Microsoft Research

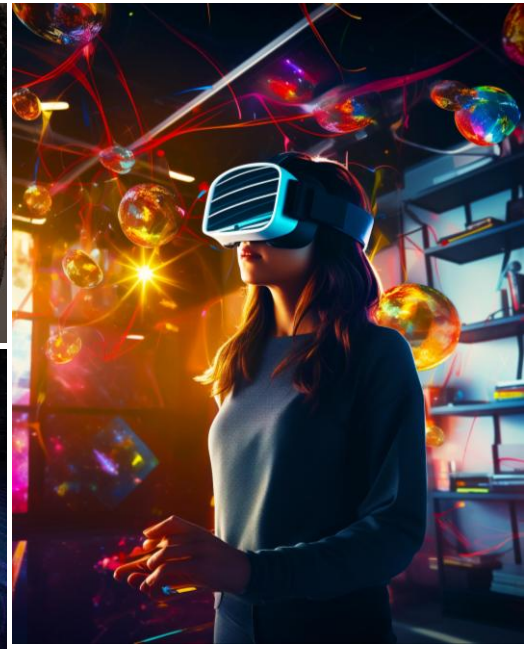
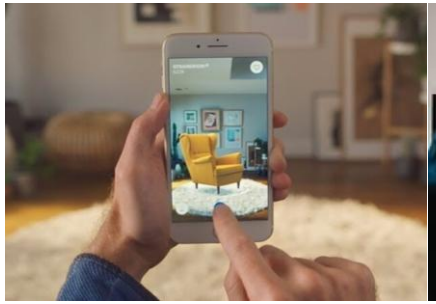
3D AIGC: From 3D GAN to Diffusion Models
三维生成: 从对抗生成模型到扩散模型探索

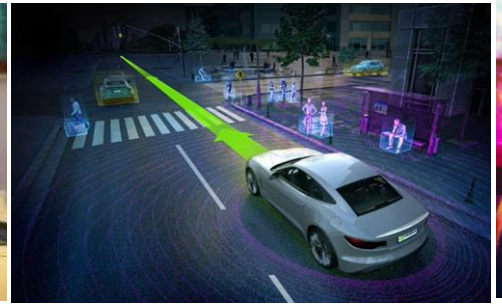
杨蛟龙
微软亚洲研究院
jiaoyan@microsoft.com

三维内容生成技术应用

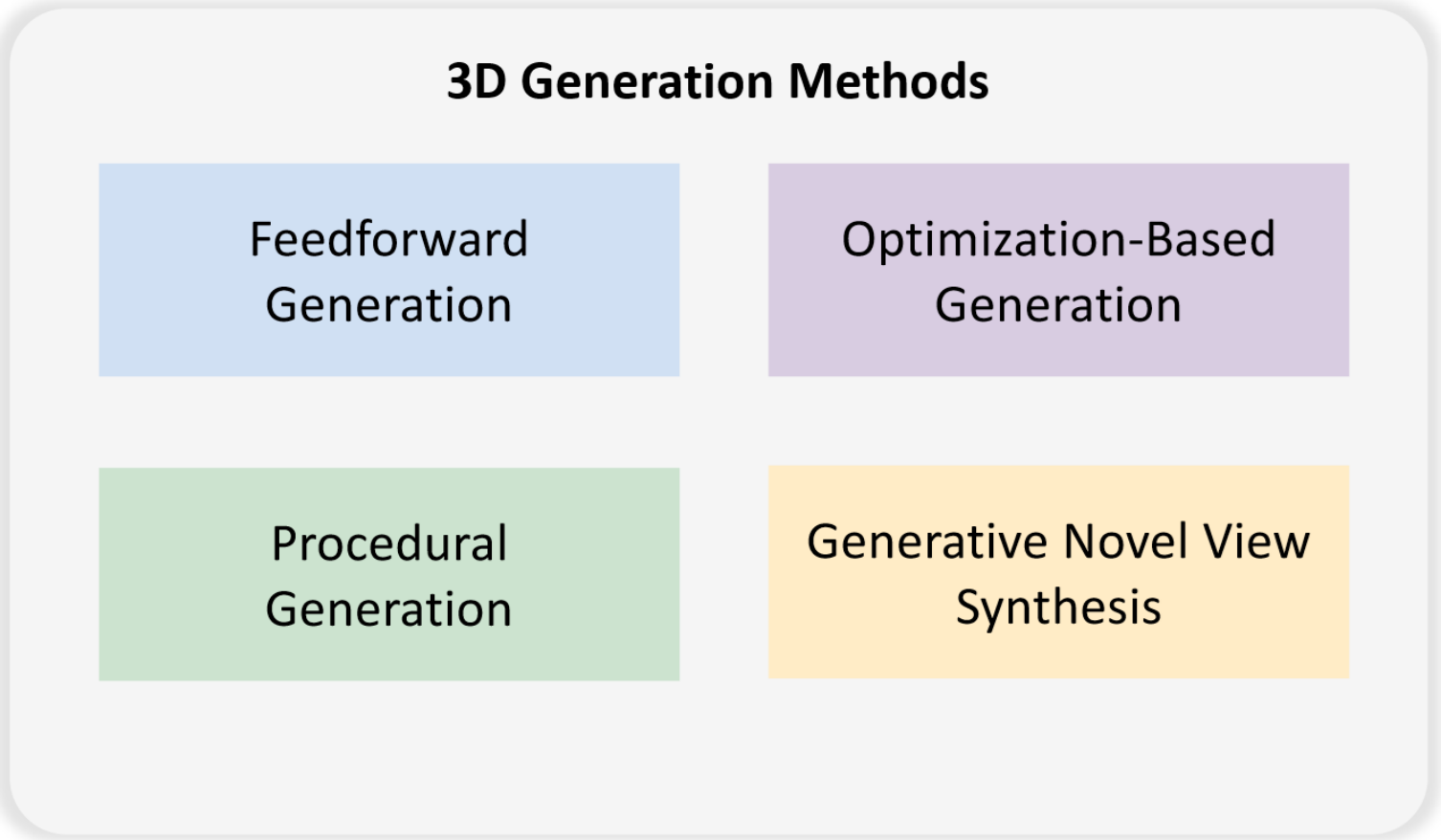
- 游戏设计
- 视频创作
- 影视制作
- 数字人
- AR/VR
- 自动驾驶
- …



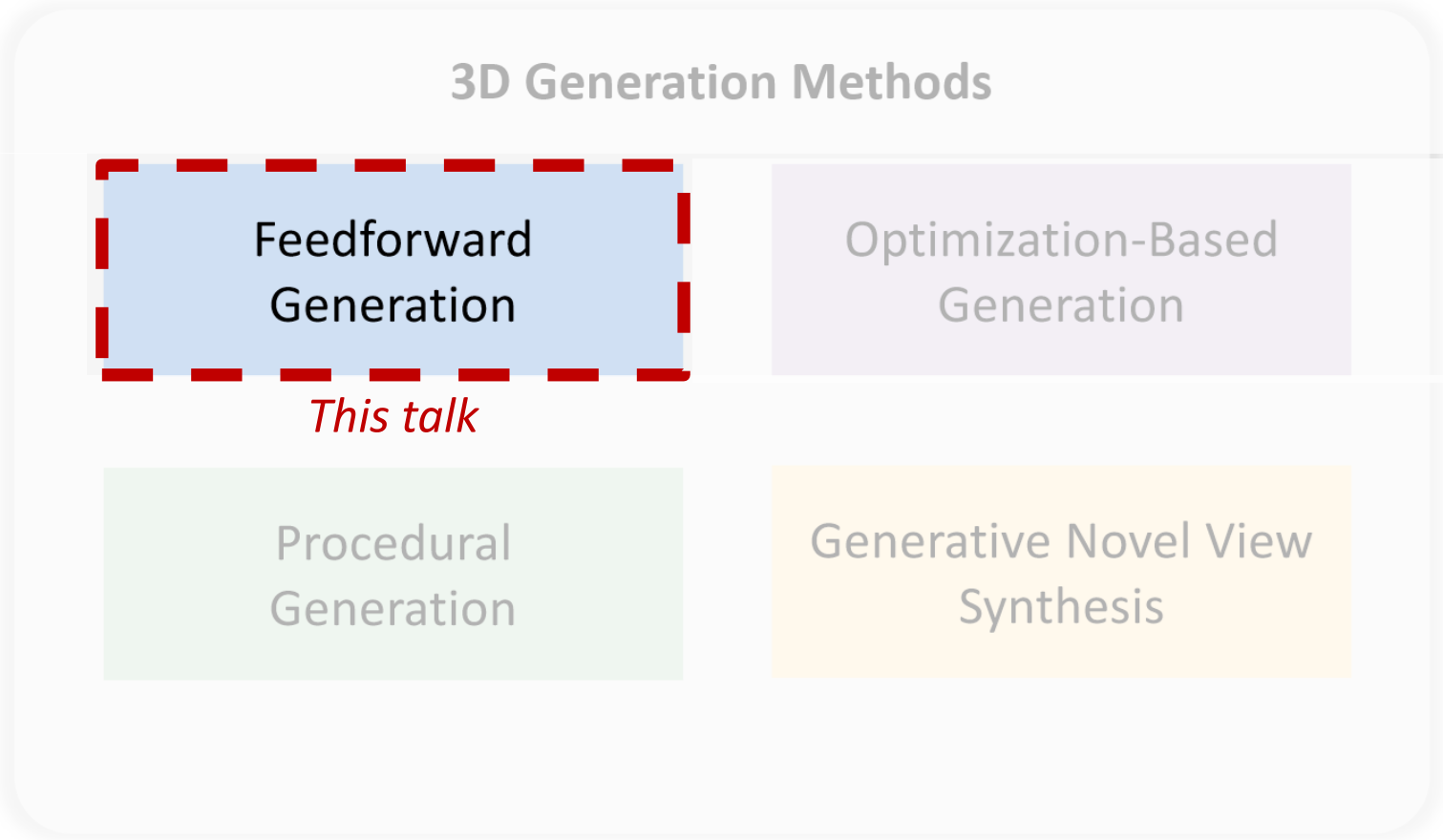


三维生成方法分类



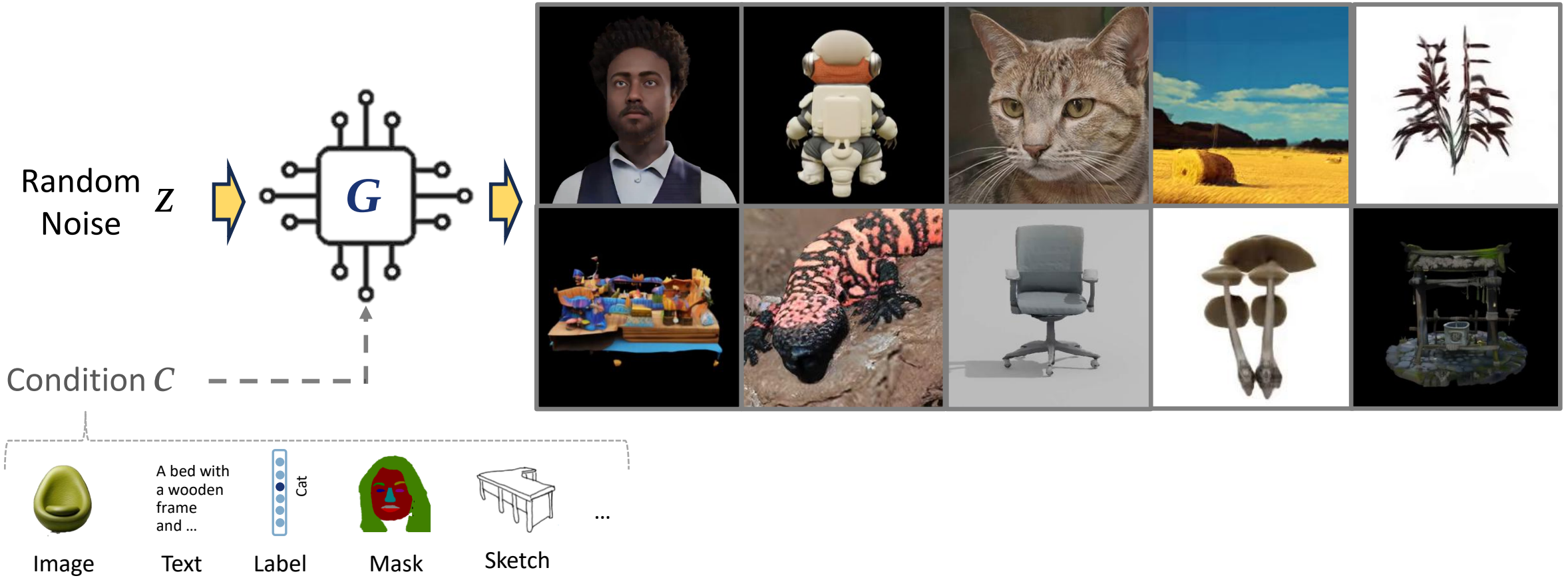
Li et al., Advances in 3D Generation: A Survey, arXiv 2024

三维生成方法分类



Li et al., Advances in 3D Generation: A Survey, arXiv 2024

基于生成式模型的三维生成



Unconditional & Conditional Generation

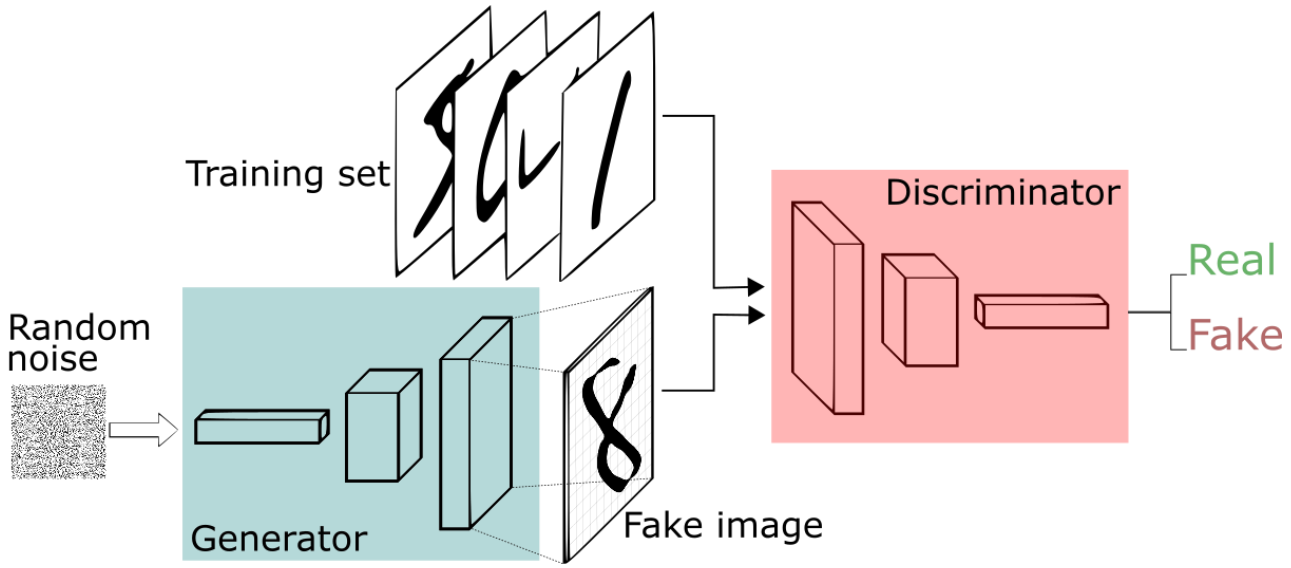
三维生成的关键要素

- **三维表达：**
  - ~~Mesh~~
  - ~~Volume (explicit)~~
  - **NeRF & Hybrid-Nerf**
    - Pi-GAN, GRAM
    - Triplane (EG3D)
  - **3D Gaussians**
- **生成模型：**
  - ~~VAE~~
  - ~~Flow Model~~
  - **GAN**
    - 3D(-aware) GAN
  - **Diffusion**
    - 3D Diffusion
    - Multiview 2D diffusion (+ 3D reconstruction)

Generative Adversarial Networks (GAN)

- Goodfellow et al. “Generative Adversarial Nets.” NIPS 2014.
  - Two-player minimax game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$



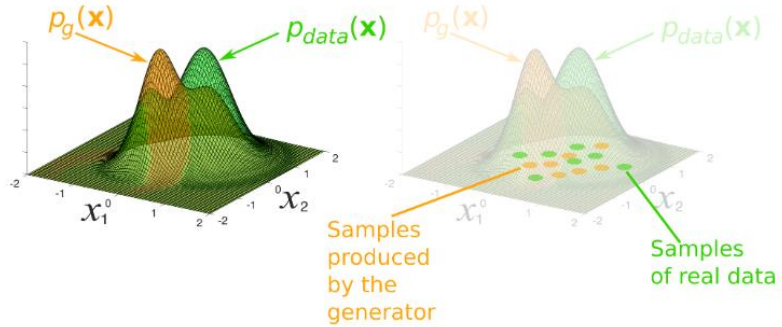
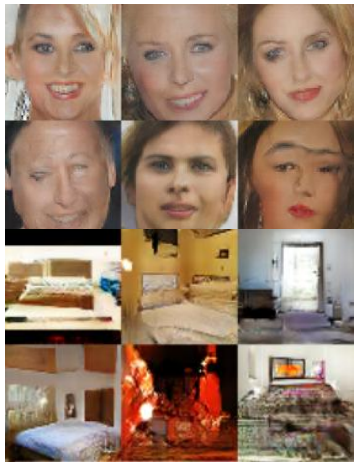


Image source: https://sthalles.github.io/intro-to-gans/ generative adversarial networks: an overview

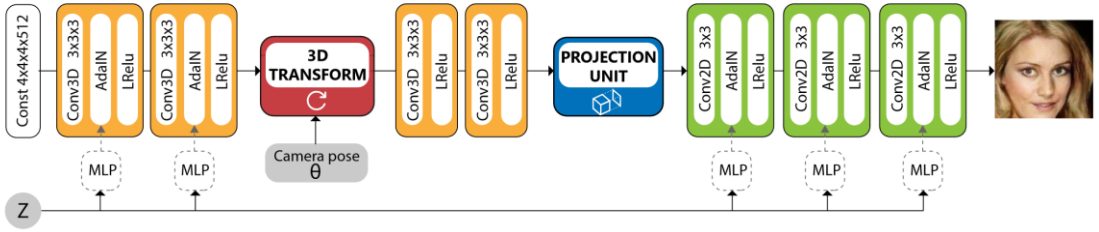
2D GANs

GAN 14’ DC-GAN 15’ PG-GAN 18’ StyleGAN 19’ StyleGAN2 20’

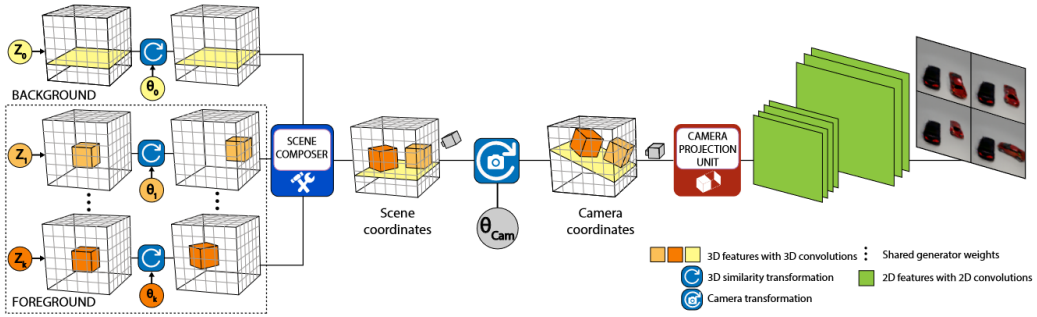


3D GANs – Earlier works

- Earlier 3D GANs
  - *3D-data-free scheme* ☺
  - *No explicit 3D representation* ☹



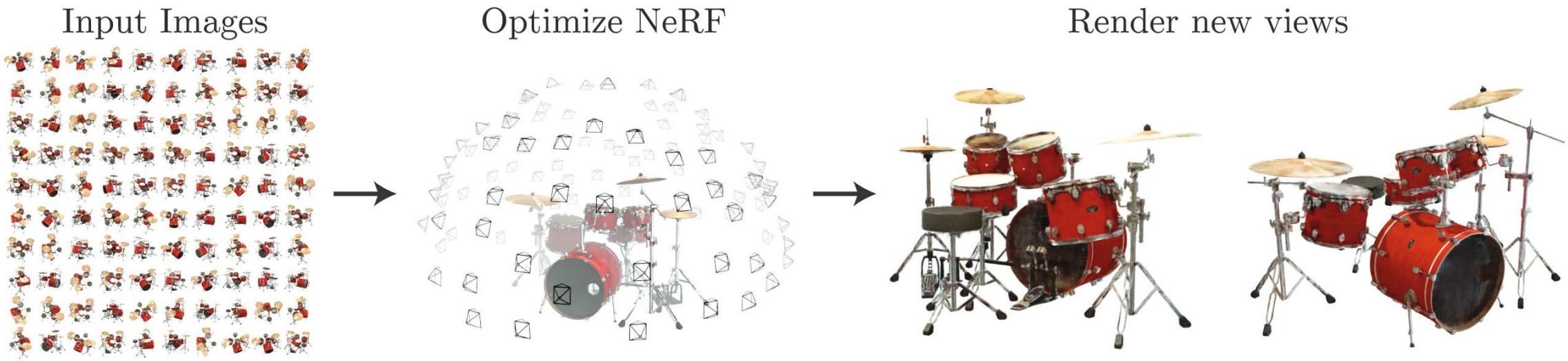
HoloGAN ICCV 2019



BlockGAN NeurIPS 2020

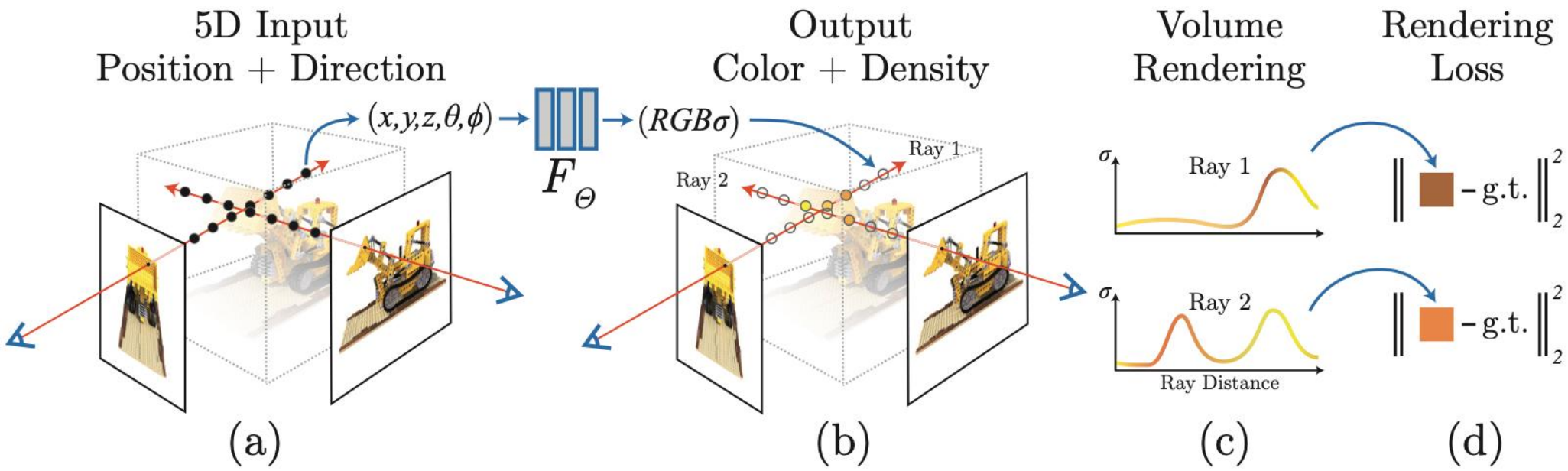
Neural Radiance Field (NeRF)

- Mildenhall et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV 2020 (Best Paper Honorable Mention)
- Task: Input Multiview images -> Output rendering at arbitrary novel views.



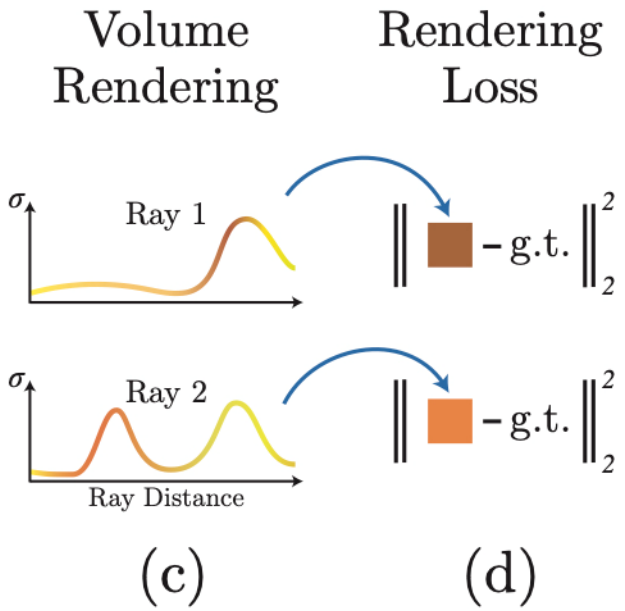
Neural Radiance Field (NeRF)

- Use network (MLP) to model the 5D radiance field (view direction, rgb, density) of a scene



Neural Radiance Field (NeRF)

- An explicit, physics-based rendering process



The color $C(\mathbf{r})$ of camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt\,, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$$

$T(t)$ denotes the accumulated transmittance from t_n to t

Use a discrete set of samples to estimate the integral:

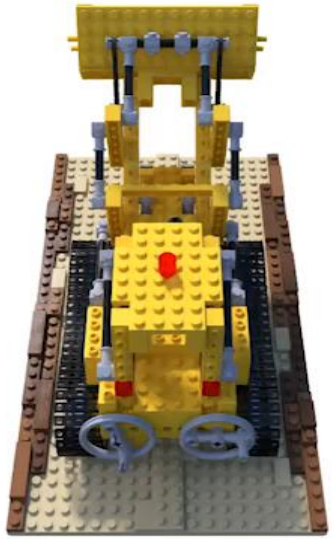
$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i\,, \text{ where } T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j\delta_j\right)$$

Neural Radiance Field (NeRF)

Real forward-facing scenes
(20-60 images for training)

Neural Radiance Field (NeRF)

Real forward-facing scenes
(20-60 images for training)

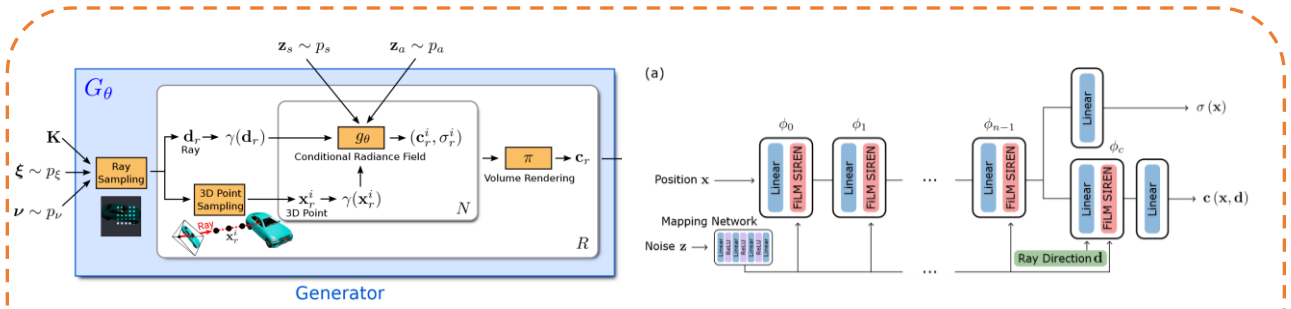







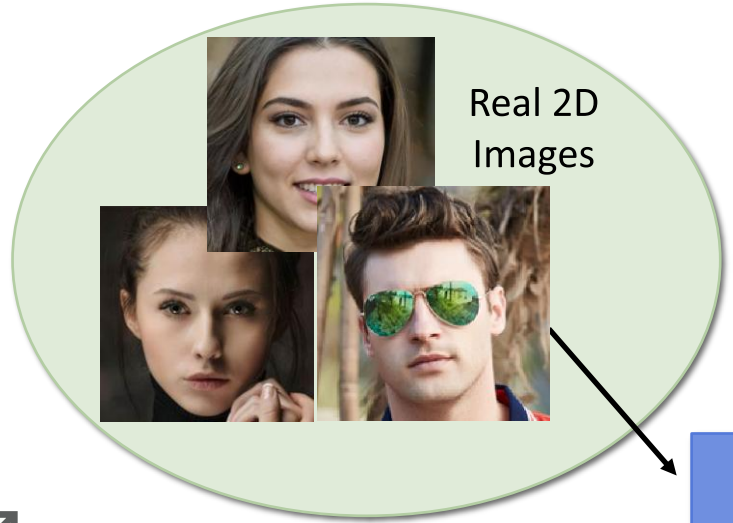
Synthetic 360° scenes
(100 images for training)

3D GAN with NeRF



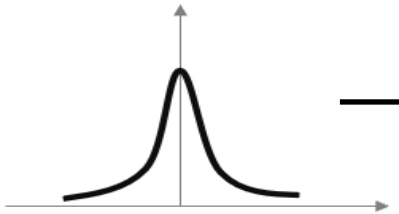
GRAF NeurIPS 2020

Pi-GAN CVPR'21



Discriminator

Real or generated?



Random noise

NeRF Generator

Rendering

Rendered Image

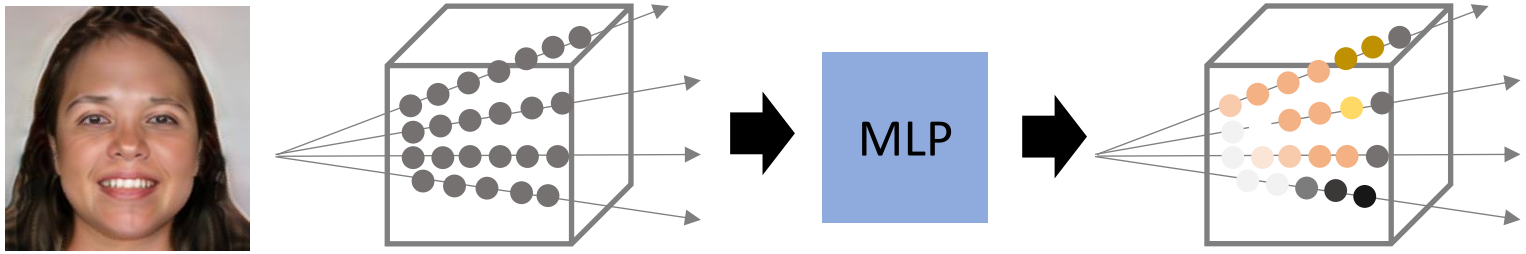
Random viewpoint

3D GAN with NeRF – Pros and Cons

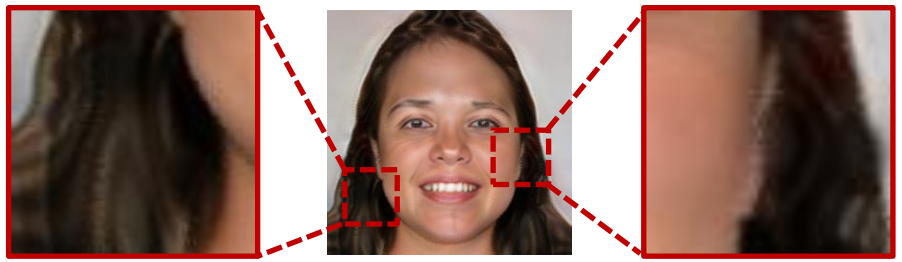
- Pros:
  - **High-quality** image
  - **3D-consistent** rendering

3D GAN with NeRF – Pros and Cons

- Cons:
  - Formidable computation cost for GAN training
    - (256×256) × 24 × (8×256) × 4B = **12GB** memory cost to render one 256^2 image!
      - Image resolution
      - Samples per ray
      - MLP layers
    - Double cost, i.e., **24G,** for training due to gradients



  - Monte Carlo sampling loses details & brings noise



The Dilemma

- Direct Volumetric Rendering
  - e.g., GRAF, Pi-GAN
  - Low quality

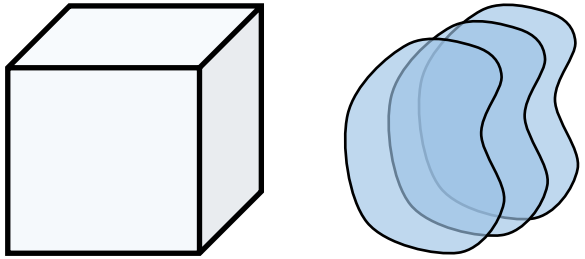

Pi-GAN
CVPR 2021

- Low-Res Vormetric Rendering + Conv Upsampling
  - e.g., GIRAFFE, StyleNeRF
  - No strict 3D consistency


GIRAFFE
CVPR 2021

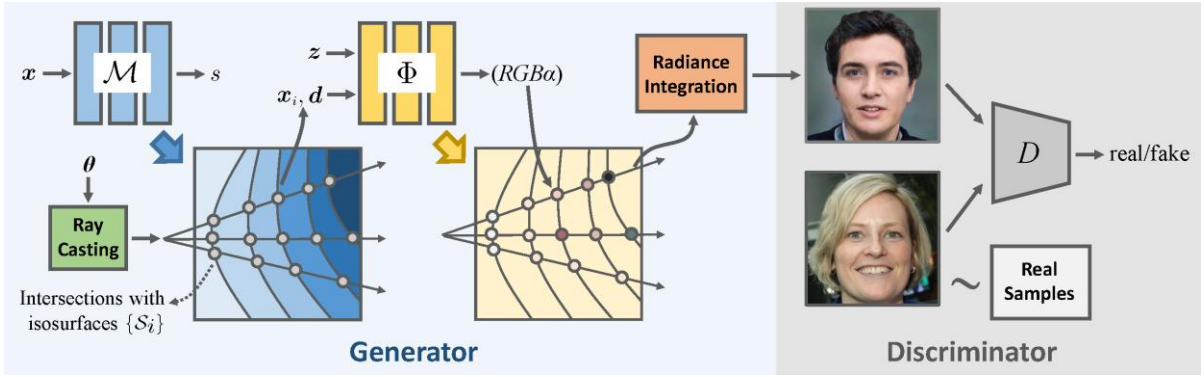
GRAM – *Generative Radiance Manifolds*

- 一种新的三维表示: **神经辐射流形**
  - 紧致、可自适应于所学习物体类别的几何
  - 可实现细节学习
  - 避免采样噪声

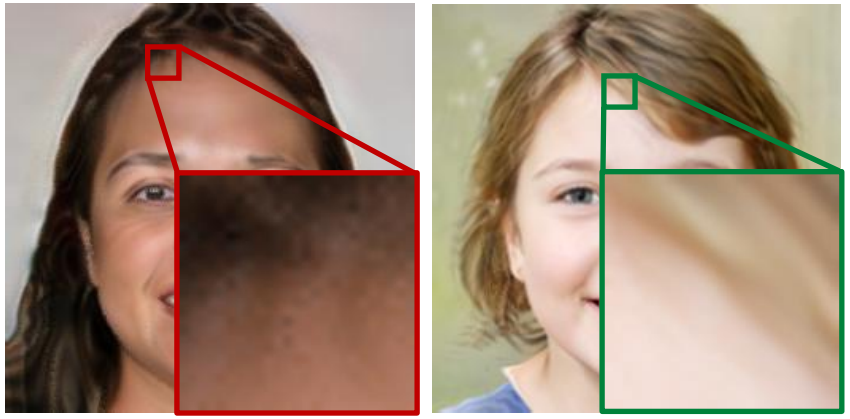


Volume ✗ Surface manifolds ✓

细节生成能力



算法流程



去噪声能力

GRAM – *Generative Radiance Manifolds*

- 实验：256x256三维人脸(FFHQ)，猫脸，汽车模型生成
  - **高质量、高三维一致性**的生成效果，显著优于先前技术
  - 逐帧神经网络渲染速度：2fps
  - 将辐射流形提取成mesh渲染：180fps

Methods	FFHQ 256^2		Cats 256^2		CARLA 128^2	
	FID	KID	FID	KID	FID	KID
StyleGAN2	12.2	0.18	10.5	0.30	12.6	0.45
GRAF	78.5	5.92	61.4	4.60	34.2†	1.81†
pi-GAN	61.3	4.23	55.4[1]	4.33	38.7†	2.11†
GIRAFFE	38.6†	2.25†	22.3	1.12	109[2]	7.29
Ours	**29.8**	**1.16**	**16.7**	**0.75**	**28.5**	**1.09**




GRAM – *Generative Radiance Manifolds*

GRAM – *Generative Radiance Manifolds*

- 缺陷：内存消耗依然过大

256×256 ⟶ 1024×1024

(**12GB/24GB**) (**192GB/384GB**)

渲染 训练 渲染 训练

GRAM-HD – 高分辨率神经辐射流形

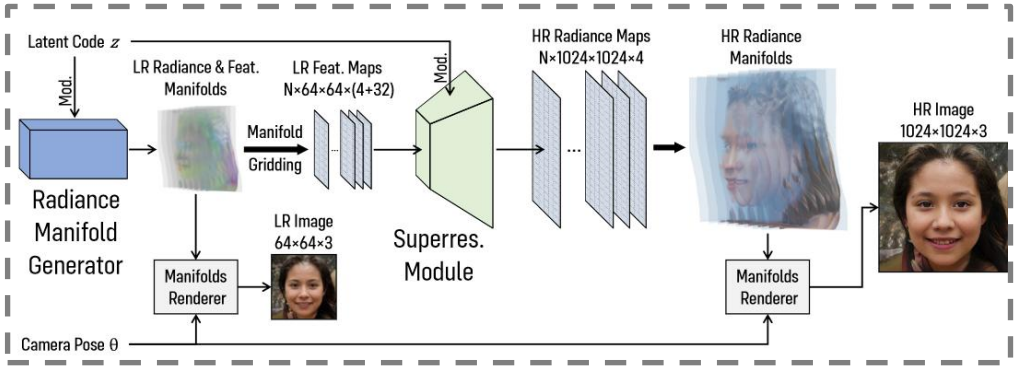
- 解决方案：使用高效的**二维卷积**对神经辐射流形进行上采样实现**三维超分辨率**

低分辨率
辐射流形

超分

高分辨率
辐射流形

渲染



GRAM-HD – 高分辨率神经辐射流形

- 实验结果：大幅减小内存消耗、提升渲染速度、生成质量更高

Table 1: **Left:** Comparison of memory cost when training. Whole computational graph retained with AMP enabled. **Right:** Comparison of time consumed per generated image.

Methods	256^2	1024^2
GRAM	22.2G	OOM (~400G)
GRAM-HD	5.4G	9.0G
	4x	44x

Methods	256^2	1024^2
GRAM	0.43s	6.69s
GRAM-HD	0.22s	0.36s
	2x	19x

Table 2: Quantitative comparison between GRAM and GRAM-HD.

Method	FFHQ256 FID	FFHQ256 KID	CATS256 FID	CATS256 KID
GRAM	15.0	6.55	12.9	7.37
GRAM-HD	**13.0**	**5.14**	**7.05**	**2.53**
	Higher quality!			

GRAM-HD – 高分辨率神经辐射流形

- 实验结果：1024x1024人脸生成 (FFHQ)

GRAM-HD – 高分辨率神经辐射流形

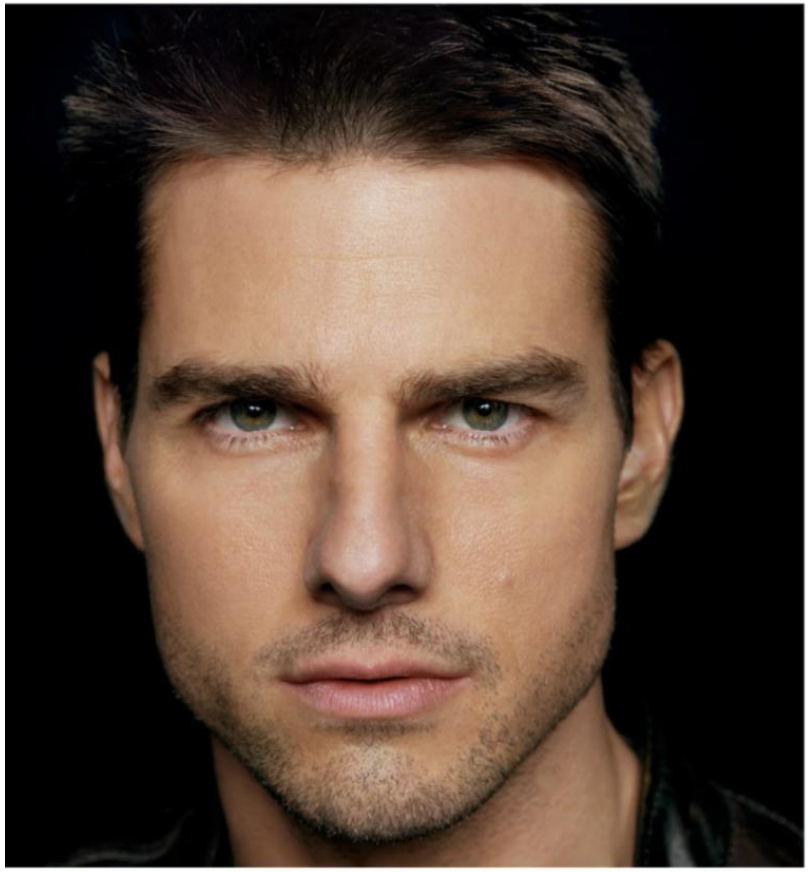
- 实验结果：1024x1024人脸生成 (FFHQ)

GRAM-HD – 高分辨率神经辐射流形

- 实验结果：512x512猫脸生成 (AFHQ)

GRAM-HD – 高分辨率神经辐射流形

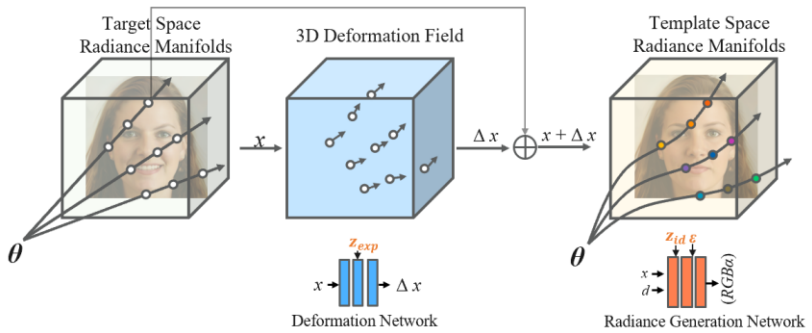
- 实验结果：真实人脸图片嵌入与3D渲染



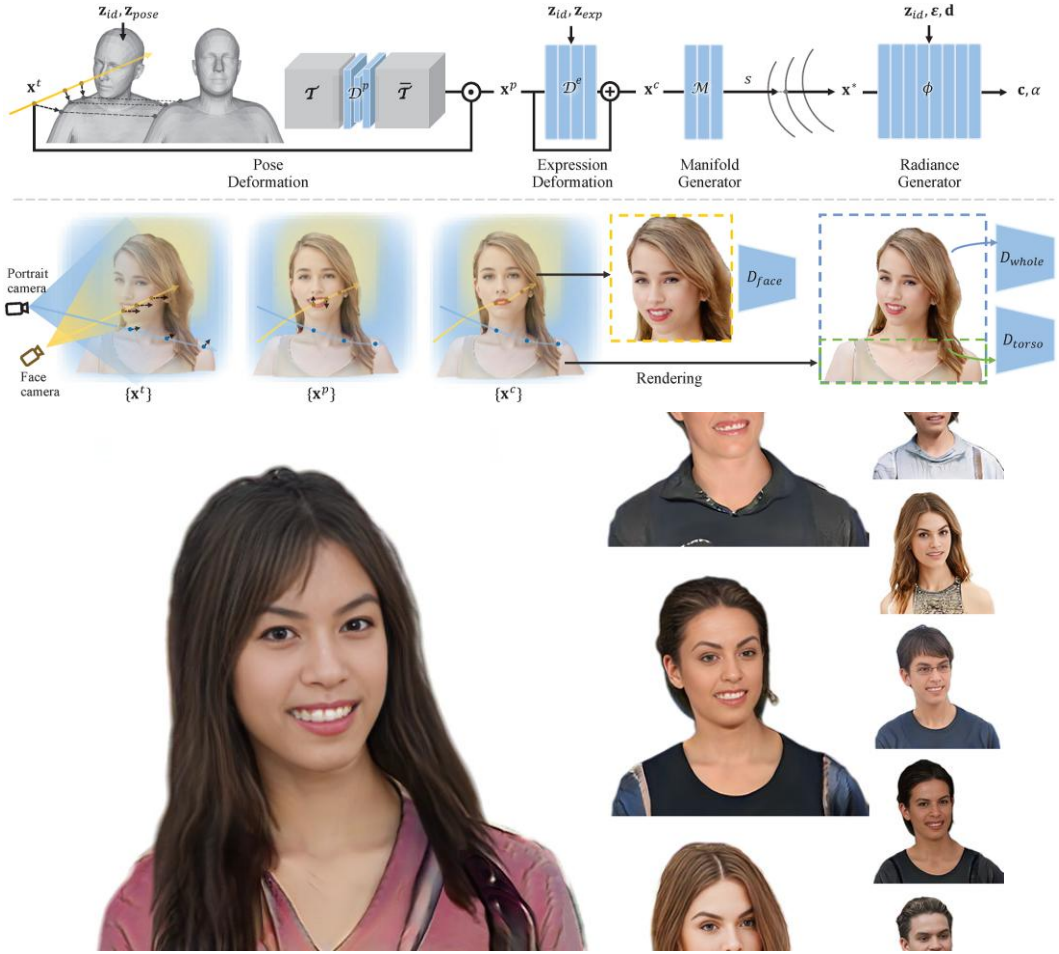
Target

Initial

GRAM + 形变场: 可控3D人像生成



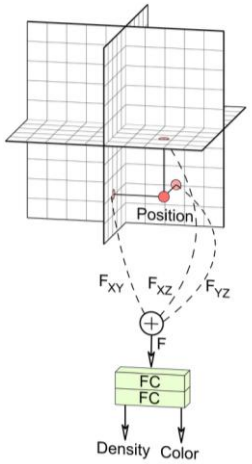
Y. Wu et al. AniFaceGAN
NeurIPS 2022 (**Spotlight**)



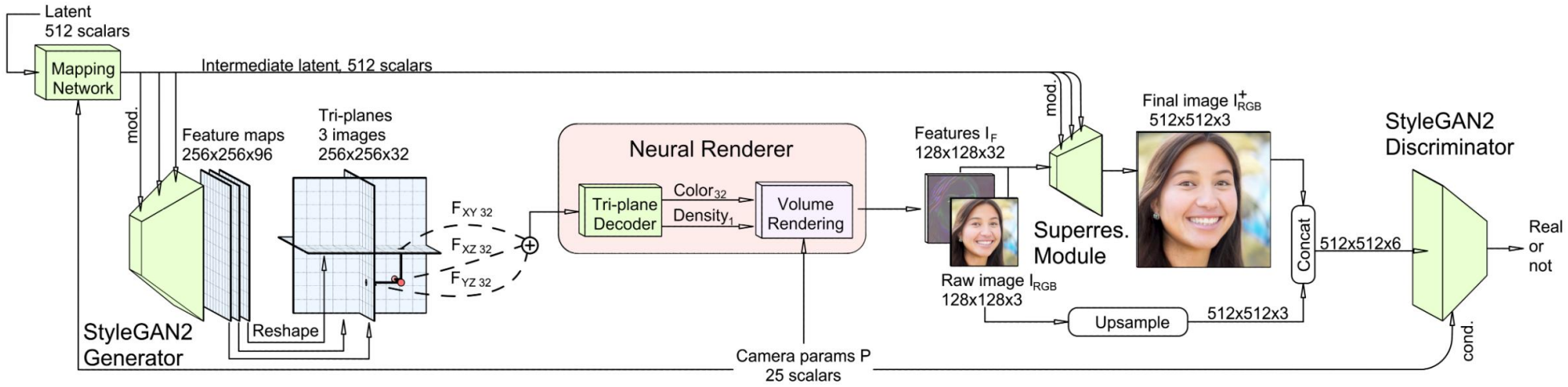
Y. Wu et al. AniPortraitGAN
SIGGRAPH Asia 2023

EG3D – Triplane + 3D-aware GAN

- 三维表示: **Triplane (三平面)**
  - 显示+隐式混合表达
  - 卷积网络(StyleGAN2)生成三平面特征
    + NeRF MLP小网络体渲染
    + 2D卷积图像超分
  - 高质量纹理与几何
  - 高生成和渲染速度(实时渲染)



高质量纹理与几何展示



算法流程

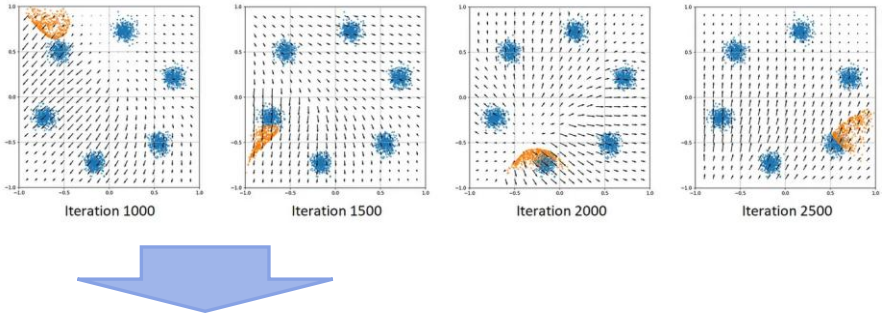
EG3D – Triplane + 3D-aware GAN

- 实验：512x512三维人脸(FFHQ)与猫脸(AFHQ)
  - 高质量生成效果，显著优于先前技术
  - 渲染速度：实时(>25fps, single RTX 3090)

	FFHQ				Cats
	FID↓	ID↑	Depth↓	Pose↓	FID↓
GIRAFFE 256^2	31.5	0.64	0.94	.089	16.1
π-GAN 128^2	29.9	0.67	0.44	.021	16.0
Lift. SG 256^2	29.8	0.58	0.40	.023	—
Ours 256^2	**4.8**	0.76	**0.31**	**.005**	**3.88**
Ours 512^2	**4.7**	**0.77**	0.39	**.005**	**2.77**†

3D GAN 优缺点

- 优点：
  - 无需3D数据（可用2D图像训练）
- 缺点：
  - 建模能力弱
  - 训练稳定性差（Mode collapse issue）
  - 难以scale-up



- 只在单类别、简单物体上较为成功
  - 人像、动物脸、汽车、家具、水果等

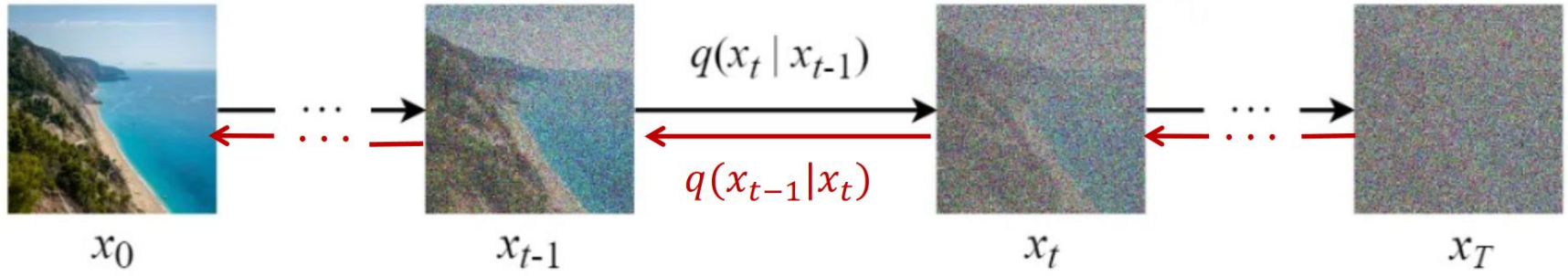
VQ3D ICCV 2023

3DGP ICLR 2023

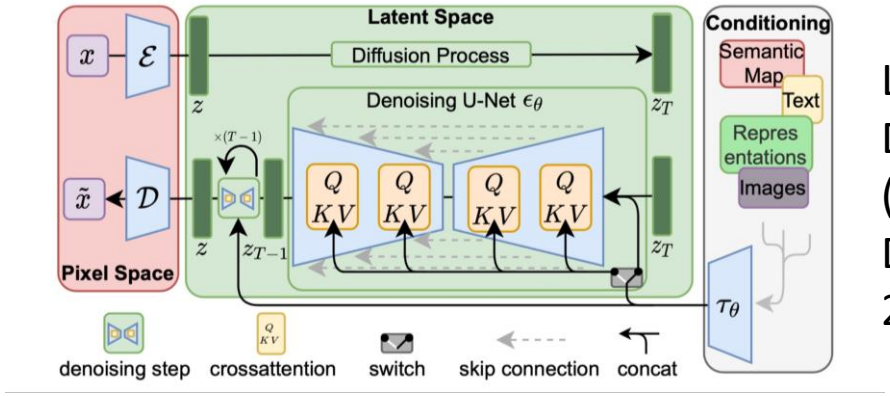
ImageNet物体建模困难

Diffusion Models

- Forward process introduces noise (data generation) and inverse process (decoder) denoises the data
- Convert a simple base distribution (e.g., a Gaussian) to the target (data) distribution iteratively
- Stable training; does not suffer from model collapse



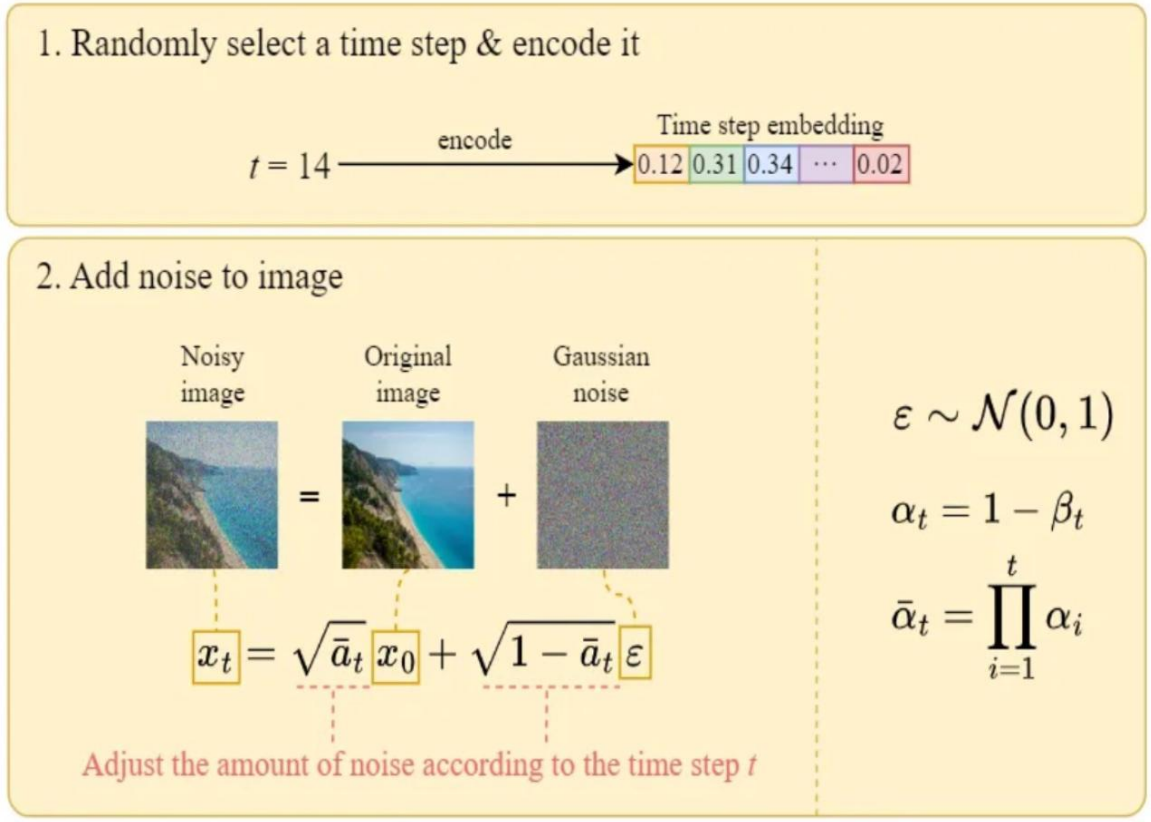
"Diffusion Models Beats GANs"

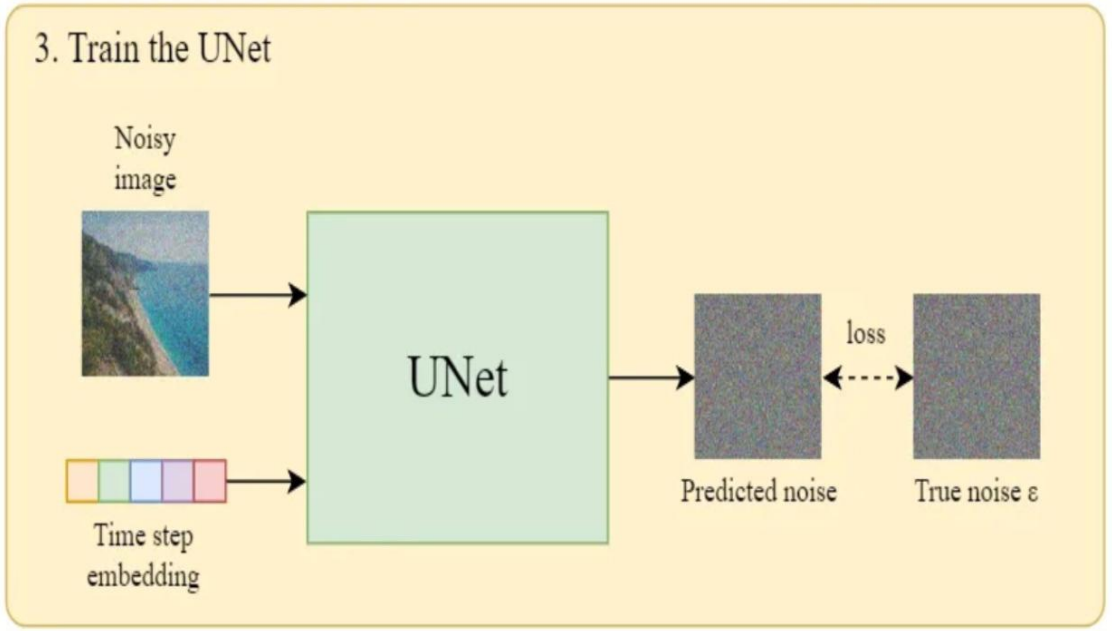


Latent Diffusion (Stable Diffusion), 2022

Diffusion Models

- Training algorithm



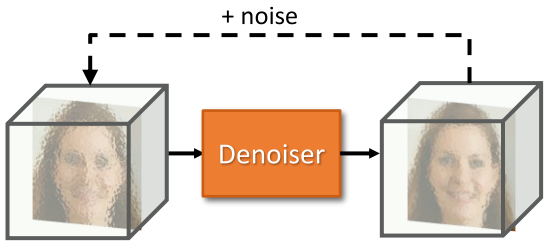


From Steins (medium.com)

3D Diffusion 优缺点 (vs. 3D GAN)

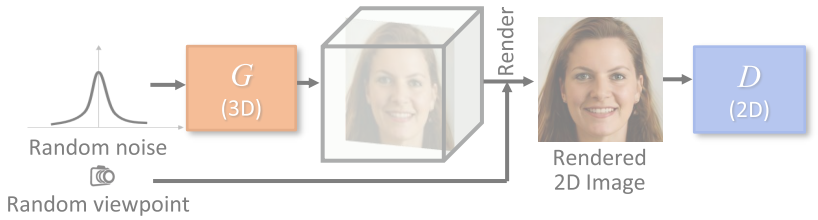
3D Diffusion

- 优点:
  - 建模能力强
  - 训练稳定性好
  - 容易scale-up
- 缺点:
  - 需要3D训练数据
  - 需要大算力



3D GAN

- 缺点:
  - 建模能力弱
  - 训练稳定性差
  - 难以scale-up
- 优点:
  - 无需3D训练数据 (可用2D图像)

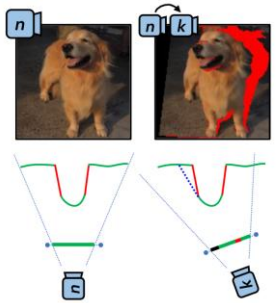


IVID – 使用2D扩散模型进行3D生成

- 迭代视角采样生成 (Iterative View Sampling)

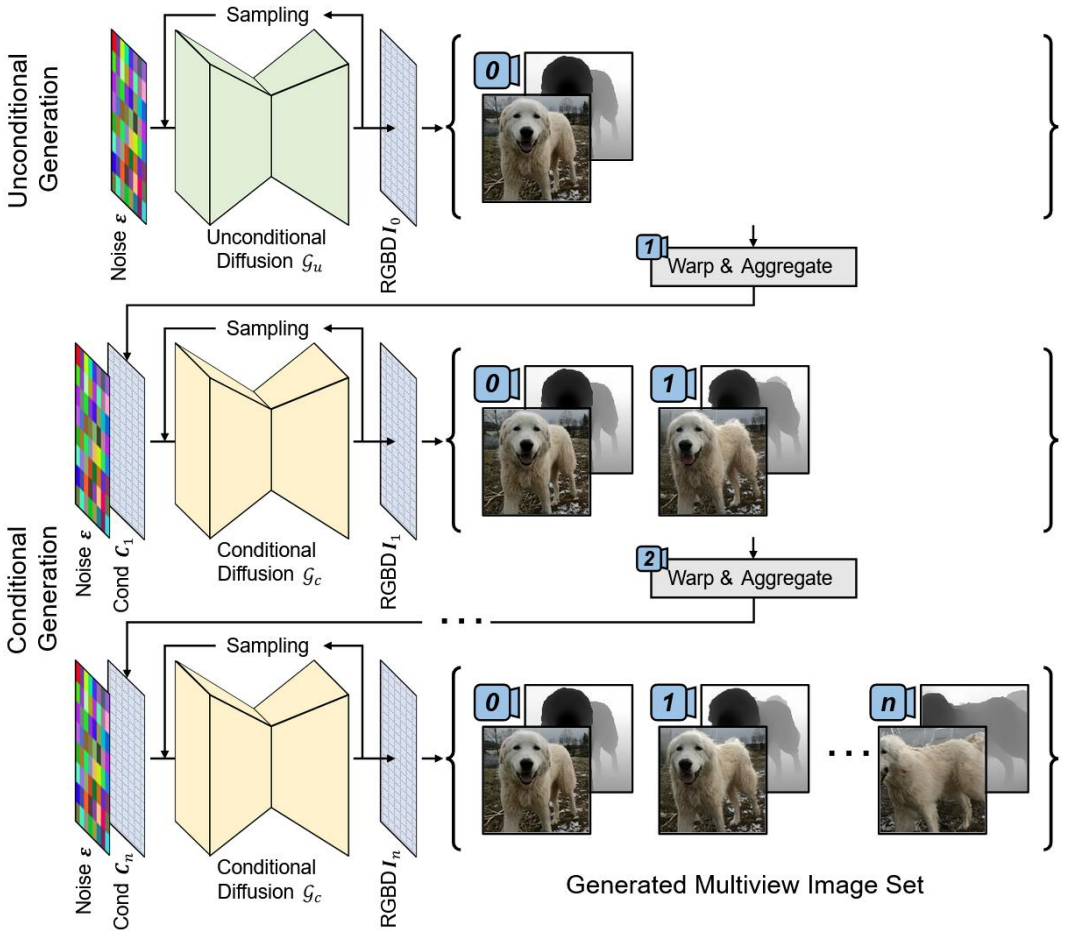
$$
\begin{aligned}
q_a(\mathbf{x}) = \; & q_i(\Gamma(\mathbf{x}, \boldsymbol{\pi}_0))\cdot && \textit{第1个视角图像 (Unconditional)} \\
& q_i(\Gamma(\mathbf{x}, \boldsymbol{\pi}_1)|\Gamma(\mathbf{x}, \boldsymbol{\pi}_0))\cdot && \textit{第2个视角图像 (Conditional)} \\
& \cdots \\
& q_i(\Gamma(\mathbf{x}, \boldsymbol{\pi}_N)|\Gamma(\mathbf{x}, \boldsymbol{\pi}_0), \cdots, \Gamma(\mathbf{x}, \boldsymbol{\pi}_{N-1})) \\
& && \textit{第N个视角图像 (Conditional)}
\end{aligned}
$$

- RGB-D生成
  - 使用Depth Warping作为新视角生成条件
  - 单目深度估计算法提供训练数据



Depth Warping

MiDaS TPAMI2020



算法流程

IVID – 使用2D扩散模型进行3D生成

- 实验：ImageNet训练

(360度生成!)

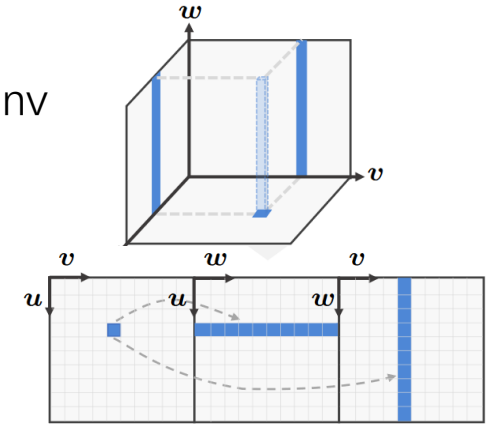
Method	ImageNet FID↓	ImageNet IS↑
pi-GAN [4]	138	6.82
EpiGRAF [46]	67.3	12.7
EG3D [3]	40.4	16.9
Ours	9.45	68.7

局限：

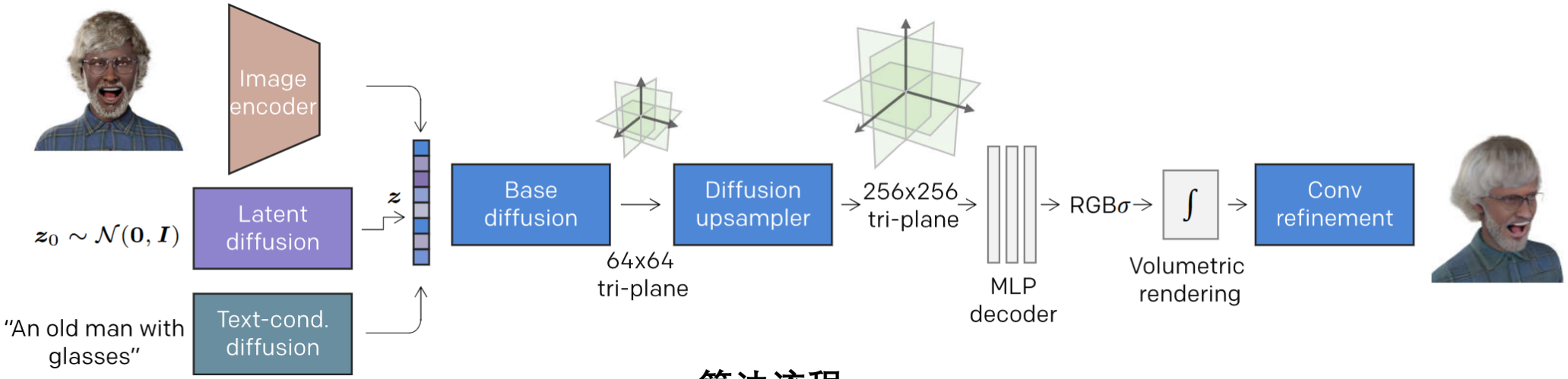
依赖深度图，最终生成质量受深度图数据质量影响严重

Rodin – *Triplane Diffusion*

- **Triplane 扩散生成模型**
  - Triplane “roll-out” + 3D-aware Conv
  - 2阶段diffusion (Base+Upsampler)
- **2D Conv图像后处理**
- ***100K 3D艺术人像数据训练***
- **支持条件生成（图像、文本）**



生成结果展示



算法流程

Rodin – *Triplane Diffusion*

- 实验结果：显著优于3D GAN方法

	Pi-GAN	GIRAFFE	EG3D	Autoencoder	Ours
FID ↓	78.3	64.6	40.5	67.4	**26.1**



T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, et al. Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. *CVPR2023*

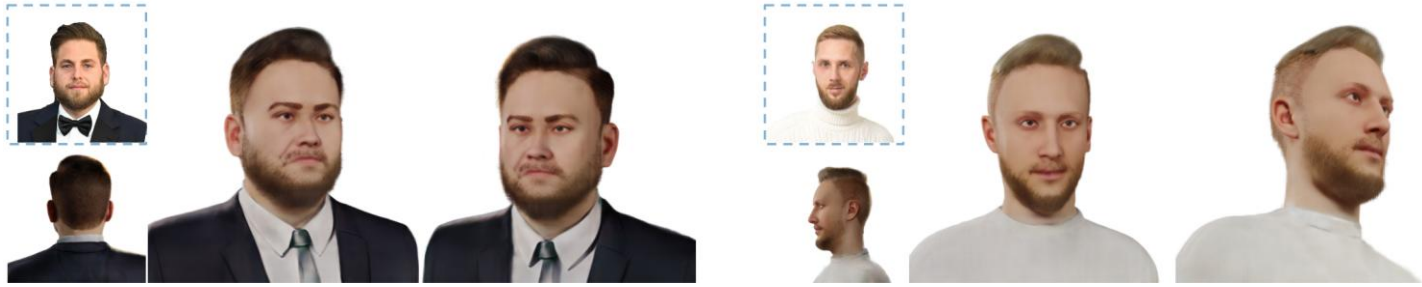
Rodin – *Triplane Diffusion*

- 实验：图像与文本条件输入

A woman with afro hairstyle wearing red

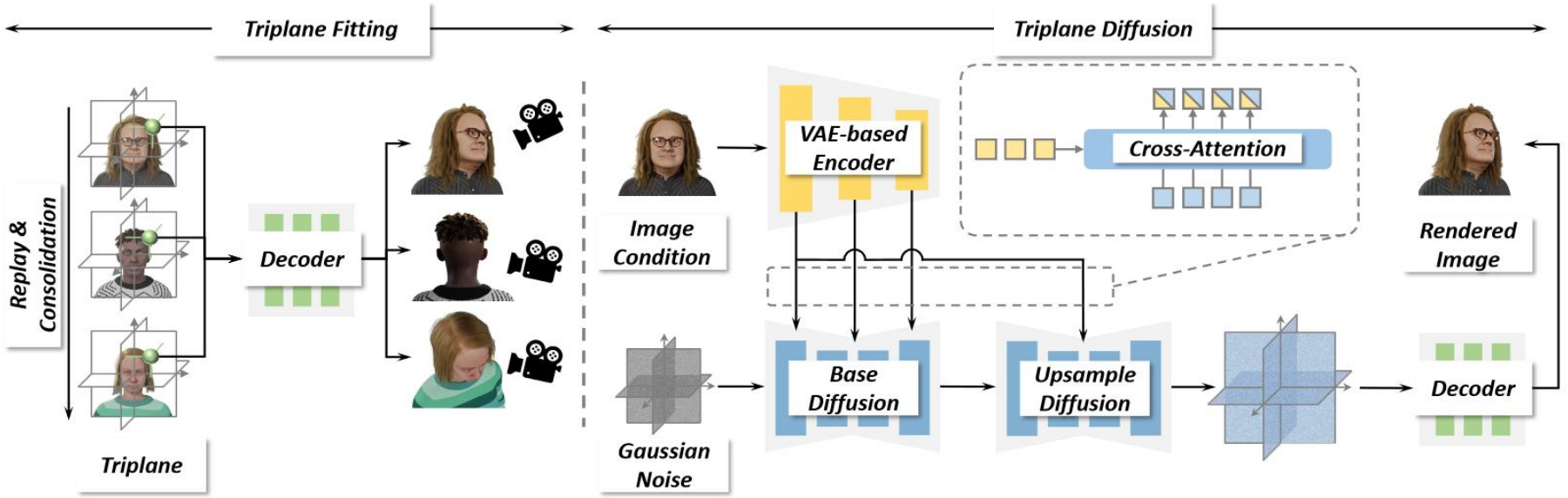
Rodin-HD – *Higer Res, Better 3D consistency*

- **分辨率提升**
  - 512x512->1024x1024
- **三维一致性提升**
  - 移除2D Conv图像后处理模块
- **算法细节优化**
  - 更好的Triplane fitting和diffusion training



生成结果展示

算法流程



Rodin-HD – *Higer Res, Better 3D consistency*

- 实验结果

	Pi-GAN	GIRAFFE	EG3D*	Rodin*	Rodin	**Ours**
FID ↓	78.3	64.6	40.5	30.29	45.70	32.62

Table 2: Quantitative results of unconditional avatar generation. **The subscript * indicates that 2D refinement is applied to the rendered images.**

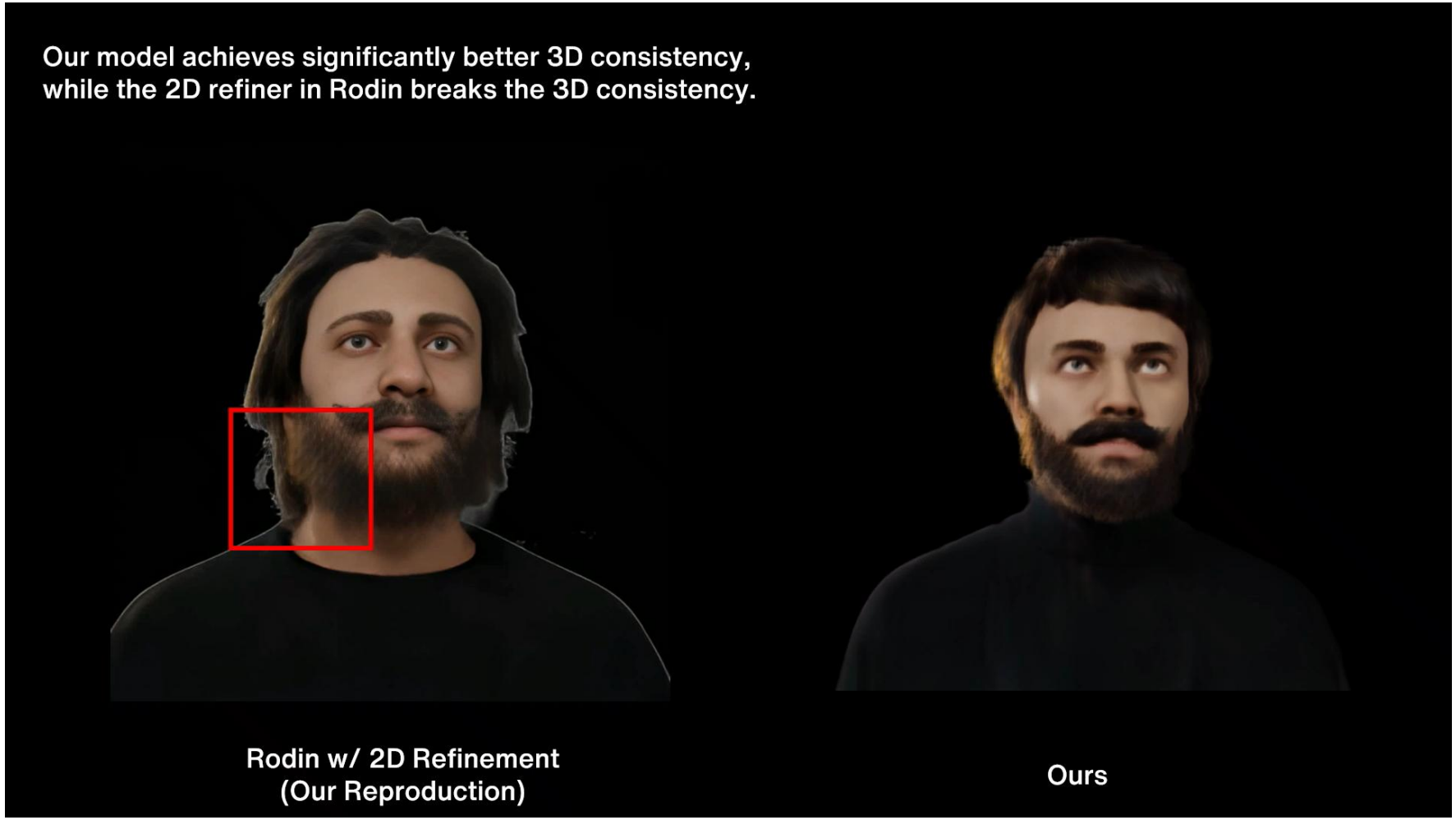
Models	PSNR↑	LPIPS↓	SSIM↑
Rodin* [71]	31.73	0.051	0.973
Ours	**35.46**	**0.041**	**0.975**

Table 3: 3D consistency measured by the fitting quality of NeuS [70].

B. Wang, Y. Cheng, C. Wang, T. Zhang, J. Yang, et al. RodinHD: High-Fidelity 3D Avatar Generation with Diffusion Models. *ECCV 2024*

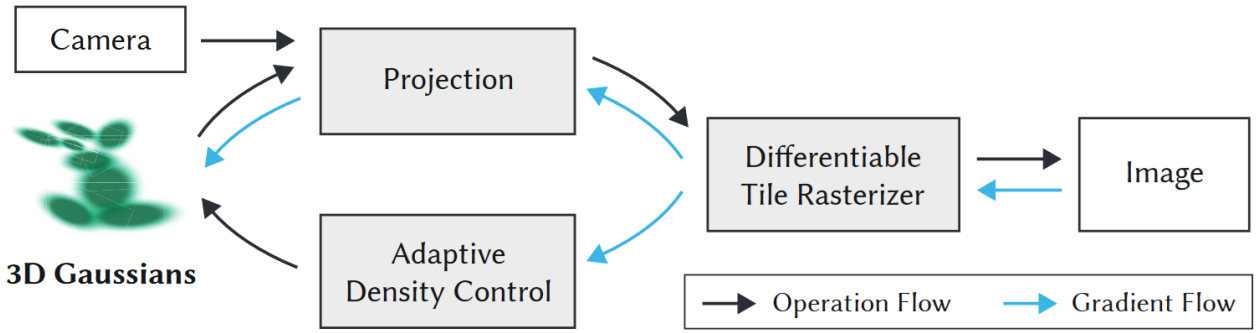
Rodin-HD – *Higer Res, Better 3D consistency*

- 实验结果



3D Gaussian Splatting (3DGS)

- Represent the scene with 3D Gaussians with properties of continuous volumetric radiance fields (*position, scale, rotation, opacity, view-dependent color*)
- An adaptive optimization scheme
- NeRF-parity rendering quality;
- Significantly faster rendering speed (>30fps for 1080p)



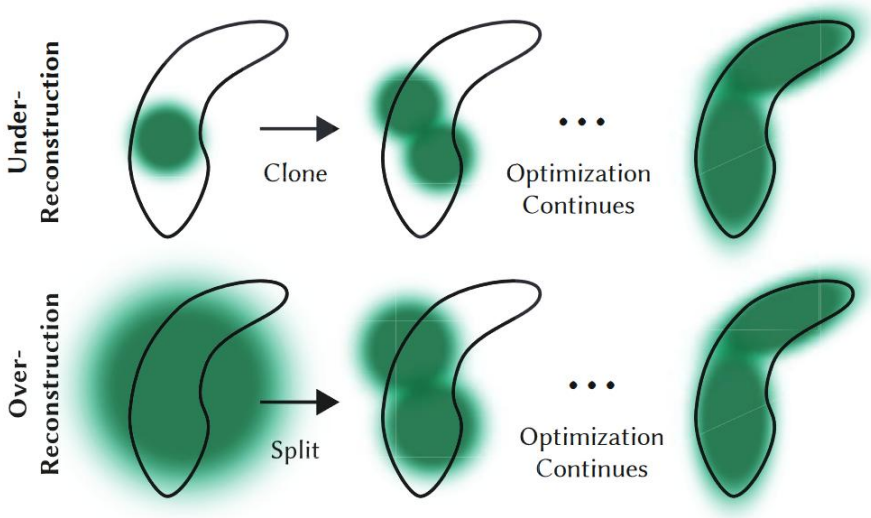


Fig. 4. Our adaptive Gaussian densification scheme. *Top row (under-reconstruction)*: When small-scale geometry (black outline) is insufficiently covered, we clone the respective Gaussian. *Bottom row (over-reconstruction)*: If small-scale geometry is represented by one large splat, we split it in two.

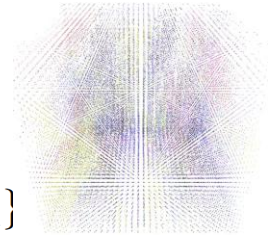
3D Gaussian Splatting (3DGS)

B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *SIGGRAPH 2023*

GaussianCube – 用于三维生成的结构化3D高斯

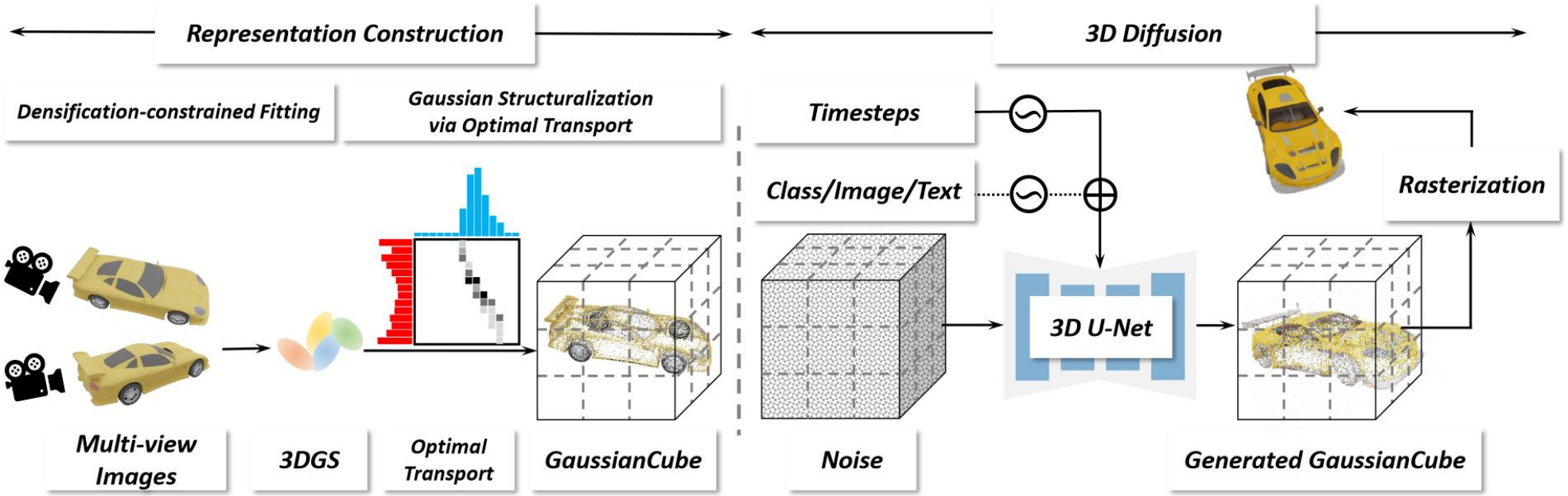
- 用固定点数 (32^3) 的高斯做场景拟合，再用最优传输(Optimal Transport)算法将优化好的高斯分配到规整三维网格(并记录偏移)

$$
\begin{aligned}
\underset{\mathbf{T}}{\text{minimize}} \quad & \sum_{i=1}^{N_{\max}} \sum_{j=1}^{N_{\max}} \mathbf{T}_{ij} \mathbf{D}_{ij} \\
\text{subject to} \quad & \sum_{j=1}^{N_{\max}} \mathbf{T}_{ij} = 1 \quad \forall i \in \{1, \ldots, N_{\max}\} \\
& \sum_{i=1}^{N_{\max}} \mathbf{T}_{ij} = 1 \quad \forall j \in \{1, \ldots, N_{\max}\} \\
& \mathbf{T}_{ij} \in \{0, 1\} \quad \forall (i, j) \in \{1, \ldots, N_{\max}\} \times \{1, \ldots, N_{\max}\}
\end{aligned}
$$

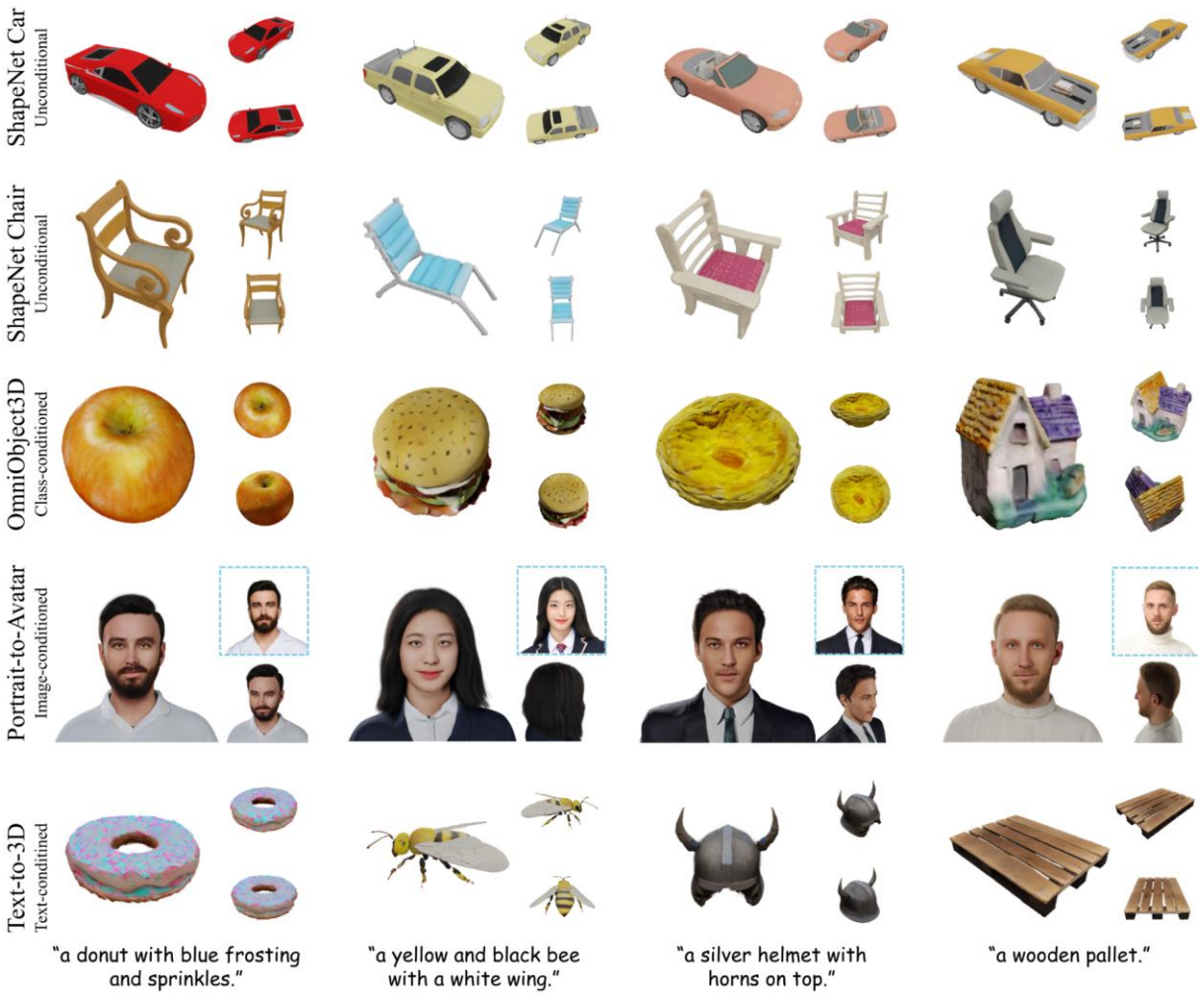


- 标准3D-Unet 扩散模型建模

算法流程



GaussianCube – 用于三维生成的结构化3D高斯



Method	ShapeNet Car		ShapeNet Chair		OmniObject3D	
	FID-50K↓	KID-50K(‰)↓	FID-50K↓	KID-50K(‰)↓	FID-50K↓	KID-50K(‰)↓
EG3D	30.48	20.42	27.98	16.01	-	-
GET3D	17.15	9.58	19.24	10.95	-	-
DiffTF	51.88	41.10	47.08	31.29	46.06	22.86
Ours	**13.01**	**8.46**	**15.99**	**9.95**	**11.62**	**2.78**

Table 3: Quantitative results of unconditional generation on ShapeNet Car and Chair [9] and class-conditioned generation on OmniObject3D [59].

Method	PSNR↑	LPIPS↓	SSIM↑	CSIM↑	FID-5K↓	KID-5K(‰)↓
Rodin w/o 2D SR	18.80	0.2842	0.7439	0.6594	32.07	24.78
Rodin	18.59	0.2821	0.7373	0.6466	20.02	9.24
Ours	**21.87**	**0.1768**	**0.7703**	**0.7821**	**8.32**	**2.67**

Table 4: Quantitative results of digital avatar creation conditioned on single portrait image.

	DreamGaussian	VolumeDiffusion	Shap-E	LGM	Ours
CLIP Score↑	26.38	24.41	30.52	30.06	**30.56**
Inference Time (s)↓	~ 120	5	7	**2**	5

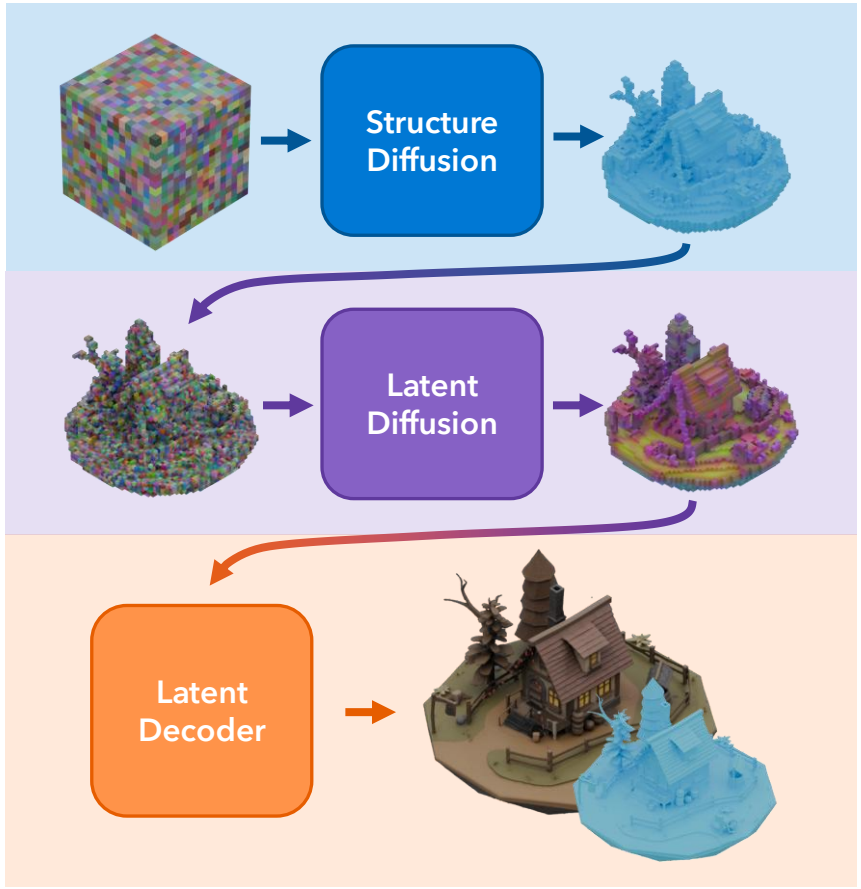
Table 5: Quantitative results of text-to-3D creation. Inference time is measured on a single A100 GPU. While Shape-E, LGM achieve comparable CLIP scores as ours, they either utilize millions of training data or leverage 2D diffusion prior.

GaussianCube – 用于三维生成的结构化3D高斯

B. Wang, Y. Cheng, J. Yang, C. Wang, et al. GaussianCube: A Structured and Explicit Radiance Representation for 3D Generative Modeling. *arXiv 2024*

Sparse Latent Diffusion (To appear)

- 通用多任务3D生成大模型
  - Text-to-3D / Image-to-3D
  - Gaussian/Mesh/Triplane output
- 三阶段生成框架
  - Structure Diffusion
  - Latent Diffusion
  - 3D Decoder
- 多数据集大规模训练
  - Objaverse等



块状、橙色和青色机器人，具有铰接式四肢

一个配备太阳能电池板和对接舱的未来空间站

身着传统盔甲的武士

一间质朴的小木屋，有石烟囱和木制门廊

时尚、具有未来感的银色和蓝色宇宙飞船模型

带有木质框架和毛绒床垫的床

陶瓷花盆中的盆景树，具有精细的叶子和树枝

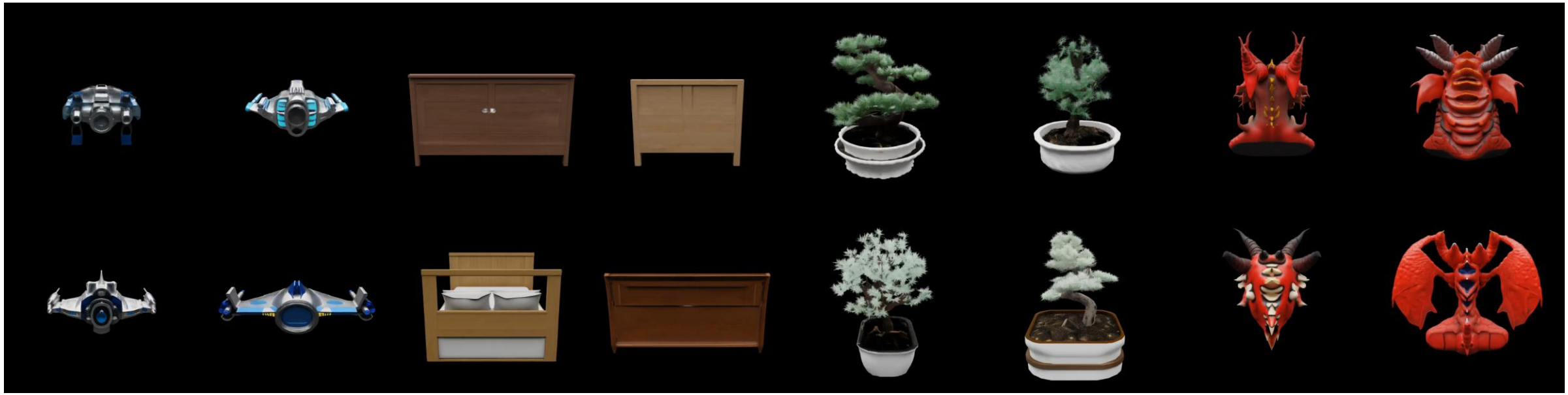
狰狞的红色龙头，长有尖牙利角，呈咆哮姿态

块状、橙色和青色机器人，具有铰接式四肢

一个配备太阳能电池板和对接舱的未来空间站

身着传统盔甲的武士

一间质朴的小木屋，有石烟囱和木制门廊



时尚、具有未来感的银色和蓝色宇宙飞船模型

带有木质框架和毛绒床垫的床

陶瓷花盆中的盆景树，具有精细的叶子和树枝

狰狞的红色龙头，长有尖牙利角，呈咆哮姿态

Scene composed of 7 assets generated by our model.

总结与一些思考

- 扩散模型成为3D生成主流方法
  - GAN方法基本被抛弃
- 数据仍然是瓶颈
  - 算力充足情况下模型能力未见顶
  - 高质量、大规模三维数据是关键
- 物体生成质量接近产品级，融入现有工作流程是关键
  - Mesh生成与高质量布线、PBR材质、形状纹理细节控制、便捷编辑与驱动等
- 4D生成、大场景生成是下一步重点
  - 对标视频生成：相近视觉质量、更三维一致、更可控

Thanks