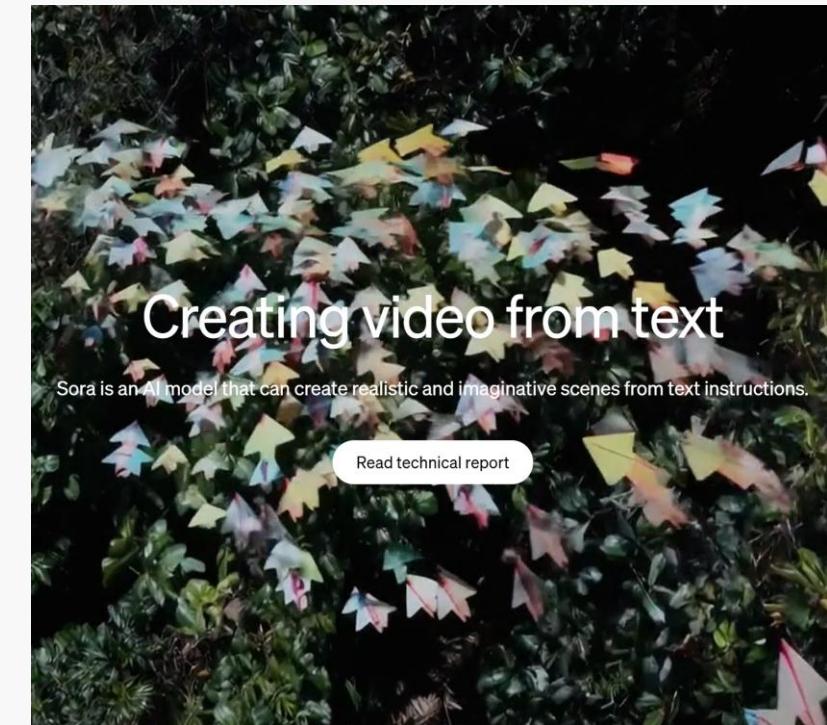
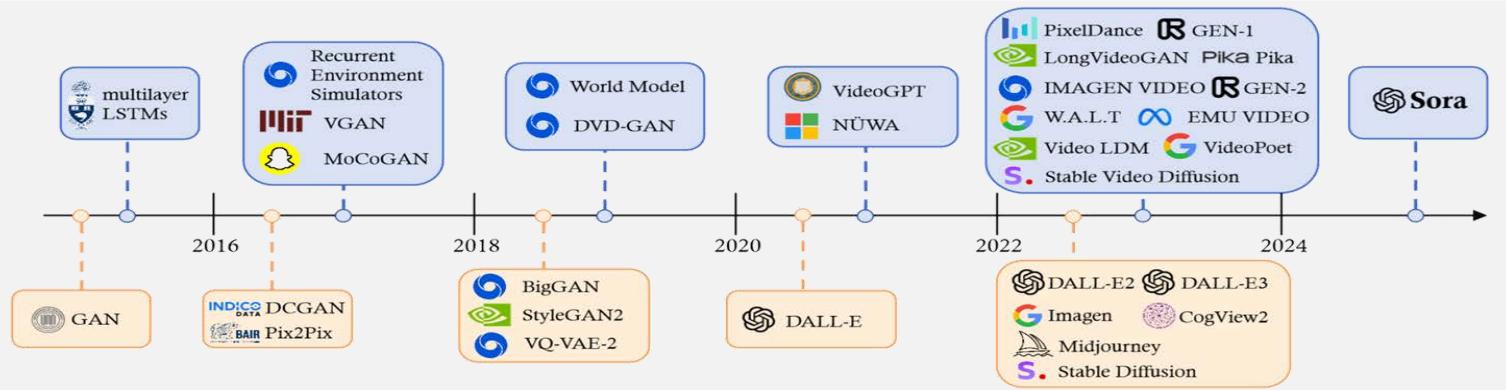
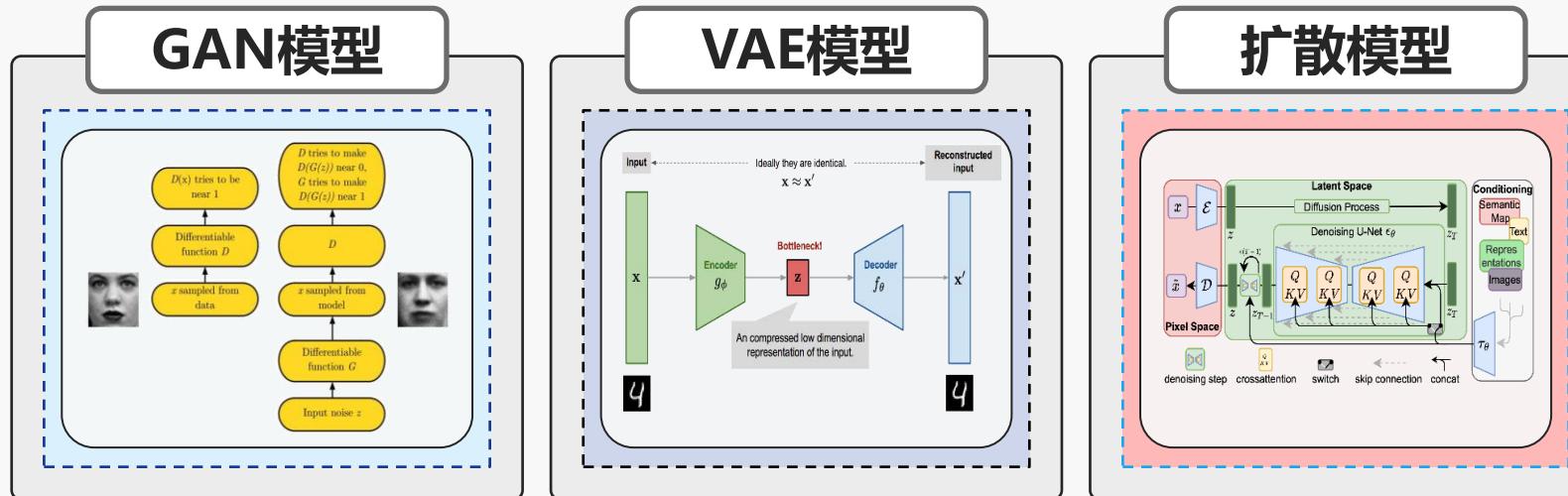


三维视觉生成

- 一、三维视觉生成概述
- 二、三维视觉生成主流方法

三维视觉生成概述

- 视觉内容生成的目的是生成高质量且多样的视觉数据
- 随着DALL-E、Sora等模型的提出，视觉内容生成受到广泛关注



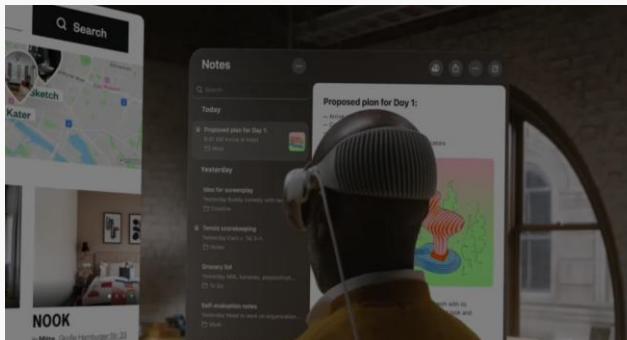
Sora

三维视觉生成概述

- 当前视觉内容生成研究大多集中在二维图片或视频，随着**VR/AR、自动驾驶、具身智能**等应用领域的发展，**三维资产需求日益增加**



虚拟现实



增强现实
VR/AR



边角案例



数据扩展
自动驾驶



仿真模拟

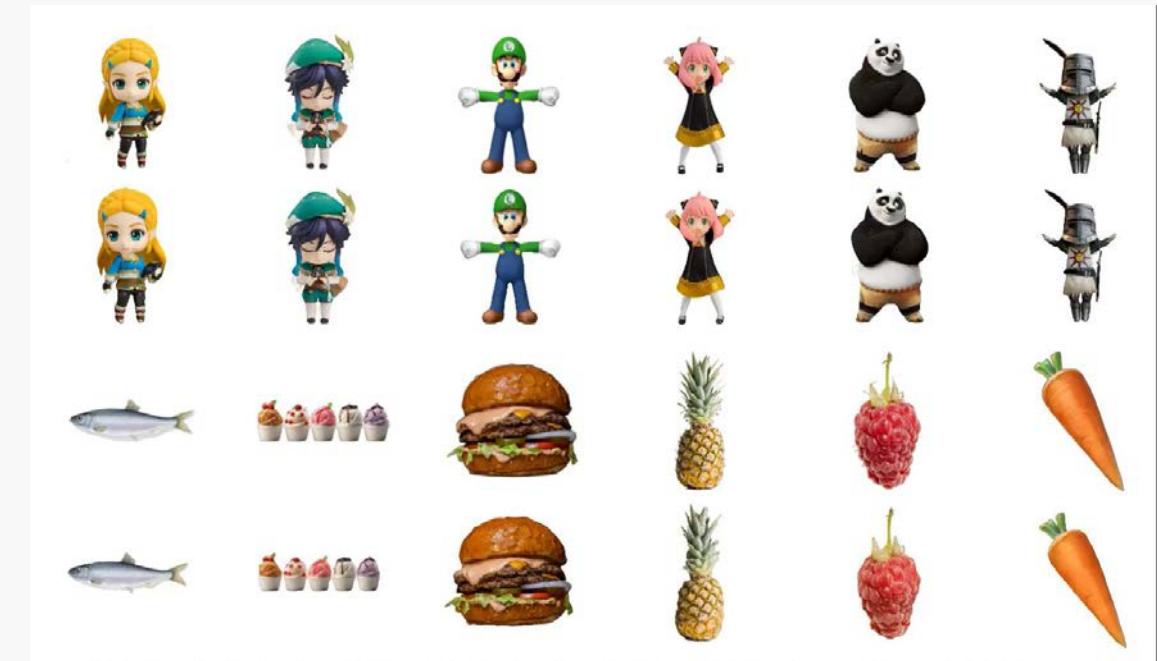
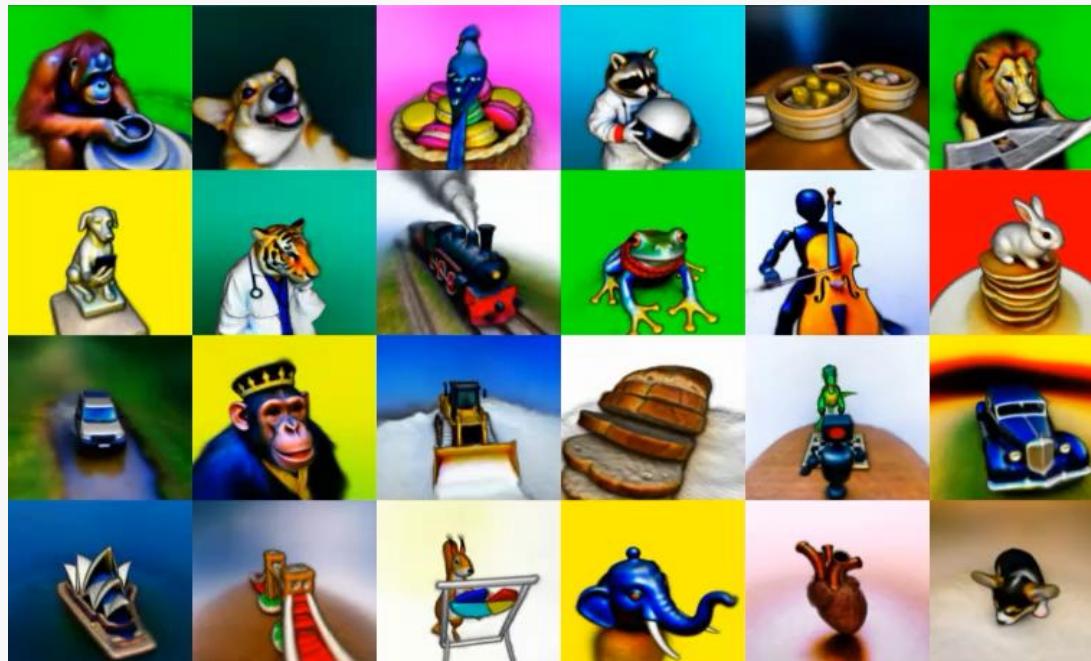


感知推理
具身智能



□ 什么是三维生成？

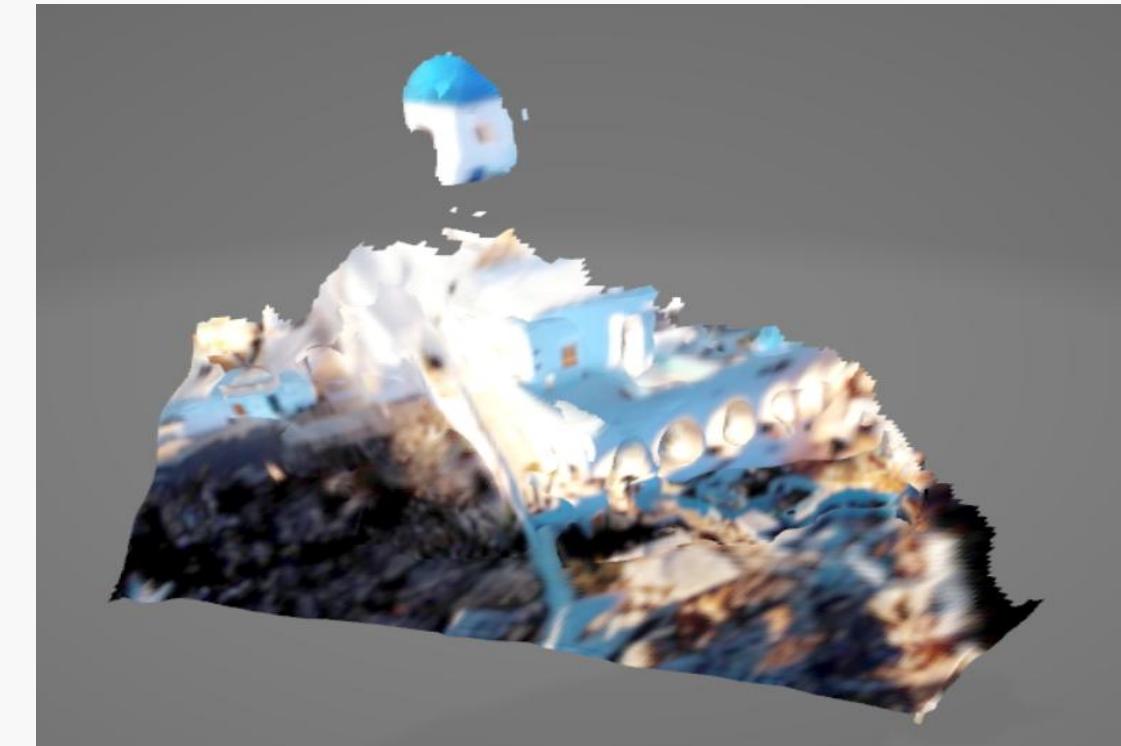
- 三维生成：给定**图像或文本提示**，输出高质量、高保真度、多视角一致的**特定三维对象表示**





三维视觉生成概述

□ Sora可以生成三维视觉内容吗？





三维目标生成

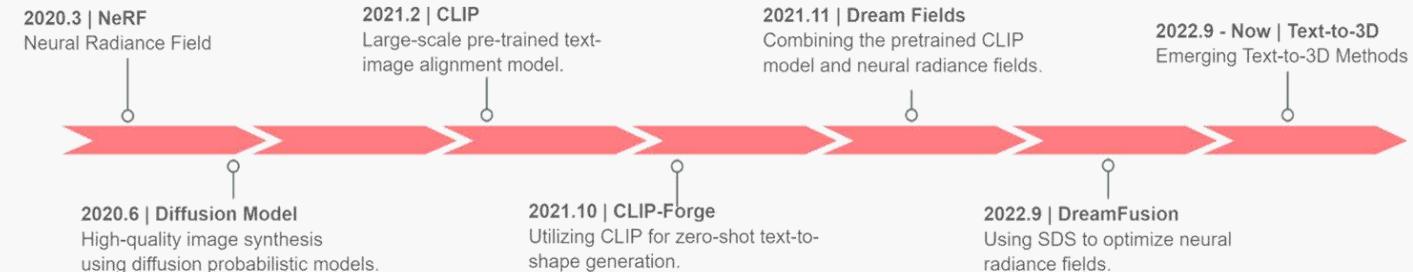
- 三维内容生成主要分为**文生3D**和**图生3D**两类
- 三维视觉内容生成能够在生动的人类想象领域与数字创作的现实世界之间架起桥梁





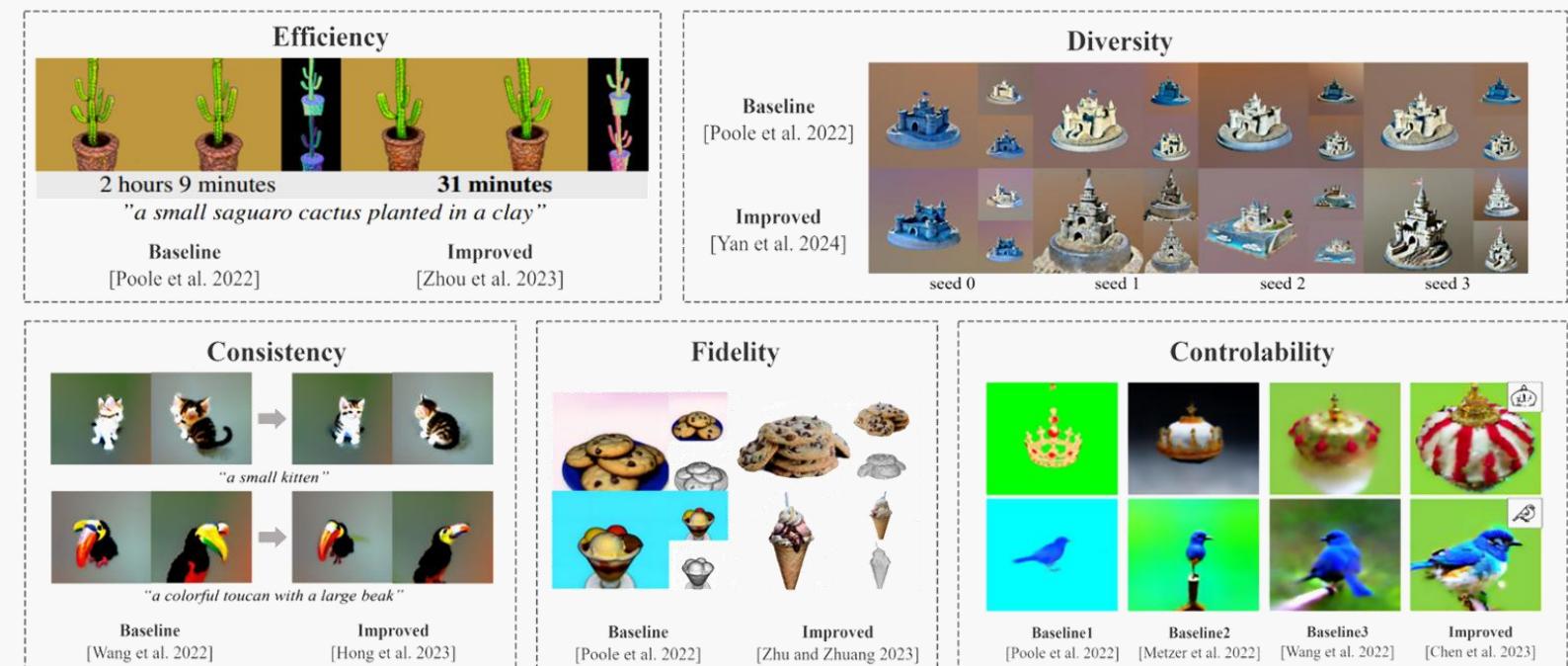
□ 三维视觉生成主流方法分类:

- 二维视觉大模型蒸馏
- 多视角生成重建
- 三维原生模型训练方法



□ 关注核心问题

- 高效性
- 多样性
- 一致性
- 真实性
- 可控性



二维大模型蒸馏三维知识

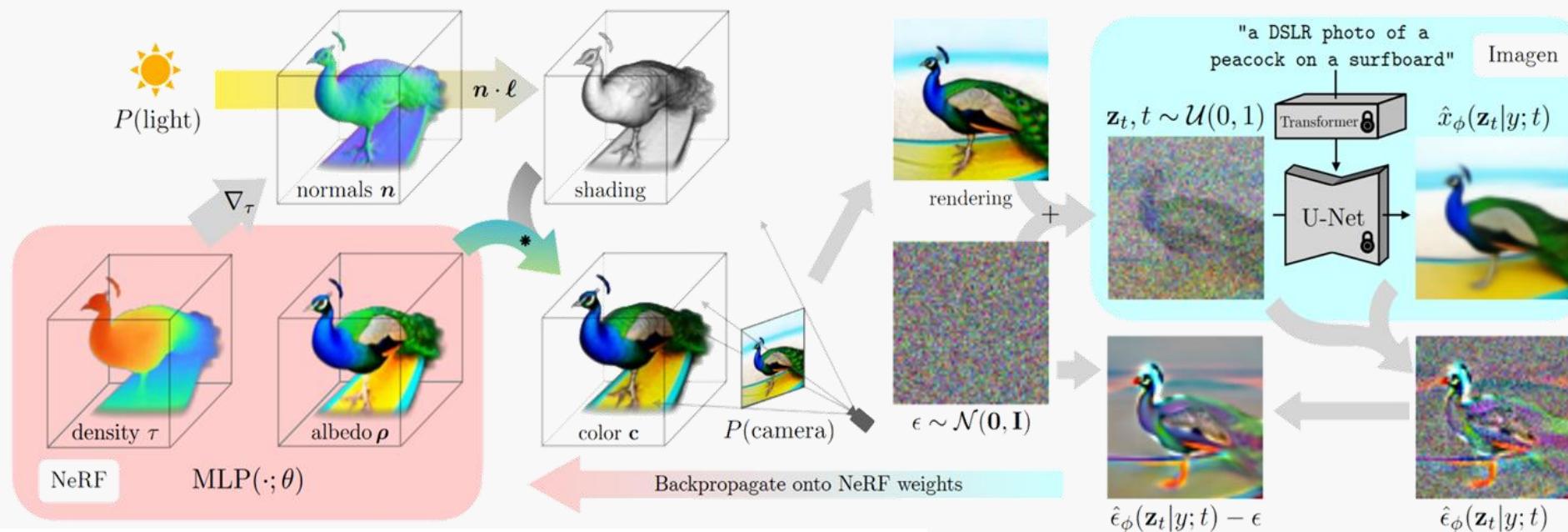
□ 不需要任何三维数据进行训练 (No 3D Data)

- 三维物体的每一个侧面 (2D Image) 都在二维扩散生成大模型 (Stable Diffusion / Imagen ...) 的数据分布当中达到最大似然，本质是二维模型蒸馏

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t, \epsilon} \left[w(t) \underbrace{(\hat{\epsilon}_{\phi}(\mathbf{z}_t; y, t) - \epsilon)}_{\text{让每个侧面的噪声预测尽可能符合预训练二维扩散模型的分布}} \frac{\partial \mathbf{x}}{\partial \theta} \right]$$

SDS: Score Distillation Sampling
(分数蒸馏采样)

让每个侧面的噪声预测尽可能符合
预训练二维扩散模型的分布





□ 主要难题：多面问题 (Multi-face/Janus problem)

- 由于二维大模型存在长尾分布，偏向于三维物体的每个侧面投影是正面

Sherpa3D
~25min



-30°

Shap-E
~10s



DreamFusion
~ 1h



Magic3D
~ 40min



Fantasia3D
~ 45min



150°



ProlificDreamer
~ 3h



"A DSLR photo of an adorable Corgi dog with a wagging tail"



□ 主要难题：多面问题 (Multi-face/Janus problem)

- CVPR 2024 清华提出Sherpa3D，通过三维结构引导二维大模型蒸馏的方式缓解多面问题，并保持高质量和三维一致性

□ 分析二维扩散模型 v.s. 三维扩散模型

■ 二维扩散模型：

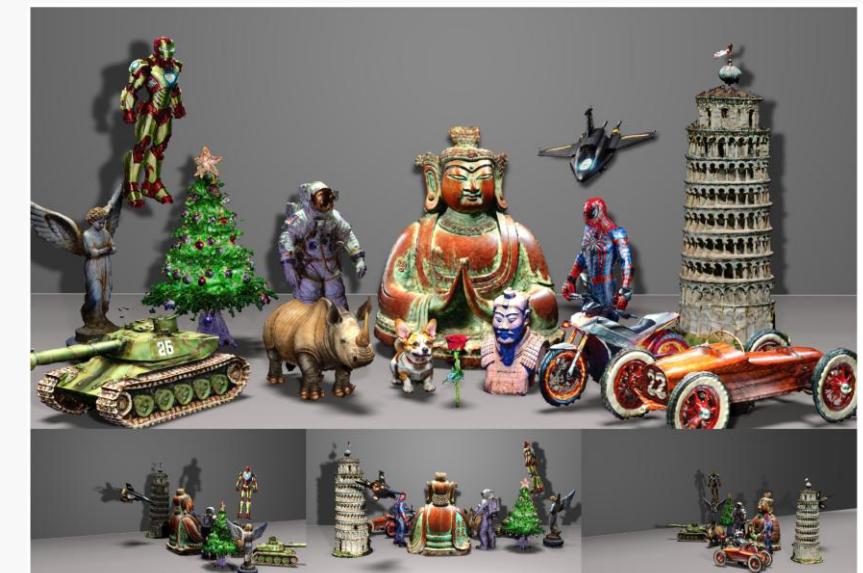
- 强泛化性，纹理细节丰富，但有多面问题

■ 三维扩散模型：

- 多视角几何一致性，但质量低且泛化性有限

■ Sherpa3D的目标：

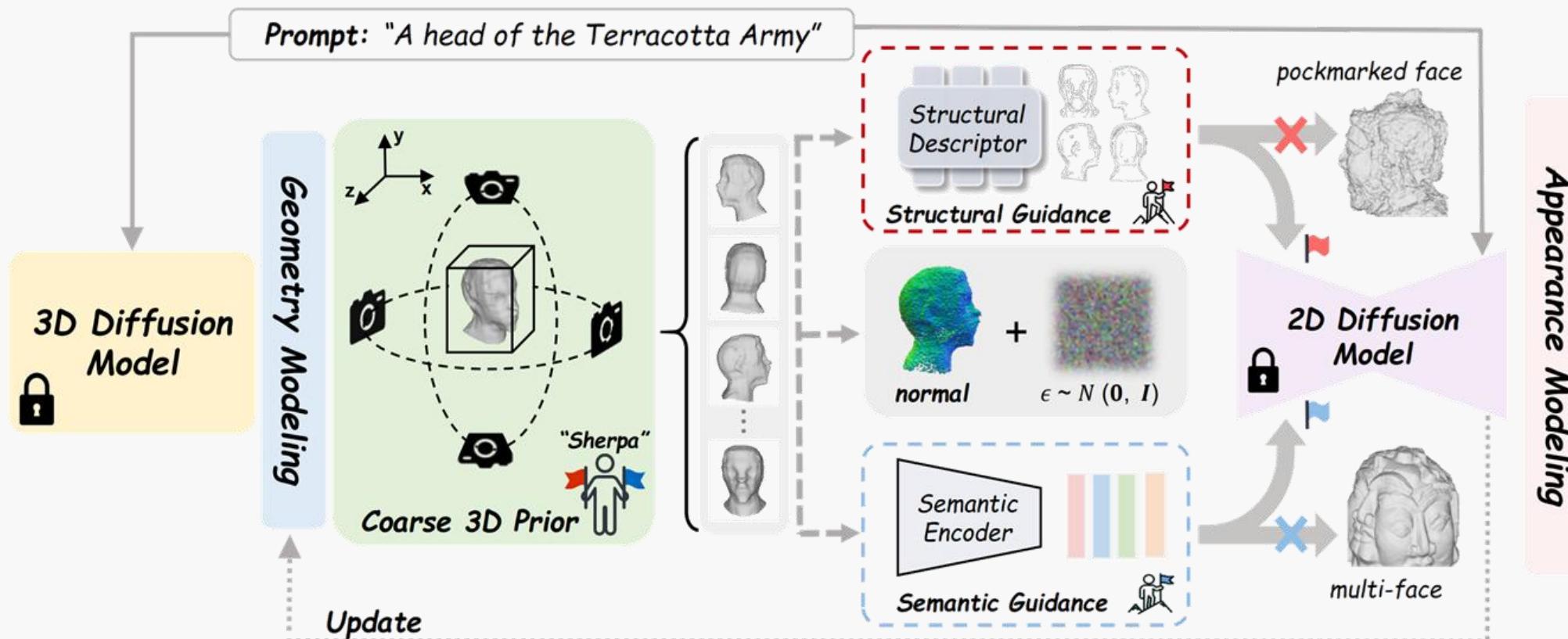
- 充分利用二维和三维模型的优势，生成高保真度，强泛化性，几何一致（即没有多脸问题）的三维物体





□ 主要难题：多面问题 (Multi-face/Janus problem)

- CVPR 2024 清华提出Sherpa3D，通过三维结构引导二维大模型蒸馏的方式缓解多面问题，并保持高质量和三维一致性





□ 主要难题：多面问题 (Multi-face/Janus problem)

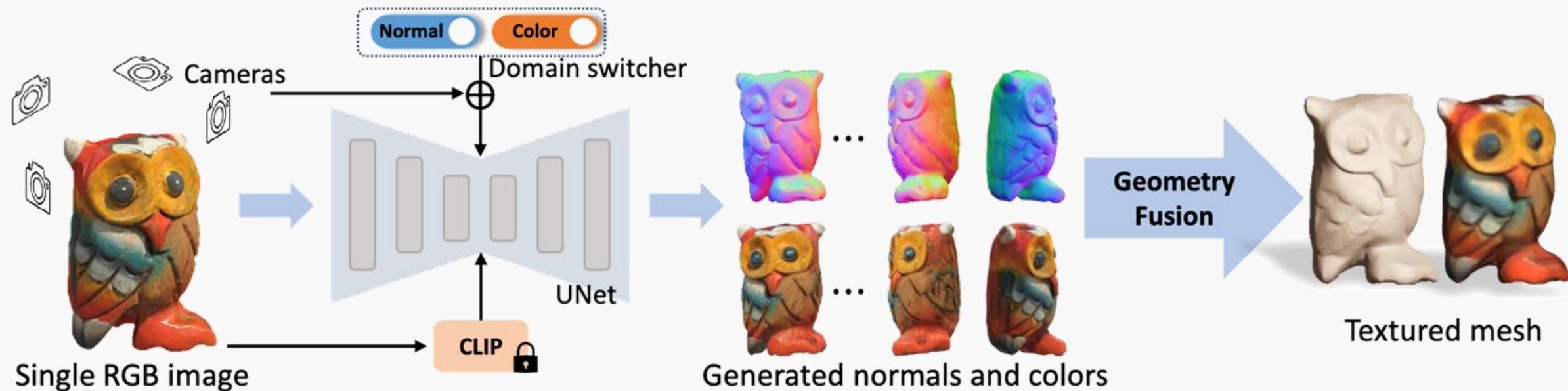
- CVPR 2024 清华提出Sherpa3D，通过三维结构引导二维大模型蒸馏的方式缓解多面问题，并保持高质量和三维一致性



多视角生成重建

核心想法：

- 先从单张图片/文本生成多视角（比如6个正交视角），然后根据多视角进行重建
- 单图生六视图：基于二维大模型（Stable Diffusion）**微调**的多视角扩散模型
- 多视图重建：可以基于SDF / NeRF / Mesh / Gaussian Splatting等方法



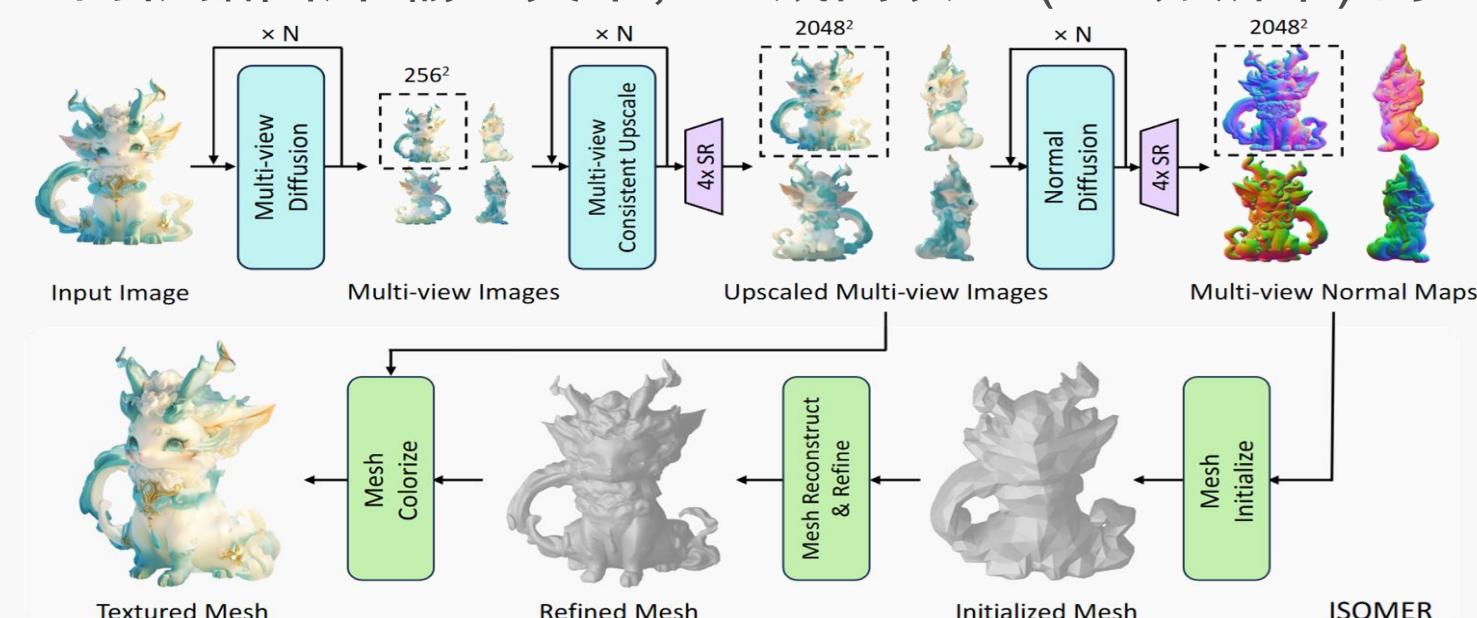
多视角生成重建

□ 主要难题：效率和质量的平衡

- 效率：由于两阶段，生成时间难以缩短至1分钟内
- 质量：由于SDF/NeRF等方法的限制，分辨率很难达到512以上

□ 清华大学NeurIPS 2024提出Unique3D

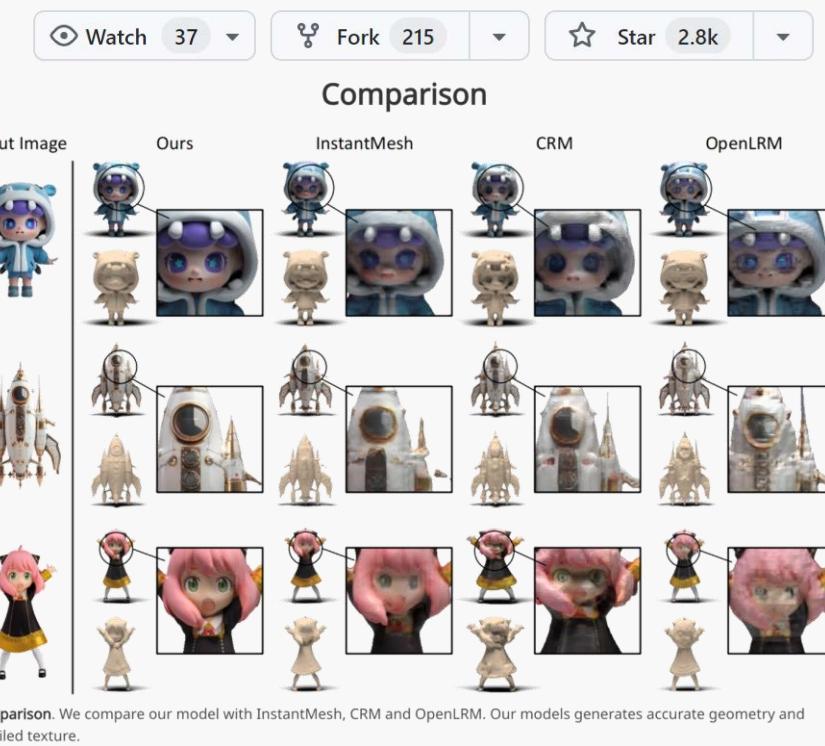
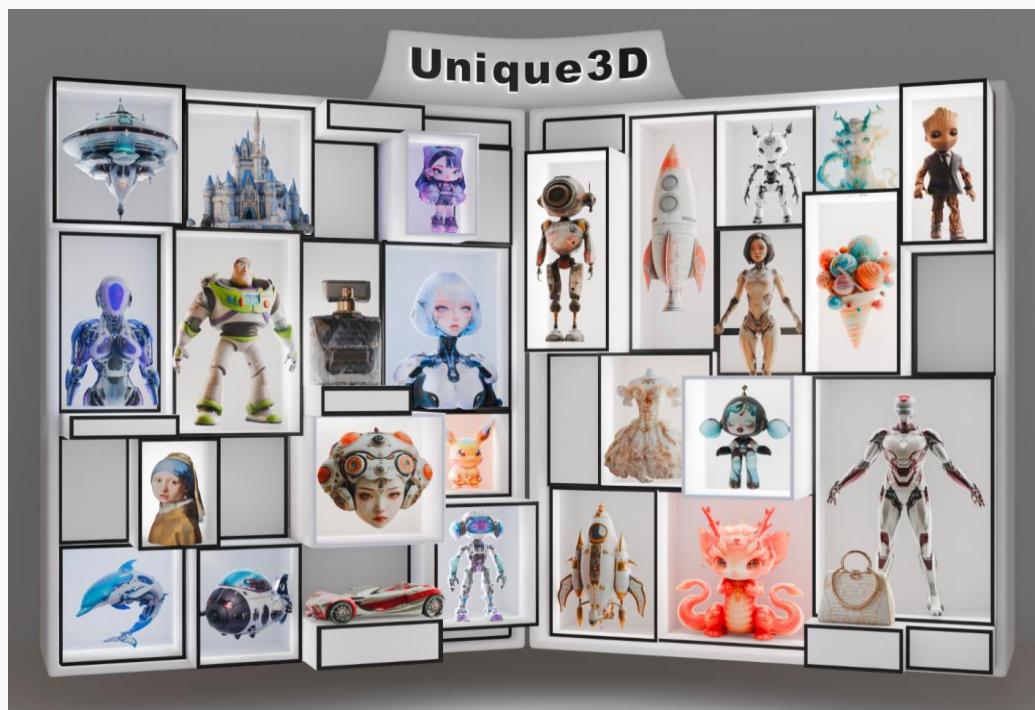
- 能够在30s内根据图片/输入文本，生成高质量 (2K 分辨率) 的三维物体 (Mesh)





□ 清华大学NeurIPS 2024提出Unique3D

- 能够在30s内根据图片/输入文本，生成高质量 (2K 分辨率) 的三维物体 (Mesh)
- 使用**高效率原生Mesh优化**代替NeRF/SDF等隐式表达
- 逐阶段的一致超分多视图将分辨率提升至2K





多视角生成重建

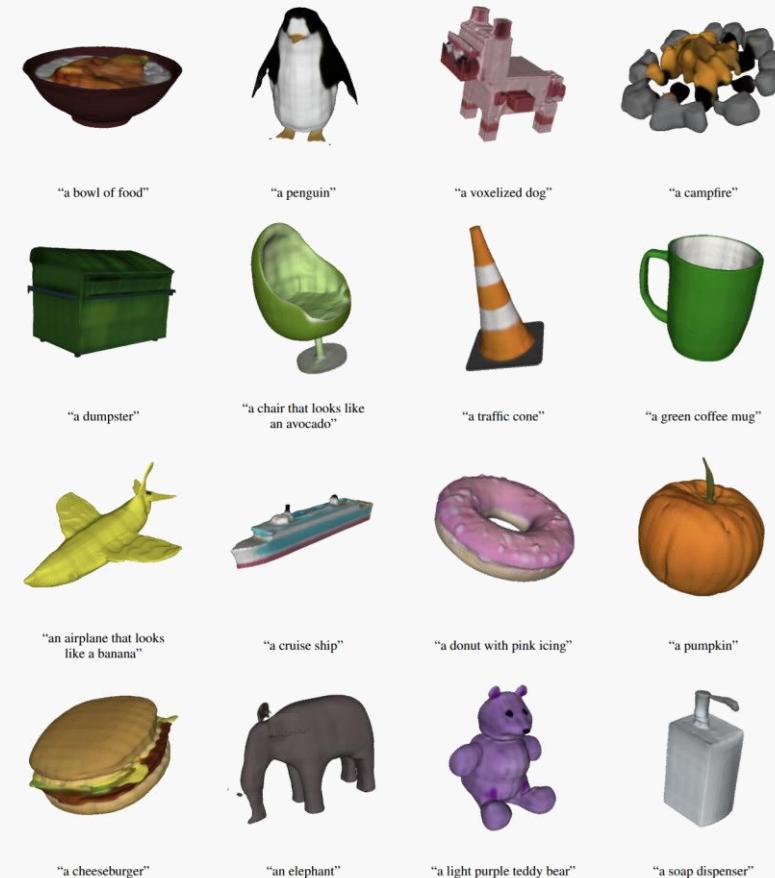
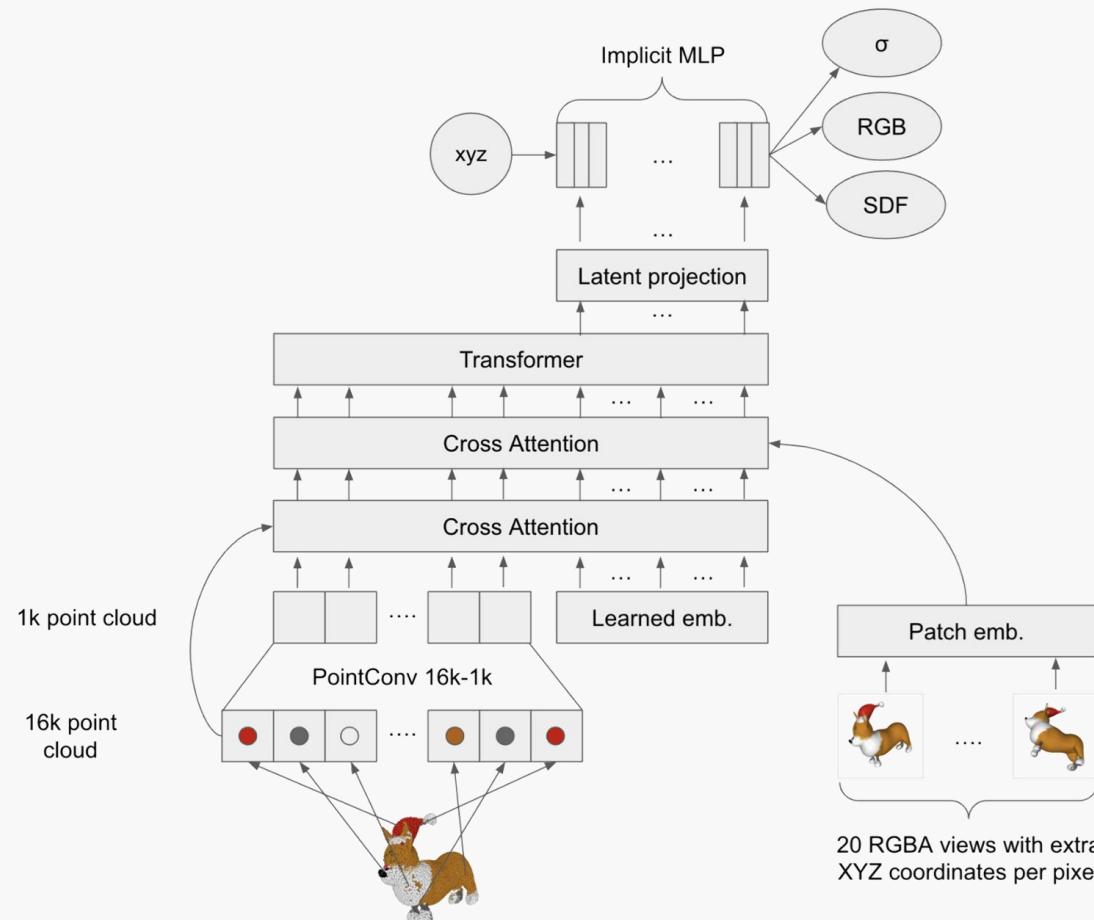
□ 清华大学NeurIPS 2024提出Unique3D





□ 核心想法：

- 大量三维数据驱动训练原生三维大模型，不需要经过中间2D的模态



- 企业 (Adobe / NVIDIA / Clay / Tripo / Meshy...) 私有3D高质量数据和10M+开源数据Objaverse-XL使得原生三维生成逐渐成为主流



原生三维模型

□ 原生三维模型经典范式

- 先训练VAE将数据压缩至隐空间 (Latent Space)
- 然后在隐空间中训练三维扩散大模型

