# 内容生成的新时代——3D AIGC

**Merhyi**

**摘要**

生成式人工智能 (Artificial Intelligence Generated Content, AIGC) 是人工智能领域中里程碑式的技术革新。其涉及文本、图像、语音、视频等诸多媒体形式，极大推动了生产关系的变革和社会生产力的发展。随着 VAE、GAN、Diffusion Model 等生成模型的出现，AIGC 在二维图像生成领域已经出现了井喷式的突破。而二维图像生成技术的成熟，呼唤着 AIGC 在三维建模的应用。本篇文章将从 3D AIGC 的产生背景出发，指出 3D AIGC 的技术基础，随后划分 3D AIGC 的主要流派，给出一些典型模型，并简要分析它们的原理。

## 1 技术背景

### 1.1 需求和供给的冲突

3D 内容于未来占据内容市场将是不可逆转的趋势。3D 内容无论是在展示事物本身的结构与形体，还是在表现物体的运动，抑或是人机交互等领域具有 2D 内容不可比拟的优势。

在展现事物结构的方面，3D 内容能够帮助人类更大程度地还原真实世界的环境以及物体。在建筑设计领域，只有 3D 内容才能展现房屋每个具体的部件对其整体结构以及物理性态的影响；在科学研究方面，3D 内容能更加完整地表示实验对象以及模型的各个性质；而场景搭建的真实性则更加需要 3D 内容构建技术的参与。

在表现物体运动的方面，3D 内容相较于 2D 有过之而无不及。其独特的骨骼模型以及对物理引擎和动作引擎的适配性，使得三维模型在动作方面更加符合现实逻辑；同时，配合 3D 内容独有的光线追踪和渲染技术，物体的运动和光线的映衬使得其展现的场景更加真实。

而在人机交互领域，AR（Augmented Reality，增强现实）技术和 VR（Virtual Reality，虚拟现实）技术与 3D 内容的结合，使得参与者能够亲身体验到来自听觉、视觉、触觉等多模态的体验，这些都是 2D 内容所无法做到的。

由此可见，3D 内容在当今的内容产业，尤其是电影业及游戏业的需求量将会只增不减，甚至有逐渐取代 2D 内容之势。

然而，如今 3D 内容的生产仍然面临着生产效率低下和生产成本高两大困难。

一方面，静态的 3D 制作相比 2D 就已经增加了摄像机角度的调节、光线角度调节等环节；对于 3D 动画，则需要经历建模、贴图、绑定、骨骼动画、特效、渲染等诸多步骤，而这些对于制作人员的技能以及流程间的协调有了更高的要求。繁杂的环节步骤使得工期大大延长，人工生产效率低下。

另一方面，3D 内容的生产有两大主要方式。第一是利用专业工具进行生产，如 Autodesk Maya、Houdini、Blender 等，在进行建模、贴图、渲染、修饰等过程中，需要大量的在生产软件中切换的操作，这不仅对生产者的技能有了更高的要求，其操作的繁杂本身也增加了时间成本；第二则是通过硬件扫描现实世界的实物进行移植。这种方式通常需要成本高昂的激光雷达扫描仪、数据处理的软件及设备等，而且场地往往受限，机动性差。不仅如此，在扫描导入之后，仍需要进行大量的修饰以及改善才能真正投入使用，这无疑又大大增加了生产的成本。

在 3D 内容的需求和生产之间日益激烈的矛盾下，3D AIGC 技术应运而生。

## 1.2 今天的 3D AIGC 产业

今天的 3D AIGC 产业虽仍处于萌芽阶段，但就其行业本身，新的模型、论文层出不穷，许多创业公司从不同方向出发，形成了相对完整的一个产业体系。现今的 3D AIGC 产业主要由以下几个部分组成：

**底层模型生产者** 这些公司是行业的基础设施提供商。它们主要着力于构建通用的 3D AIGC 模型。目前的底层模型生产者主要是一些大型公司，比如 Openai、NVIDIA、Google、Microsoft、Stability.ai 等，它们拥有充足的资金以及大量的研究人员，同时具有几十年对 AI 方向的投入与研发经验。

**3D 资产拥有者** 这些公司拥有大量的 3D 素材以及 3D 资产。它们主要位于游戏和动画两个产业中。它们有的负责外包建模，有的则是 3D 素材的双边贸易市场。

**3D 扫描公司** 这类公司同样拥有大量的 3D 素材和资产，但不同的是它们还构建或拥有属于自己或者自己所属行业的硬件扫描 3D 成像技术。

**游戏领域的 3D AIGC 生产者** 这类公司专注于在游戏行业为 AIGC 的整个产业提供素材和资源，诸如游戏场景、模型贴图、3D 原模等。大部分这类公司是创业公司，目前正处于萌芽期。

**泛 3D AIGC 生产者** 这些公司包括了试图通过技术突破来生成 3D 模型的 AIGC 创业者，也有一些利用 AIGC 技术构造 3D 的 UGC（User Generated Content，用户原创内容）的公司。


图 1: 现今 3D AIGC 的产业地图

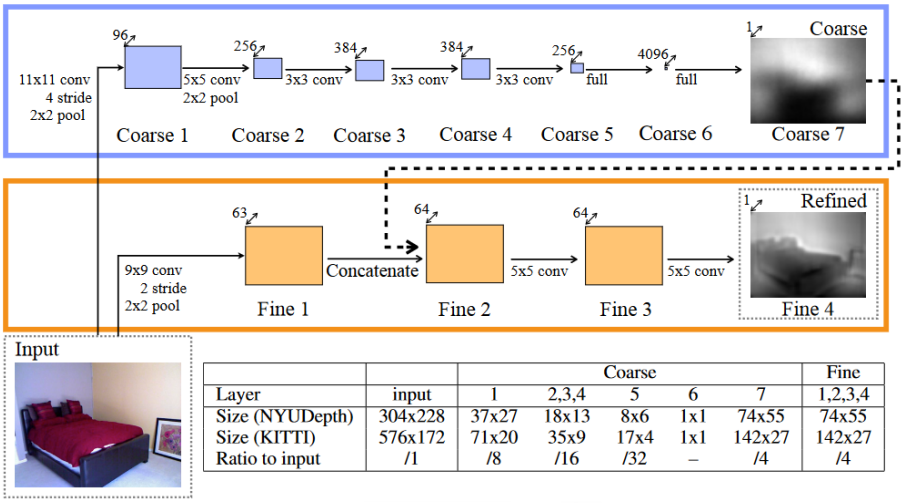
# 2 相关技术

在讨论主流模型之前，先介绍与 3D AIGC 生产相关的两大技术：3D 重建技术和生成模型。

## 2.1 与 AIGC 相关的 3D 重建技术

**基于体素的显式表达** 体素（Voxel）是三维模型中最简单的形式。通过将二维的卷积拓展到三维的形式，就能实现最简单的三维重建。David Eigen 等人首次利用深度学习方法进行三维重建，使用单张图片，通过神经网络来直接恢复深度图，并将神经网络分为全局的粗略估计和局部的精细估计形式，同时采用一个尺度不变的损失函数进行回归，从而实现简单的三维重建。[1]

Christopher 等人提出的 3D-R2N2 模型则通过 Encoder-3D-LSTM-Decoder 的网络结构建立了从 2D 图形到 3D 体素模型的映射，从而实现了基于体素的单视图多维重建。[2]




| | | Coarse | | | | | Fine |
|---|---|---|---|---|---|---|---|
| Layer | input | 1 | 2,3,4 | 5 | 6 | 7 | 1,2,3,4 |
| Size (NYUDepth) | 304x228 | 37x27 | 18x13 | 8x6 | 1x1 | 74x55 | 74x55 |
| Size (KITTI) | 576x172 | 71x20 | 35x9 | 17x4 | 1x1 | 142x27 | 142x27 |
| Ratio to input | /1 | /8 | /16 | /32 | – | /4 | /4 |

图 2: David 等人提出的模型结构

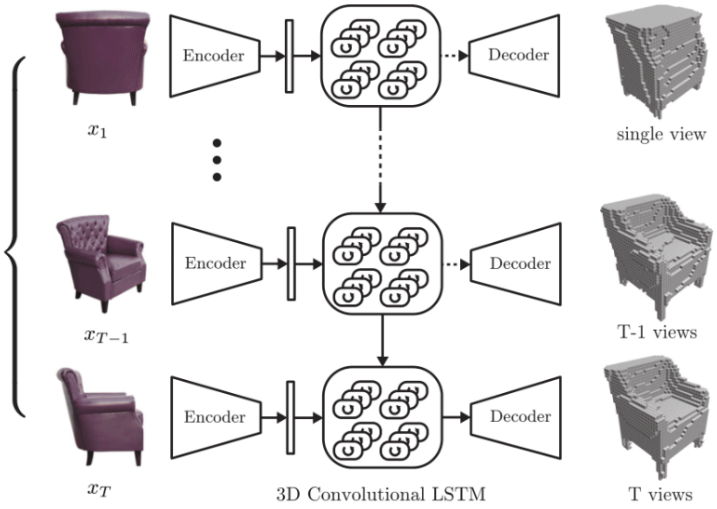



图 3: Christopher 等人提出的 3D-R2N2 模型

**基于点云的显式表达** 基于体素的方法如果要提升分辨率，则会大量增加三维卷积的计算量。相较之下，点云（Point Cloud）就是一种更加简单、更易学习的结构。另外，由于点之间的连接性在几何变换和变形时不需要进行更新，在一定程度上减少了计算量。

Leonidas Guibas 等人首次利用点云结构进行三维重建。他们选择了恰当的损失函数来进行测量，从而解决了训练点云网络时由相同近似程度的相同几何形状可通过不同点云表示所导致的损失问题。[3]

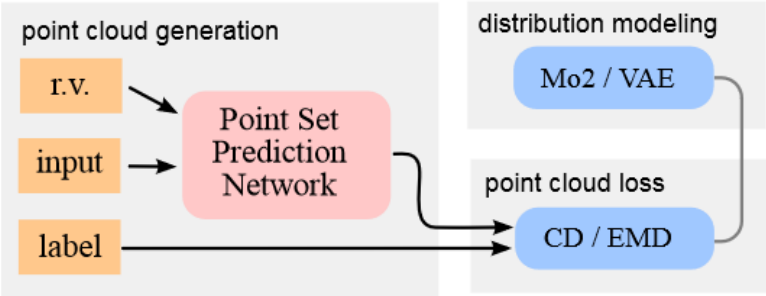



图 4: Leonidas 等人提出的 PointOutNet 系统结构

Rui Chen 等人通过对场景内的点云进行处理，进一步融合三维的深度以及二维的纹理信息，从而提高了点云的重建精度。[4]

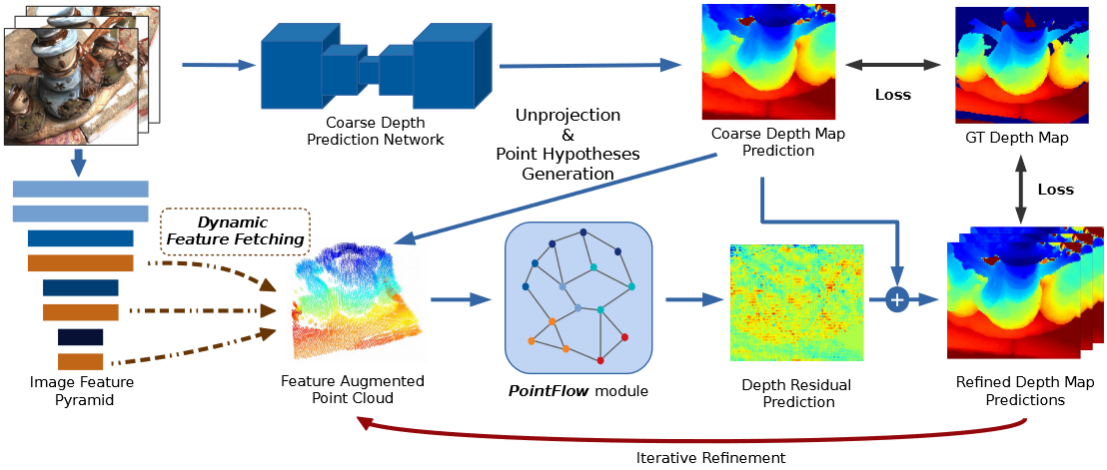



图 5: Rui Chen 等人提出的 Point-MVSNet 框架

**基于网格的显式表达** 然而，点云缺少连接性的特点在给其带来计算量的便利的同时，也丢失了物体表面的信息，进而会导致重建物体的表面不平整的问题。相较于体素和点云，基于网格（Mesh）的方法则兼有计算量小和连接性强的特点，能够展示物体的更多细节。

主要的基于网格进行 3D 重建的方法是由 Nanyang Wang 等人于 2018 年提出的 Pixel2Mesh 算法。它首先将任意的输入图像都初始化成一个椭球体，作为重建的初始三维形状。接着，它将训练的神经网络分为两个部分，一部分使用全卷积神经网络，用于提取输入图像的特征；另一部分则通过图卷积网络来表示三维的网格结构。最后，它将三维网格进行修饰和变形，最终输出物体的形状。这个模型利用了四个不同的损失函数来约束形状，通过端到端的神经网络实现了从单张彩色图片生成用网格表示的三维物体的重建操作。[5]

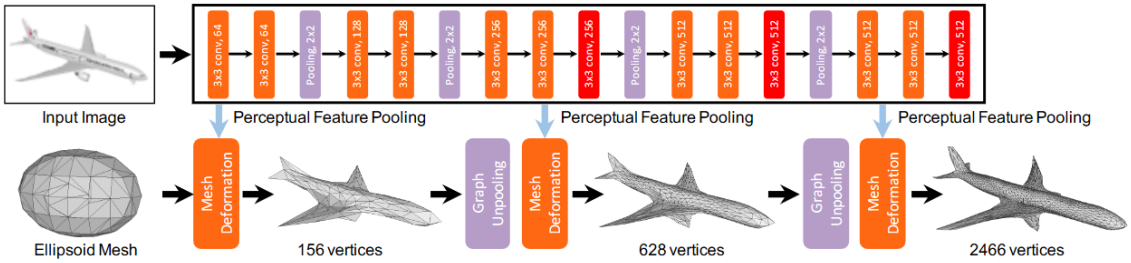



图 6: Nanyang Wang 等人提出的 Pixel2Mesh 模型

**基于辐射场的隐式表达** 基于神经辐射场（Neural Radiance Fields，NeRF）的隐式建模方法由 Ben Mildenhall 等人提出，它解决的是给定某一物体的多视角图片，重构出该物体的三维表示的问题。该方法主要利用神经网络构造五维的辐射场函数，接着运用体渲染（Volume Rendering）技术，将给定视角的辐射场渲染成单张图像。[6]

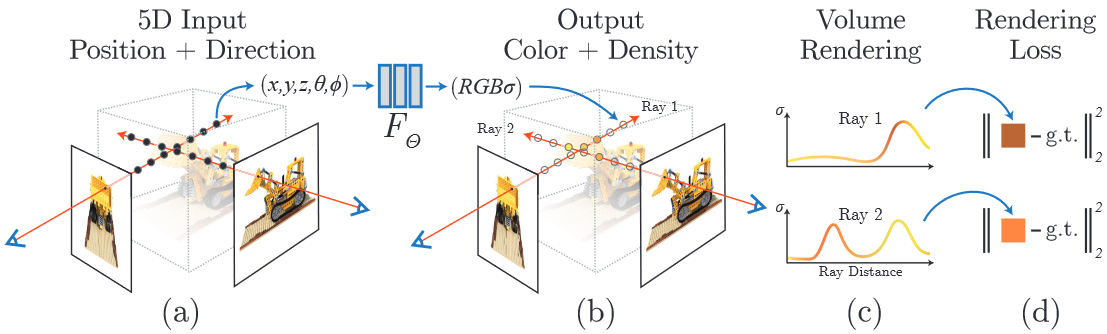



图 7: Ben Mildenhall 等人提出的 NeRF 方法

由 NVIDIA 提出的快速 NeRF 技术（Instant-NGP）是当前较为流行的基于辐射场进行建模的技术。在高质量场景的建模中，往往需要构造一个较大的网络结构，若将每个采样点都进行遍历，则会大大拖慢运行速度。Instant-NGP 和原始的 NeRF 方法最大的差别在于其用稀疏的体素格栅结构（Voxel Grid）来表达场景，进而优化了执行速度。[7]

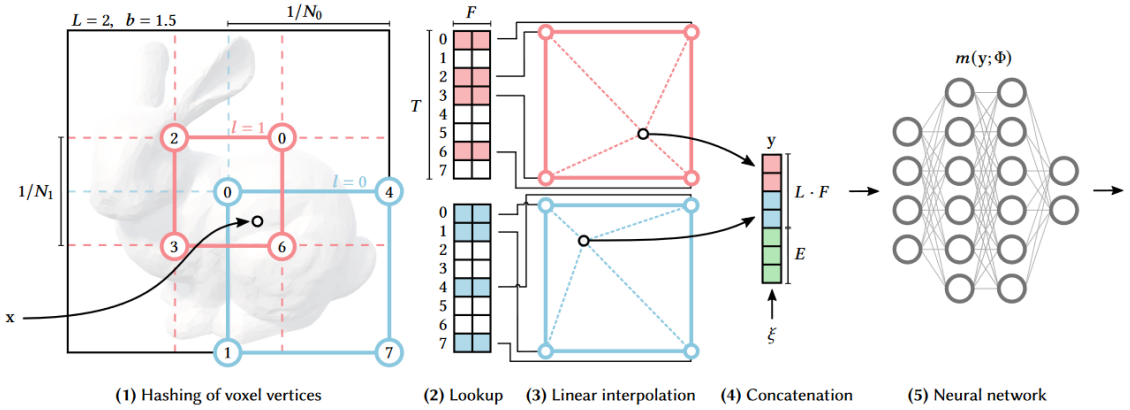



图 8: Instant-NGP 的多分辨率哈希编码过程

**DMTet** 深度移动四边形算法（Deep Marching Tetrahedra，DMTet）是由 NVIDIA 提出的，基于移动四边形算法（Marching Tetrahedra，MT）的深度学习版本。它以夹杂噪声的点云或者粗略的体素模型作为输入，首先基于深度学习方法预测有向距离场（Signed Distance Field，SDF）隐式地精细化模型形体，接着利用移动四边形算法来生成模型的显式表面网格，得到模型的参数表面（Parametric Surfaces）。这一方法能够结合隐式和显式的表达方式，且相较于 NeRF 具有较快的运行速度。[8]

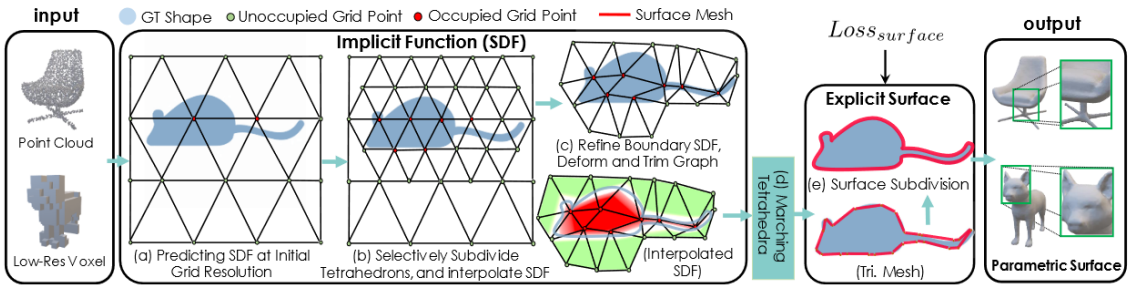



图 9: DMTet 的流程图

## 2.2 与图像生成相关的生成模型

**VAE** 变分自编码器（Variational Auto-Encoder，VAE）是一种使用了正则化方法的自编码器（Auto-Encoder），用于解决编码器（Encoder）部分对输入数据编码时产生的过拟合问题。由于最终生成内容需要解码器（Decoder）从隐空间的分布中采样获得，VAE 采用了将每张输入内容编码成分布的方法，来让解码形成的各个子分布能够尽可能拼接成期望的分布；并在损失函数中添加对子分布的约束，即 KL 散度（Kullback-Leibler divergence），让它们尽可能接近标准正态分布。[9]

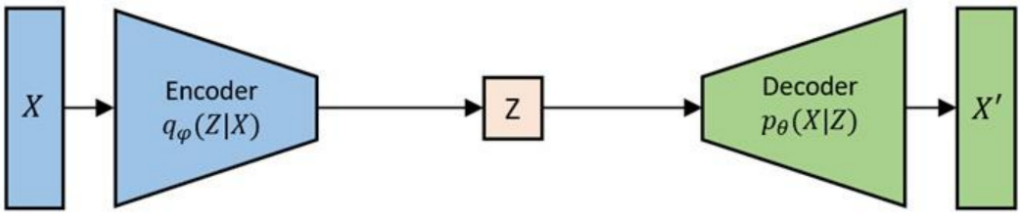



图 10: VAE 的自编码器结构

**GAN** 生成式对抗网络（Generative Adversarial Networks，GAN）是一种处于对抗博弈，相互促进状态的生成模型。模型的主要结构由两部分组成：生成器（Generator）和判别器（Discriminator）。生成器将接收一个随机噪声 $z$，并通过噪声生成图片。而判别器则用于判别生成器生成的图片是否是真实的：它首先接收一张真实图片 $x$，并计算这张图片的真实概率 $D(x)$；接着将判断由生成器构造的图片 $G(z)$，计算概率 $D(G(z))$。$D(x)$ 与 $D(G(z))$ 越接近，则说明生成器生成的图片真实性越强。[10]

在不断的博弈过程中，生成器和判别器都会受到训练。最终，生成器生成的图片能够“欺骗”判别器，从而能够生成“以假乱真”的期望图片。

随着进一步的研究，GAN 的优化模型逐渐被提出。由 Goodfellow 提出的 DCGAN 模型将深度卷积神经网络（Convolutional Neural Networks，CNN）和 GAN 相结合；[11] 由 Tero Karras 等人提出的 PGGAN 模型采用一种独特的渐进提升方式（Progressive Growing）来训练 GAN，通过逐渐增大训练网络的规模，提高并稳定训练速度的同时，大大提高了生成图像的质量；[12] BigGAN 由 Andrew Brock 等人提出，其基于 ImageNet 数据集，且具有令人惊叹的生成质量；[13] StyleGAN 同样由 Tero Karras 等人创造，其借鉴了自适应实例标准化（AdaIN）机制来控制潜在空间向量。[14]

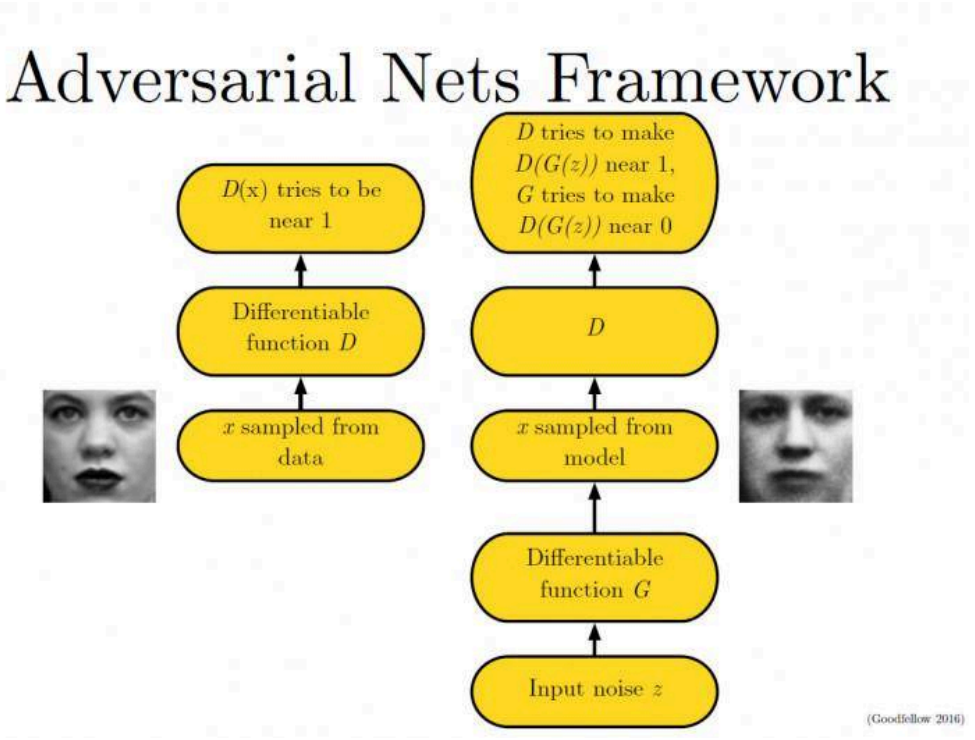



图 11: GAN 的博弈结构

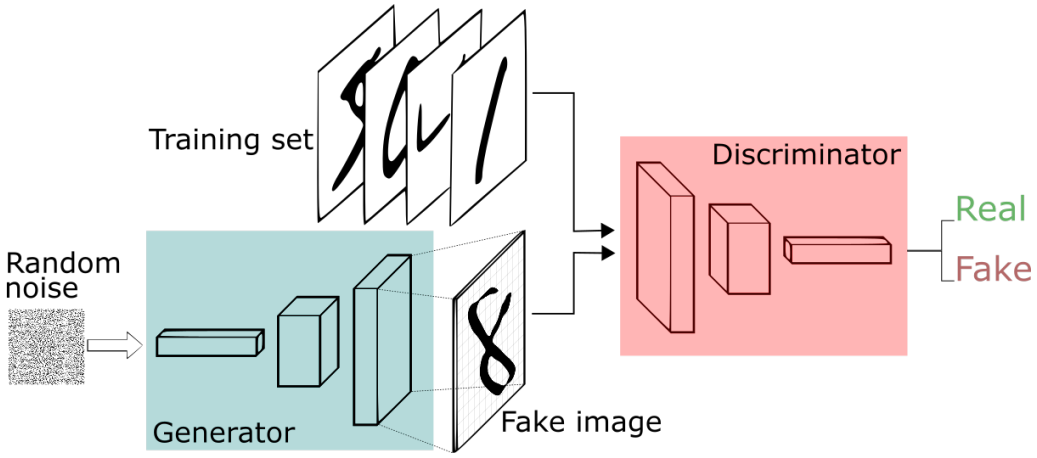



图 12: GAN 的整体结构

**Diffusion** 扩散模型（Diffusion Model）最早由 Jascha Sohl-Dickstein 等人提出，将热力学与统计力学中扩散的概念迁移到图像处理领域。[15] 而目前主要应用的扩散模型是 Jonathan Ho 等于 2020 年引入的去噪扩散模型（Denoising Diffusion Probabilistic Models，DDPM）、Robin Rombach 等于 2021 提出的 Latent Diffusion 模型以及 2022 年由 Stability AI 基于此研发的 Stable Diffusion 模型。

DDPM 主要分为两个过程，前向过程（Forward Process）和反向过程（Reverse Process）。这两个过程都由一条参数化的马尔科夫链（Markov Chain）相连，前向过程向给定的图片一步步添加高斯噪声，最终生成完全的随机噪声图片；反向过程则是一个去噪声的过程，将给定的噪声一步步减噪为原始图片。DDPM 首先选取原图，对其加噪形成原始噪声；接着随机生成噪声并计算二者的损失函数。通过训练减少损失函数值，DDPM 对噪声的预测会逐渐准确，最终能够形成较为清晰的图片。[16]

为了减少高维图片采样时巨大的计算量，Latent Diffusion 引入了隐空间（Latent Space）的概念。模型的编码器首先将图片进行压缩（即进入隐空间），并在隐空间中完成扩散模型的训练与生成工作；接着，再通过解码器将隐空间中生成的图片解压缩至原来的大小。[17]

Stable Diffusion 是 Latent Diffusion 在文字生成图片（Text-to-Image）领域的应用。它通过引入 Open AI 研发的 CLIP 模型进行文字语义和图像特征理解，进而通过提供的文本信息生成图片。

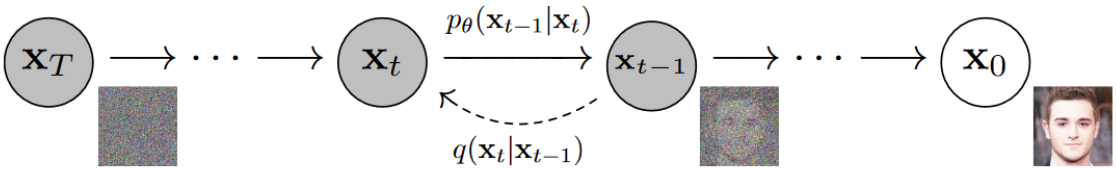



图 13: DDPM 的原理

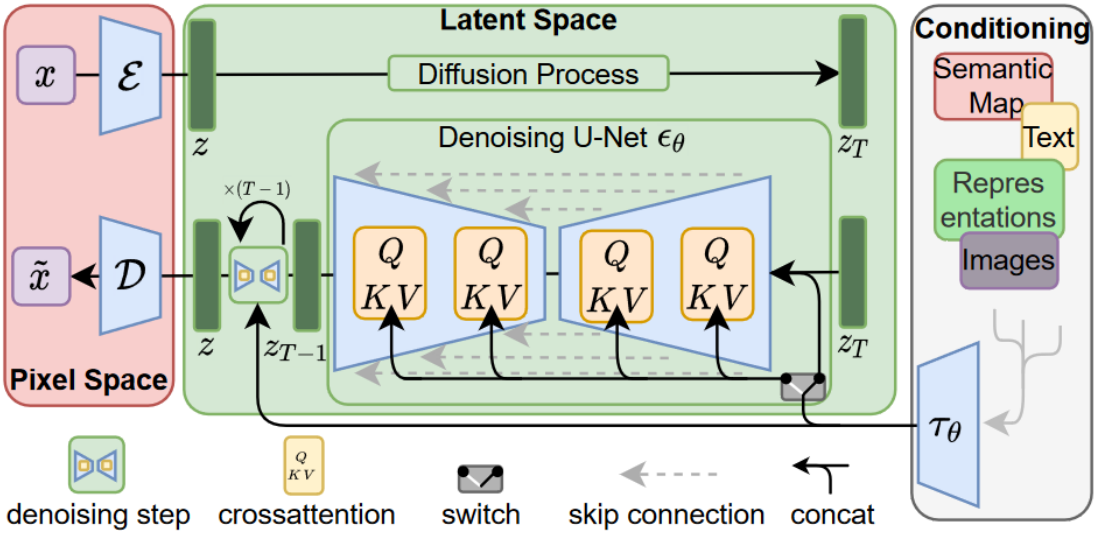



图 14: Latent Diffusion 的结构

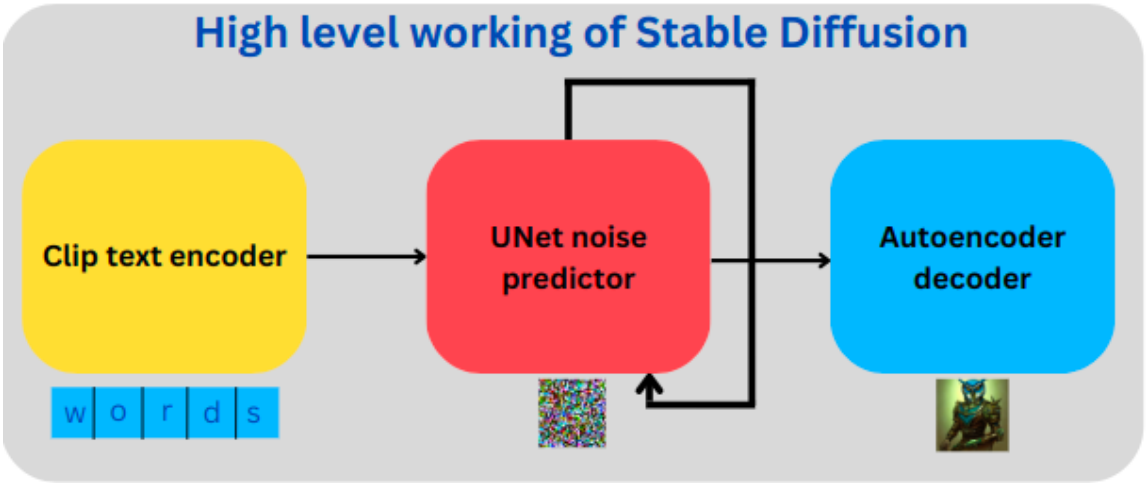



图 15: Stable Diffusion 结构图

# 3 3D AIGC 流派及模型

接下来将介绍目前 3D AIGC 的一些流派以及它们的代表模型，并简要介绍它们的原理。

### 3.1 仿 3D 派

**Zero-1-to-3** Zero-1-to-3 由 Ruoshi Liu 等人提出，它可以通过某一视角的物体图片生成另一视角下的同一物体，甚至对物体进行三维建模。这一模型通过微调其依赖的 Latent Diffusion 模型，使其能够学习对不同摄像机相对视角的控制和识别，从而在大量的数据集训练基础上，它能够生成改变摄像机观察角度的图像。它还能够与 NeRF 结合，对物体进行进一步的三维建模。[18]

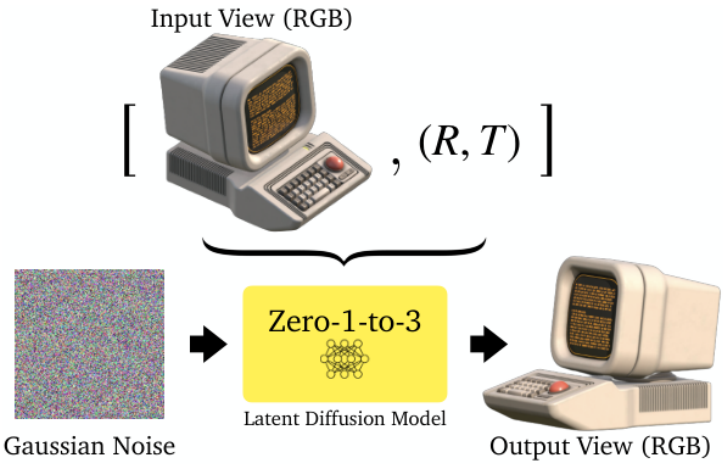



图 16: Zero-1-to-3 的流程图

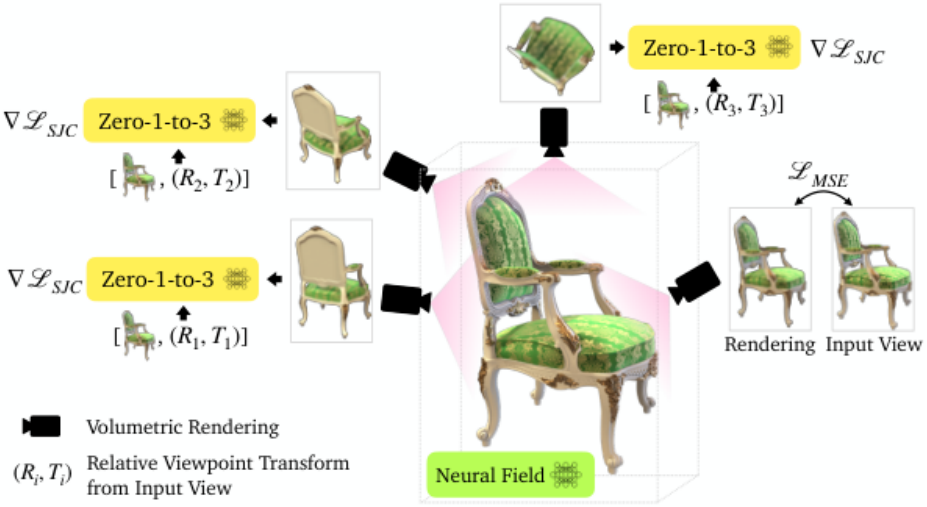



图 17: Zero-1-to-3 与 NeRF 的结合

**基于 Diffusion 的三维感知模型** 这一方法由 Jianfeng Xiang 等人引入，它基于二维的 Diffusion 模型，能够生成一个物体的多视角二维图像。它主要由两个 Diffusion 模型组成，其中一个是条件扩散（Conditional Diffusion），一个是无条件扩散（Unconditional Diffusion）。无条件模型用于随机生成第一个视角的视图，而条件扩散模型则用于进一步生成其它视角的图像。这一方法在大型数据集 ImageNet 的训练下，能够 360° 无死角地形成感知图像。[19]

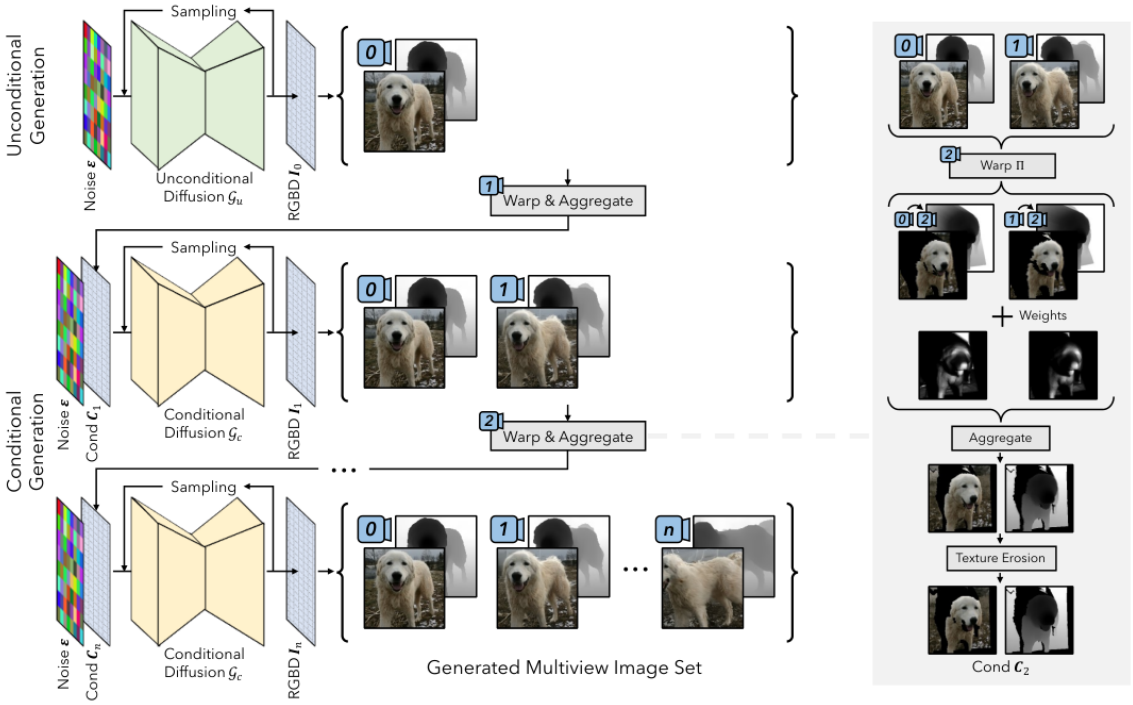



图 18: 基于扩散模型的 3D 感知生成

## 3.2 转化 3D 派

**GANCraft** GANCraft 由 Zekun Hao 等人引入，提供了一种娱乐性强的三维生成方式，能够解决数据集不足的情况下的生成困难问题。该方法基于生成式对抗网络 SPADE（一种经过改进的、适用于三维生成的 GAN 模型），输入来自 Minecraft 游戏地形的投影分割图，再利用 SPADE 将其转化为伪地面真实图像（Pseudo-ground Truth Image）。[20]

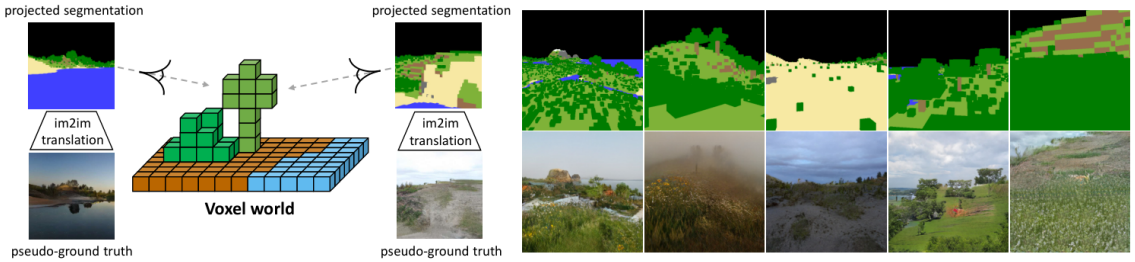



图 19: GANCraft

**MPI** 多平面图像（Multiplane Image）是一种生成思想，它主要通过输入多张表达同一场景不同深度内容的图片，对摄像机的视锥（Frustum）进行三维重建。这一方法与深度学习的结合产生了比如 MINE 等模型。Bo Zhang 等人基于这一思想提出了一种由身体姿势引导的多平面图像合成技术，用于在真实的场景中呈现角色的动作，进而生成动画。[21]


(a) Input images (b) Pose guided novel view synthesis (c) Motion transfer

图 20: 由身体姿势引导的 MPI 合成

## 3.3 2D 升维派

**Dream Fields** Dream Fields 由 Google 在 2022 年提出，它的主要原理是将文字生成图片模型与三维重建模型进行结合，即将 CLIP 模型与基于辐射场的三维建模 NeRF 结合。首先依托 CLIP 引导 NeRF 进行三维重构，再通过 CLIP 判断生成模型的多视角图片是否

符合文本描述的特征，经过反复调节，最终形成期望的三维内容。[22]

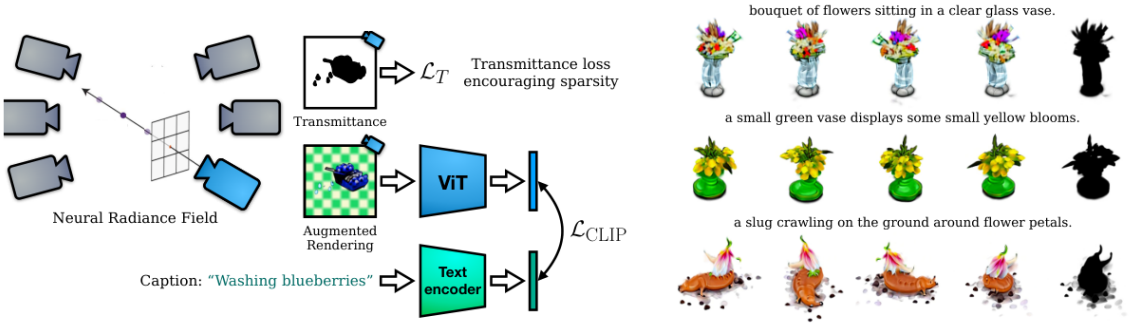



图 21: Dream Fields 的原理

**DreamFusion** DreamFusion 同样由 Google 研发，可以认为是 Dream Fields 的改进版本。它利用了 Google 自研的文本生成图像模型 Imagen 生成期望物体的多视角图片，接着通过 NeRF 对多视角图片进行 3D 重构，形成三维模型。[23]

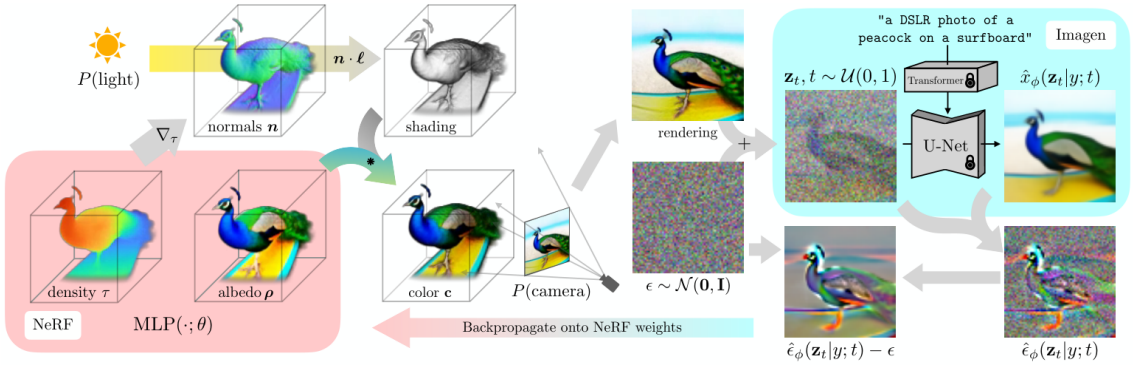



图 22: DreamFusion 的结构图

**CLIP-Mesh** CLIP-Mesh 由 Nasir Mohammad Khalid 等人提出，是结合 CLIP 模型和基于网格的三维重建技术的模型。它将初始椭球网格进行切割细分，修改其材质以及贴图，并输出多视角图；接着，由 CLIP 模型进行判断，计算损失函数；将这一过程反复迭代，使得网格模型逐渐接近期望的物体。[24]

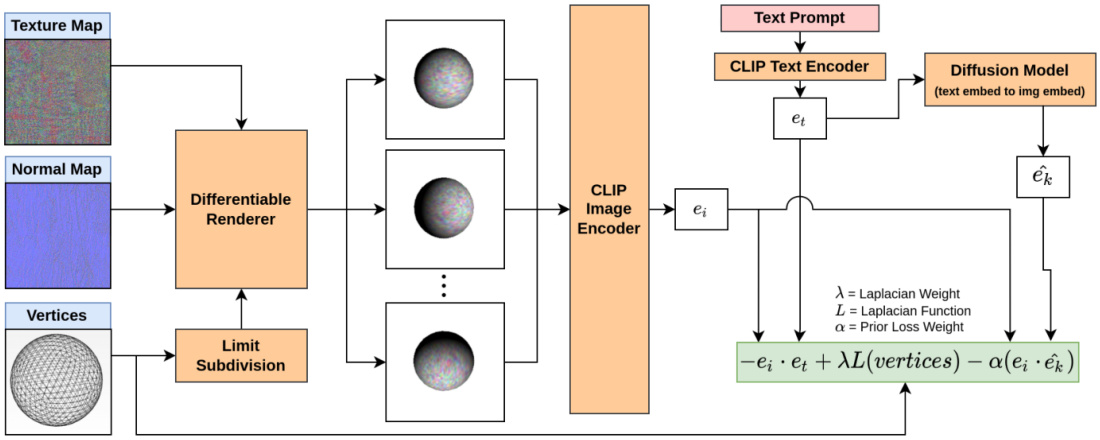



图 23: CLIP-Mesh 的流程图

**Magic3D** Magic3D 是由 NVIDIA 研发的 3D 生成模型。它的生成过程主要分为两个阶段：第一阶段，利用优化过后的自研模型 Instant NGP，通过重复采样以及渲染低分辨率图像的方法，生成初始的 3D 模型。接着，使用 DMTet 方法将其变为初始的 3D 网格模型，作为第二阶段的输入值。在第二阶段，结合 Latent Diffusion 方法，在隐空间中进行采样、渲染以及修饰，最终输出高分辨率的 3D 网格模型。[25]

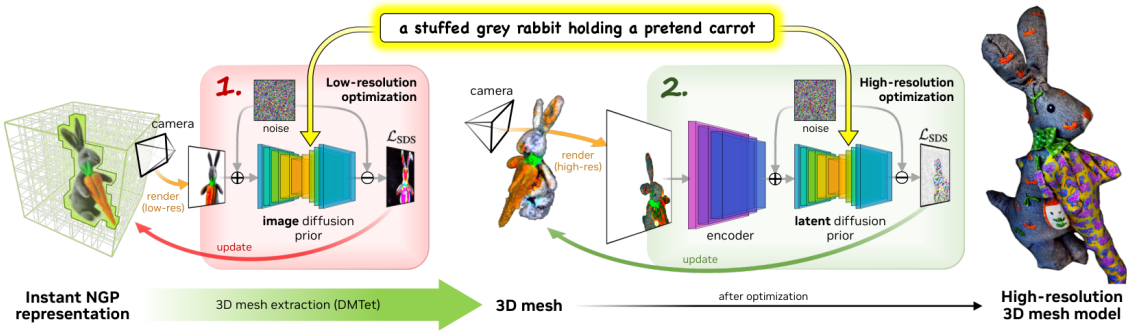



图 24: Magic3D 的流程图

**MAV3D** MAV3D 是由 MetaAI 研发的模型。它提出了一种基于描述文本来生成三维的动态场景的方法（Make-A-Video3D），这一方法利用四维的动态神经辐射场（4D NeRF），并且借助文本生成视频（Text-to-Video）模型，来优化生成的场景的外观、密度以及运动一致性。它不需要任何的三维或者四维数据既可以进行训练。[26]

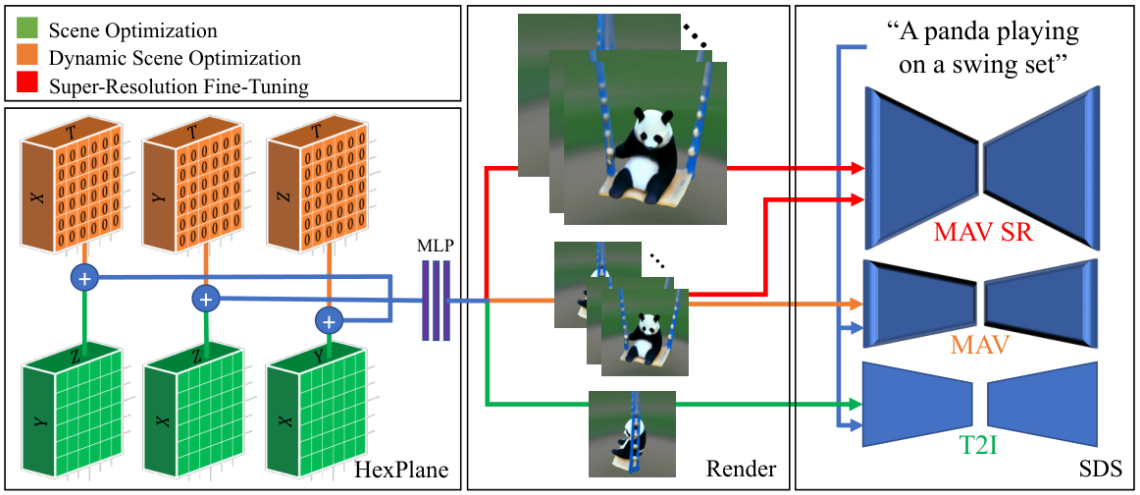



图 25: MAV3D 的流程图

**Make-It-3D** 由 Bo Zhang 等人提出的 Make-It-3D 模型基于 2D 的 Stable Diffusion，能够将一张二维的图片升维成三维的精细模型。整个过程分为两个阶段：粗略阶段（Coarse Stage）和修饰阶段（Refine Stage）。在第一阶段，先通过 NeRF 将二维图片升维成粗略的三维模型，并且保留它的几何结构与预估深度图；在第二阶段，先通过原始图片和三维模型生成材质点云（Textured Point Clouds），接着优化不可见点的纹理细节，再通过一个可学习的延迟渲染器来生成最终精细化的模型。[27]

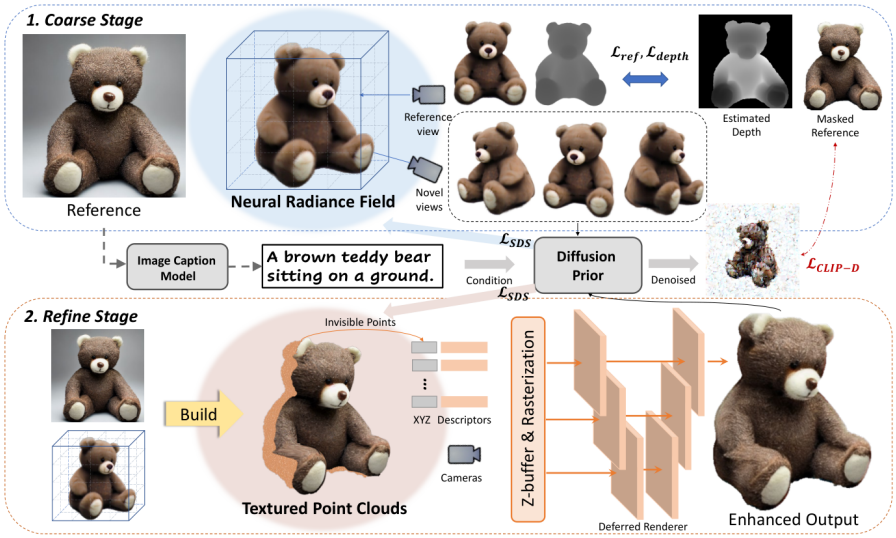



图 26: Make-It-3D 的流程图

## 3.4 原生 3D 派

**GET3D** GET3D 是由 NVIDIA 提出的模型。该模型主要由两个 GAN 结构组成，分别用于生成模型的网格结构以及材质。在这种模式下，可以对模型和贴图提出不同的要求，对物体的形状、颜色等特征分别进行限制。[28]

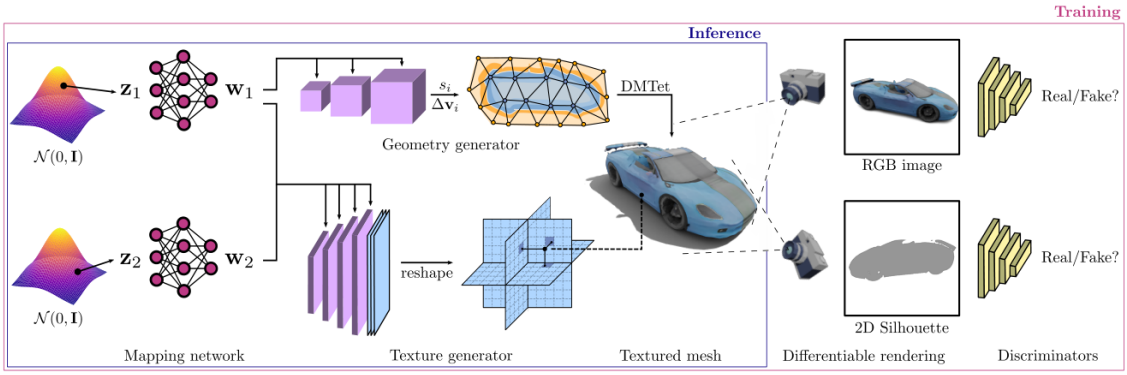



图 27: GET3D 的流程图

**Point · E** Point • E 是由 OpenAI 研发的基于点云的生成模型。它的工作分为两个部分：首先，使用文本生成图像模型将给定的文字转化为图像内容；接着，通过基于点云的三维重建技术对二维图像进行升维重构。[29]

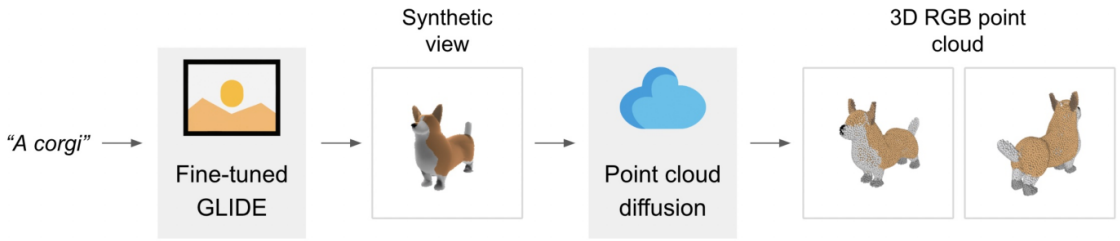



图 28: Point • E 的流程图

**Rodin** Roll-out Diffusion Network，即 Rodin，由微软亚洲研究院研发，它首次利用三维数据来训练扩散模型，并使用扩散模型直接生成 3D 虚拟人形象。它利用 NeRF 将三维空间分割成三个互相垂直的特征平面（Triplane），将给定图像展开到单个二维的特征平面，再进行感知扩散，通过将三维空间进行二维特征展开，Rodin 实现了利用二维架构进行三维感知扩散的功能。[30]

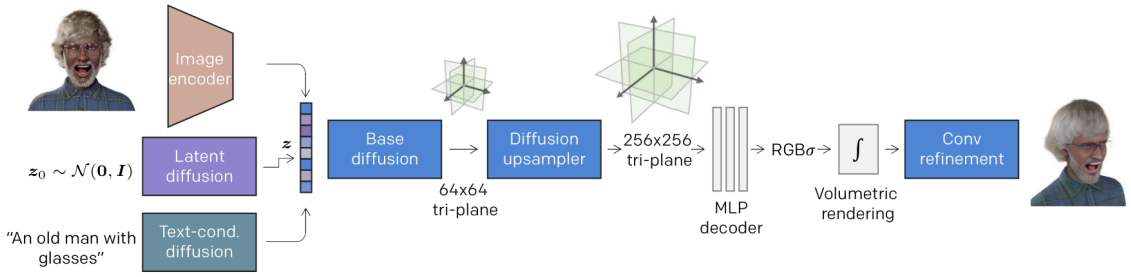



图 29: Rodin 的流程图

## 3.5 工具派

**仅使用合成数据的分析** Erroll Wood 等人指出了仅通过合成数据生成三维人脸模型的可行性。该方法基于视觉效果技术（Visual Effects，EFX）的大规模运用，对合成的人脸模型进行采样并添加标签，用于训练识别网络（Face Parsing Network）；同时维护大型的饰品库、服装库、毛发库来随机搭配，进而形成各种各样的三维人脸模型。为了解决人工添加标签的误差问题，他们还训练了一个标签自适应网络（Label Adaptation Network）。[31]

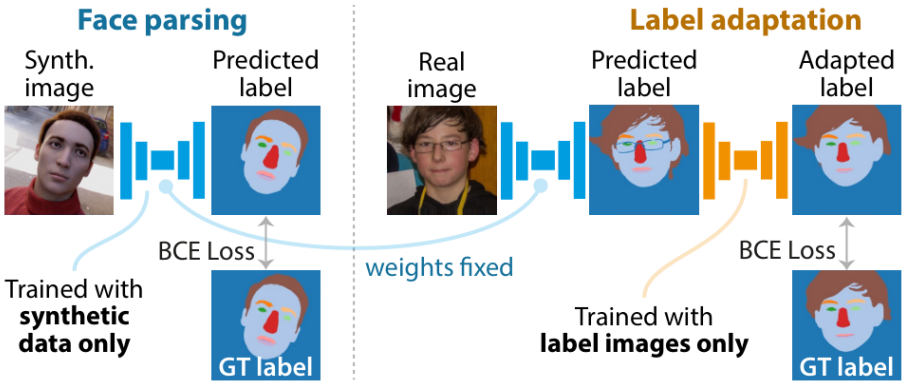



图 30: 人脸识别网络和标签自适应网络

**Ifinigen** Ifinigen 是由 Alexander Raistrick 等人研制的一套三维场景程序化生成器，用于生成大量的多样化三维训练数据。它通过集成大量的现有工具，独立生成并维护场景中各种元素的资产库（如石头、流水、天气、树木等），从零开始生成这些物体的几何形体和贴图材质，并最终在生成的场景地形中放置元素，达到生成逼真的三维自然景观的效果。[32]

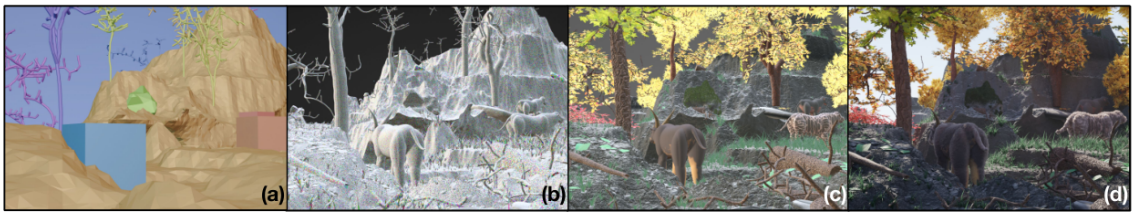



图 31: Infinigen 的场景搭建过程

# 4 总结

1. 今天的 3D 内容生产行业面临着需求和供给的强烈冲突。在这种情况下，3D AIGC 技术应运而生。如今的 3D AIGC 产业体系较为完善，主要分为底层模型端和生产贸易端。

2. 3D AIGC 的技术基础是相关的生成模型以及三维重建技术。

3. 3D AIGC 的流派主要有伪 3D 派、转化 3D 派、2D 升维派、原生 3D 派、工具派。

如今的 3D AIGC 行业，无论是底层技术，还是产业体系，都在朝着更完整、更先进的方向发展。我们相信未来的 3D AIGC 技术一定能够极大促进 3D 内容的生产，解决长期存在于 3D 内容市场的供求冲突问题。

# 参考文献

[1] Eigen, David et al. “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network.” NIPS (2014).

[2] Choy, Christopher Bongsoo et al. “3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction.” European Conference on Computer Vision (2016).

[3] Fan, Haoqiang et al. “A Point Set Generation Network for 3D Object Reconstruction from a Single Image.”2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 2463-2471.

[4] Chen, Rui et al. “Point-Based Multi-View Stereo Network.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 1538-1547.

[5] Wang, Nanyang et al. “Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images.” European Conference on Computer Vision (2018).

[6] Mildenhall, Ben et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.” ArXiv abs/2003.08934 (2020): n. pag.

[7] Müller, Thomas et al. “Instant neural graphics primitives with a multiresolution hash encoding.” ACM Transactions on Graphics (TOG) 41 (2022): 1 - 15.

[8] Shen, Tianchang et al. “Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis.” ArXiv abs/2111.04276 (2021): n. pag.

[9] Kingma, Diederik P. and Max Welling. “Auto-Encoding Variational Bayes.” CoRR abs/1312.6114 (2013): n. pag.

[10] Goodfellow, Ian J., et al. “Generative Adversarial Networks.” arXiv.Org, 10 June 2014, https://arxiv.org/abs/1406.2661.

[11] Radford, Alec et al. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” CoRR abs/1511.06434 (2015): n. pag.

[12] Karras, Tero et al. “Progressive Growing of GANs for Improved Quality, Stability, and Variation.” ArXiv abs/1710.10196 (2017): n. pag.

[13] Brock, Andrew et al. “Large Scale GAN Training for High Fidelity Natural Image Synthesis.” ArXiv abs/1809.11096 (2018): n. pag.

[14] Karras, Tero et al. “A Style-Based Generator Architecture for Generative Adversarial Networks.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018): 4396-4405.

[15] Sohl-Dickstein, Jascha Narain et al. “Deep Unsupervised Learning using Nonequilibrium Thermodynamics.” ArXiv abs/1503.03585 (2015): n. pag.

[16] Ho, Jonathan et al. “Denoising Diffusion Probabilistic Models.” ArXiv abs/2006.11239 (2020): n. pag.

[17] Rombach, Robin et al. “High-Resolution Image Synthesis with Latent Diffusion Models.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 10674-10685.

[18] Liu, Ruoshi et al. “Zero-1-to-3: Zero-shot One Image to 3D Object.” ArXiv abs/2303.11328 (2023): n. pag.

[19] Xiang, Jianfeng et al. “3D-aware Image Generation using 2D Diffusion Models.” ArXiv abs/2303.17905 (2023): n. pag.

[20] Hao, Zekun et al. “GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 14052-14062.

[21] Ouyang, Hao et al. “Real-Time Neural Character Rendering with Pose-Guided Multiplane Images.” European Conference on Computer Vision (2022).

[22] Jain, Ajay et al. “Zero-Shot Text-Guided Object Generation with Dream Fields.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 857-866.

[23] Poole, Ben et al. “DreamFusion: Text-to-3D using 2D Diffusion.” ArXiv abs/2209.14988 (2022): n. pag.

[24] Khalid, Nasir Mohammad et al. “CLIP-Mesh: Generating textured meshes from text using pretrained image-text models.” SIGGRAPH Asia 2022 Conference Papers (2022): n. pag.

[25] Lin, Chen-Hsuan et al. “Magic3D: High-Resolution Text-to-3D Content Creation.” ArXiv abs/2211.10440 (2022): n. pag.

[26] Singer, Uriel et al. “Text-To-4D Dynamic Scene Generation.” ArXiv abs/2301.11280 (2023): n. pag.

[27] Tang, Junshu et al. “Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior.” ArXiv abs/2303.14184 (2023): n. pag.

[28] Gao, Jun et al. “GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images.” ArXiv abs/2209.11163 (2022): n. pag.

[29] Nichol, Alex et al. “Point-E: A System for Generating 3D Point Clouds from Complex Prompts.” ArXiv abs/2212.08751 (2022): n. pag.

[30] Wang, Tengfei et al. “Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion.” ArXiv abs/2212.06135 (2022): n. pag.

[31] Wood, Erroll et al. “Fake it till you make it: face analysis in the wild using synthetic data alone.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 3661-3671.

[32] Raistrick, Alexander R. E. et al. “Infinite Photorealistic Worlds using Procedural Generation.” ArXiv abs/2306.09310 (2023): n. pag.