CHINA 3DV 2025 中国三维视觉大会 The 4th China 3D Vision Conference

世界模型构建：重建，生成与推演

Building World Model: Reconstruction, Generation, and Inference

张兆翔
中国科学院自动化研究所

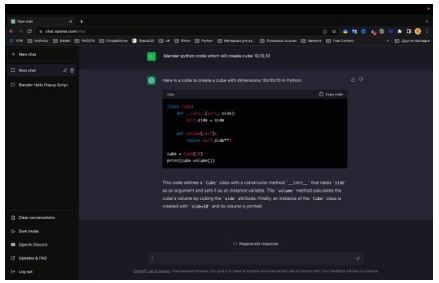
2025年4月12日，北京

人工智能蓬勃发展

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

❑人工智能创新发展不断加速，现象级成果层出不穷。

ChatGPT



- 多轮自然语言对话
- 跨领域知识理解
- 复杂逻辑推理
- 上下文记忆理解

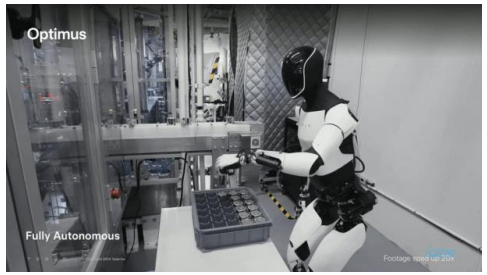
2022

Sora

- 高质量视频生成
- 多模态信息整合
- 高度时序一致性
- 高度生成可控性

2023

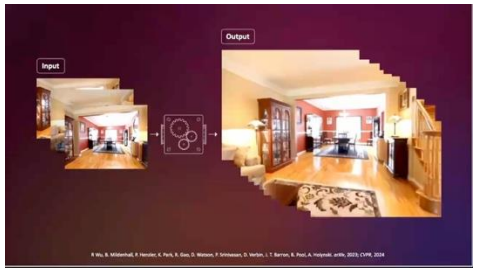
Optimus



- 感知-行动闭环
- 环境交互能力
- 持续学习与自适应
- 多模态协同

2024

空间智能



- 空间感知与定位
- 环境建模与理解
- 动态路径规划
- 自适应导航

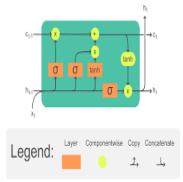
2024-

1、大模型技术成为创新驱动力

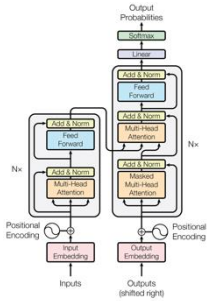
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

- 以大语言模型（LLM）为代表的大模型技术蓬勃发展，逐步通过大规模数据训练和人类反馈优化，实现更自然、更智能的人机交互。

Word2Vec



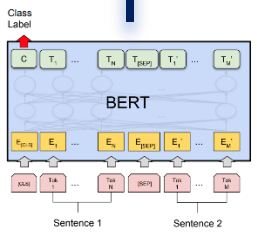
LSTM

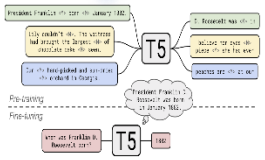


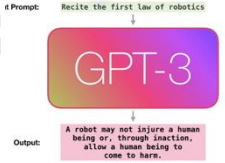
Transformer

ChatGPT GPT-4，LLaMA，Claude，Gemini，GPT-o1 ...

2013 2014 2017 2018 2019 2020 2022 2023





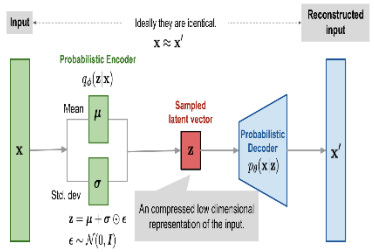


BERT & GPT　GPT2 & T5　GPT3

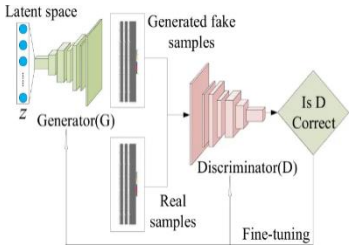
早期深度学习阶段　预训练模型的兴起　大模型时代

2、生成式技术成为创新创业热潮

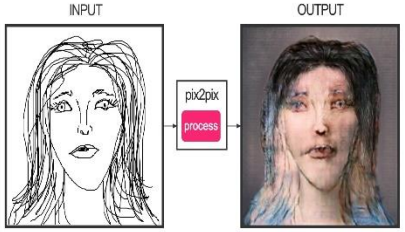
❑生成式模型（Generative Models）经历了从简单概率模型到深度学习，再到大规模预训练模型的演进，不断提升生成内容的质量和多样性。



VAE
(2013)



GAN
(2014)



Pix2Pix
(2016)

DALL·E

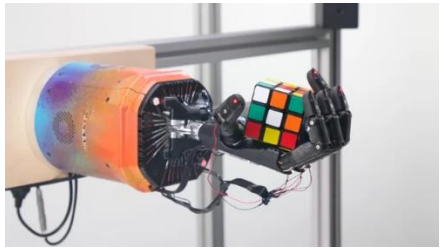
DALL-E & CLIP
(2021)

Sora
(2023)

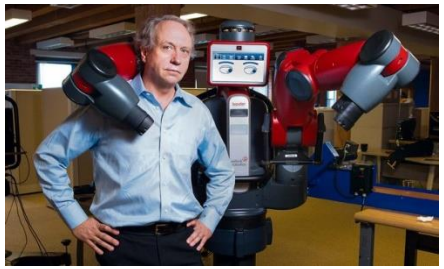
Diffusion Models
(2022)

3、具身智能成为热点研究方向

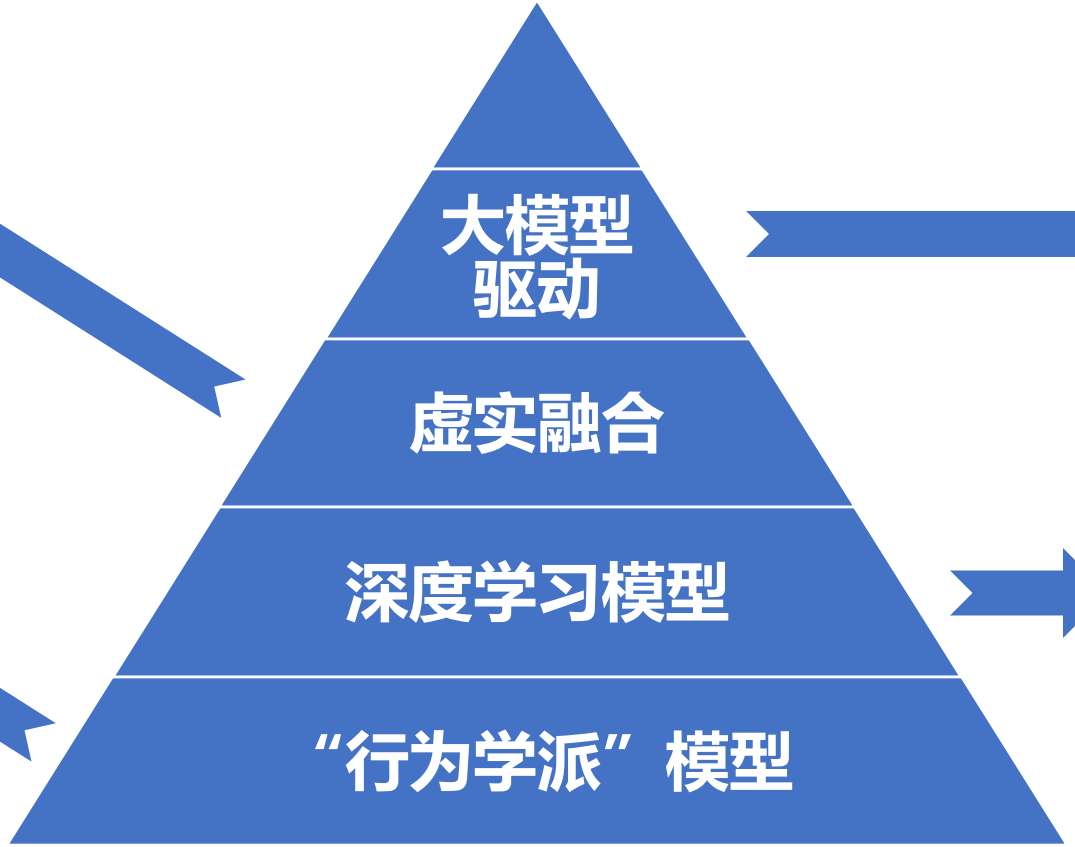
❑具身智能（Embodied Intelligence）从早期的规则和感知控制方法，逐步演进到大模型驱动的具身智能，实现了更丰富的人机互动和物理环境中的智能行为。



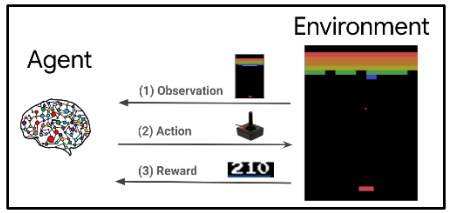
Rubik’s Cube
(2018-2021)



IRobot（80-90年代）



SayCan，CAIRaoke
(2022-)



DQN、PPO
(2010年代初期)

三个脉络的交叉点是什么？

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

世界模型

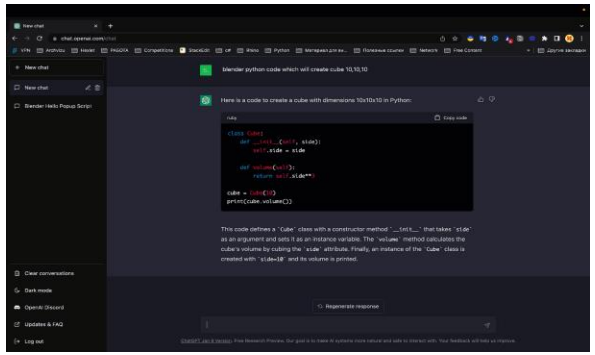
- 真实场景重建
- 多元环境生成
- 推演未来状态

提供新的内在驱动

提供新的实现手段

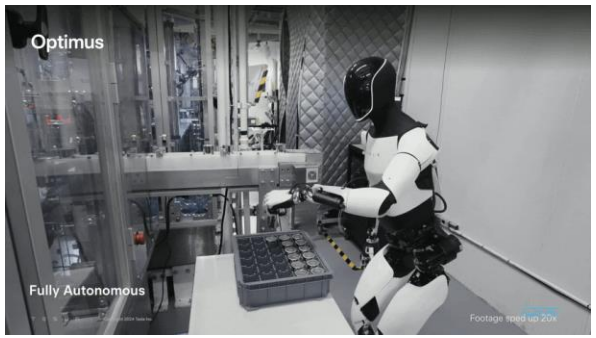
提供新的应用场景

大模型技术



生成式人工智能

具身智能



世界模型：构建必备的要素

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

世界模型

现实世界理解 | 预测运动规律 | 全新场景扩充

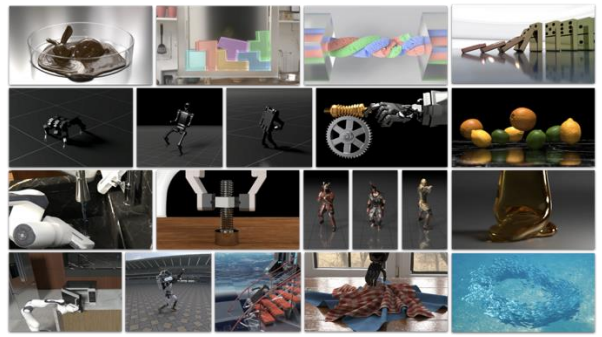
真实场景重建

- 视觉与几何精细化复刻
- 提供真实世界验证途径

推演未来状态

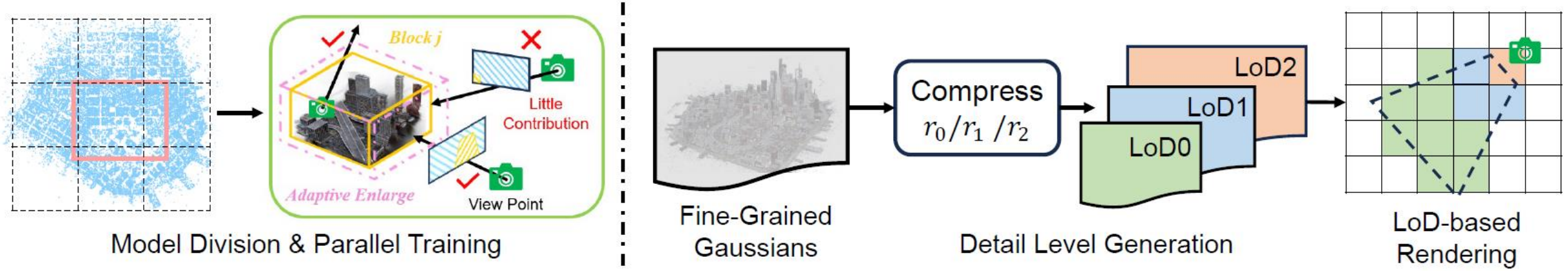
- 预测多元场景下的运动趋势
- 提供决策依据

多元环境生成



- 大数据驱动，定制需求场景
- 提升复杂场景模拟泛化能力

大规模场景重建：CityGaussian系列



- Divide and Conquer. 将 3DGS 划分为一系列子模型，自适应分配数据并用不同的GPU并行训练
- Level of Details (LoD). 根据视锥范围和奈奎斯特采样定律确定需要加载的子模型以及对应的细节层次，减少冗余运算开销，同时缓解混叠现象

LoD Selection	SSIM↑	PSNR↑	LPIPS↓
ds1, distance	**0.784**	**24.90**	**0.256**
ds1, Nyquist	0.781	24.88	0.259
ds2, distance	**0.807**	23.94	**0.193**
ds2, Nyquist	0.800	**24.24**	0.199
ds4, distance	0.713	21.46	0.211
ds4, Nyquist	**0.724**	**22.15**	**0.198**
ds8, distance	0.603	19.50	0.236
ds8, Nyquist	**0.617**	**20.04**	**0.216**

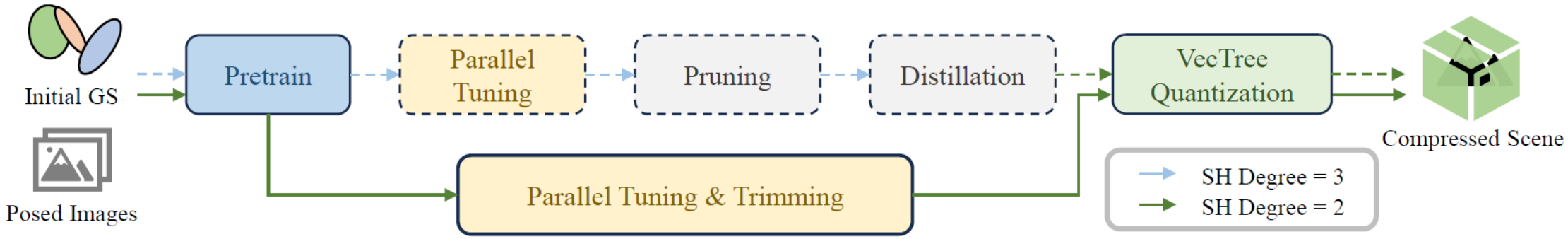
我们的LoD方案显著优于基于距离阈值的方式

CityGaussian: Real-time High-quality Large-Scale Scene Rendering with Gaussians (ECCV2024)

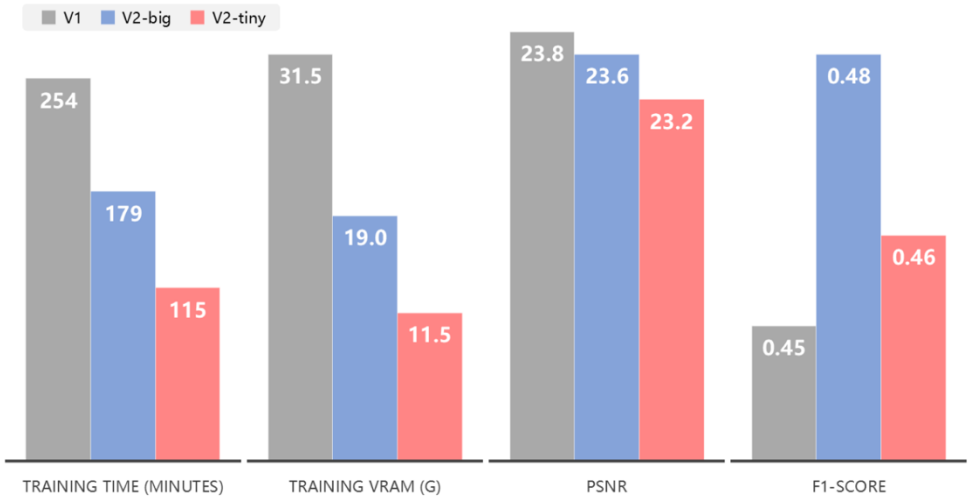
大规模场景重建：CityGaussian系列

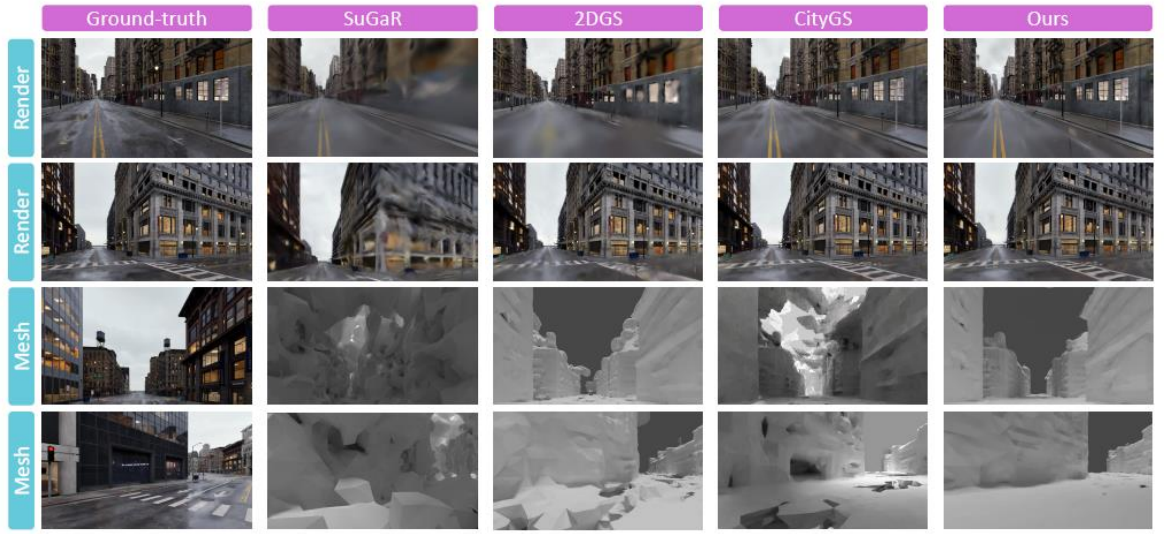
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

CityGaussianV2总体管线



新管线分别将训练用时与显存开销降低25%和50%，实现了针对2DGS的量化压缩，同时具备对街景的泛化能力

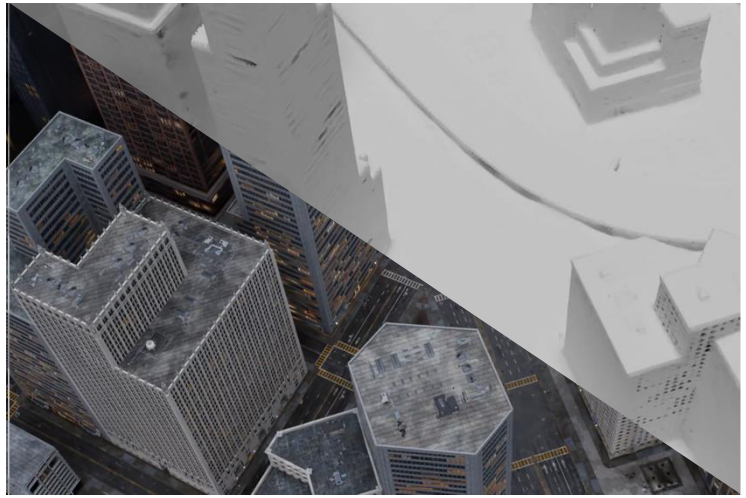
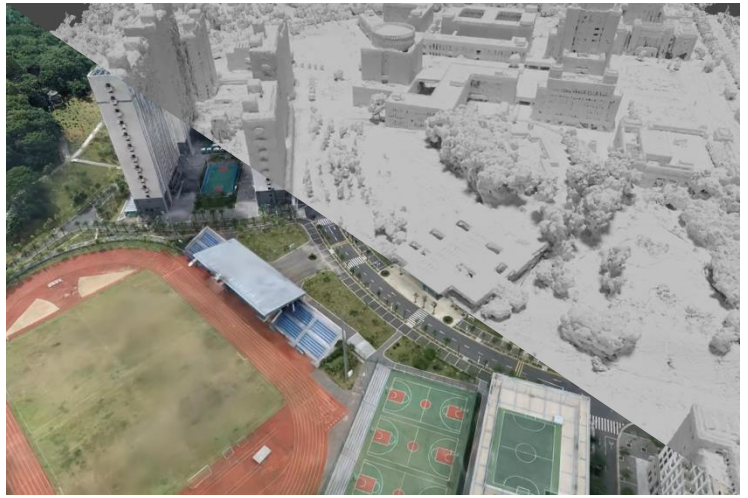




CityGaussianV2: Efficient and Geometrically Accurate Reconstruction for Large-Scale Scenes(ICLR2025)

大规模场景重建：CityGaussian系列

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

Methods	GauU-Scene				MatrixCity-Aerial				MatrixCity-Street			
	PSNR↑	P↑	R↑	F1↑	PSNR↑	P↑	R↑	F1↑	PSNR↑	P↑	R↑	F1↑
NeuS	14.46	FAIL	FAIL	FAIL	16.76	FAIL	FAIL	FAIL	12.86	FAIL	FAIL	FAIL
Neuralangelo	NaN	NaN	NaN	NaN	19.22	0.080	0.083	0.081	15.48	FAIL	FAIL	FAIL
SuGaR	22.72	0.570	0.292	0.377	OOM	OOM	OOM	OOM	19.82	0.053	0.111	0.071
GOF	22.33	0.370	0.390	0.374	17.42	FAIL	FAIL	FAIL	20.32	0.219	0.473	0.300
2DGS	23.93	0.553	0.446	0.491	21.35	0.207	0.390	0.270	21.50	0.334	0.659	0.443
CityGS	24.75	0.457	0.371	0.407	27.46	0.362	0.637	0.462	22.98	0.283	0.689	0.401
Ours	24.51	0.576	0.450	0.501	27.23	0.441	0.752	0.556	22.19	0.376	0.759	0.503

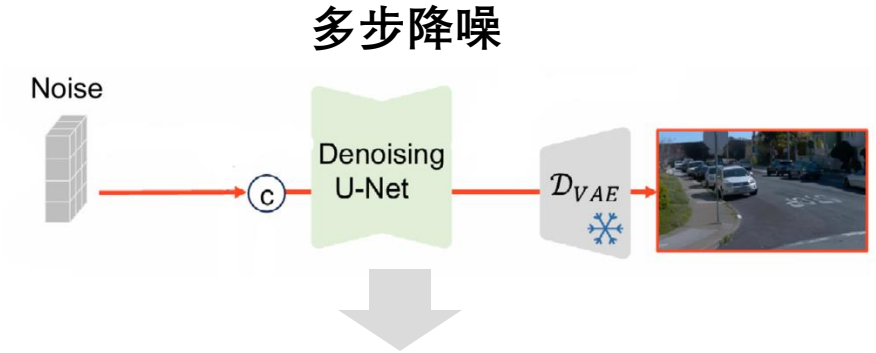
我们的算法很好地平衡了渲染质量PSNR与几何精度F1-Score

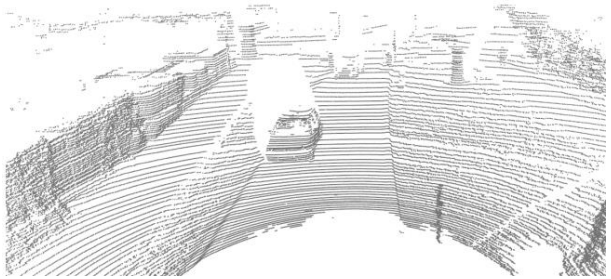
自动驾驶的重建生成一体化

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

- 现有新视角生成方法依赖重建->渲染管线，其对物理场景的观察视角受已有实际观测约束。
- 提出以**生成式模型**驱动视角生成，实现对**真实场景**的自由视角成像
- 基于三维场景的点云、颜色先验，严格约束生成内容符合真实场景的同时，合理补全缺失场景信息

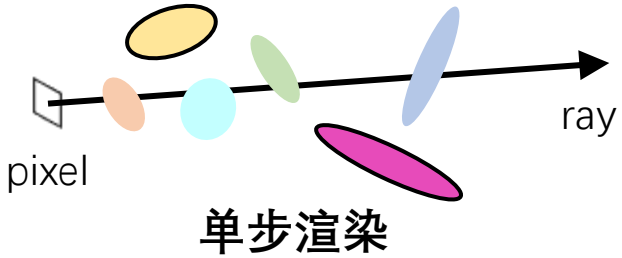
FreeVS
ICLR 2025

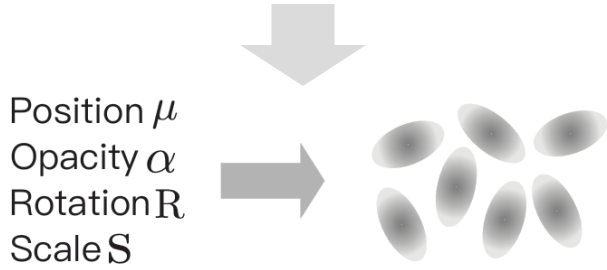




点云先验

FreeSim
CVPR 2025





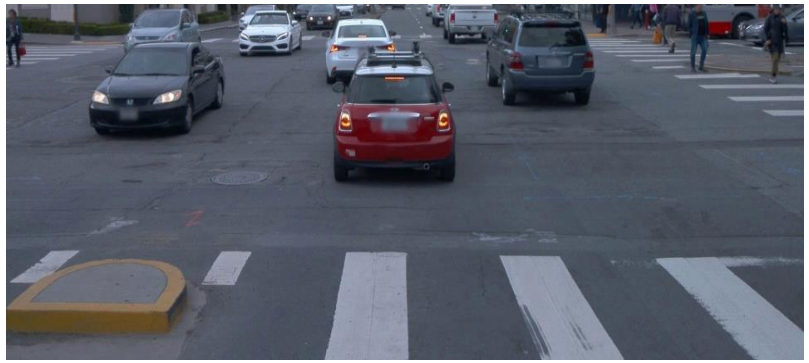
3DGS先验

- 从多步生成走向单步**实时渲染**
- 从激光雷达先验走向视觉GS先验

FreeVS: Generative View Synthesis on Free Driving Trajectory. (ICLR2025)
FreeSim: Toward Free-viewpoint Camera Simulation in Driving Scenes. (CVPR2025)

自动驾驶的重建生成一体化

- 自由合成在**真实场景**中，已记录轨迹之外的**虚构轨迹**上的相机视角
  - 可在驾驶车辆原本并未移动的场景模拟车辆移动:



  - 能模拟车辆变线行驶，甚至能模拟车辆撞向行人:

FreeVS: Generative View Synthesis on Free Driving Trajectory. (ICLR2025)

自动驾驶的重建生成一体化

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

与其他方法的对比演示:

1. 视点上升
2. 视点越过汽车
3. 向右移动

PVG

FreeSim (ours)

1. 正常驾驶
2. 快速靠边
3. 大幅转向

PVG

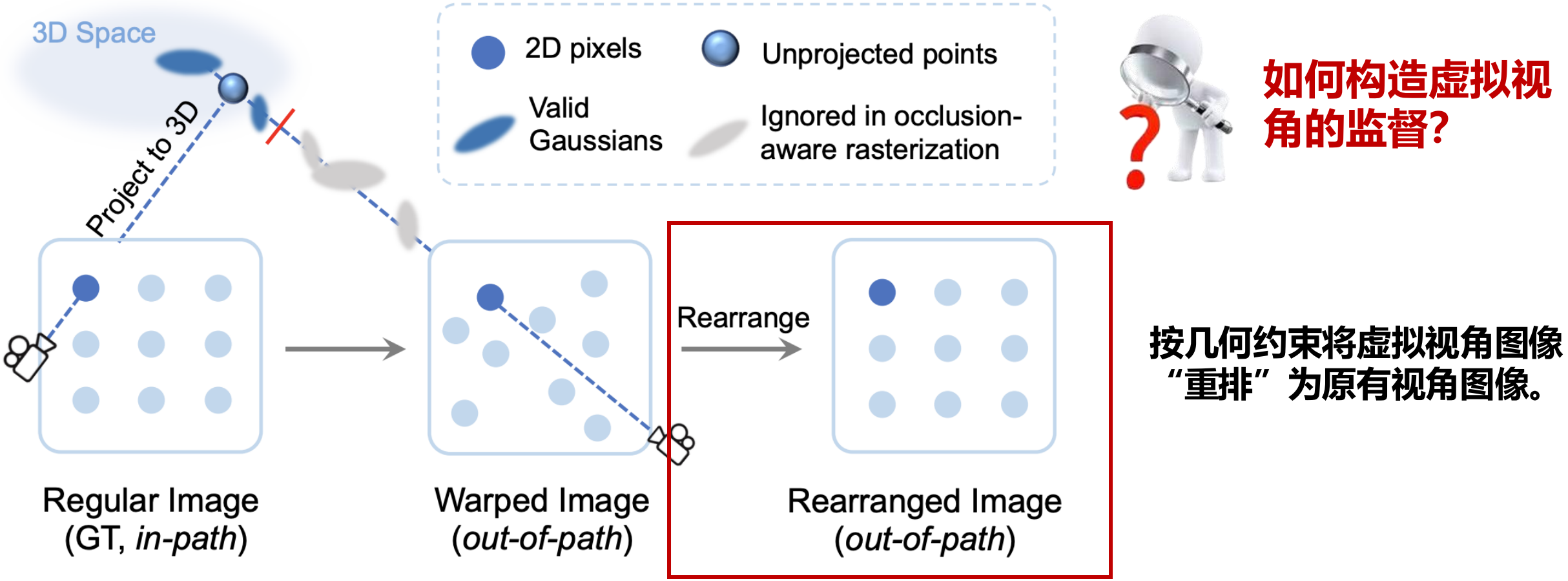
FreeSim (ours)

FreeSim: Toward Free-viewpoint Camera Simulation in Driving Scenes. (CVPR2025)

基于重建的可泛化街景仿真：FlexDrive

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

- 通过几何手段在**虚拟视角**上进行渲染



如何构造虚拟视角的监督?

按几何约束将虚拟视角图像“重排”为原有视角图像。

FlexDrive: Toward Trajectory Flexibility in Driving Scene Reconstruction and Rendering. (CVPR2025)

基于重建的可泛化街景仿真： FlexDrive

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

Cut-in 模拟 （与原有轨迹对比）

FlexDrive: Toward Trajectory Flexibility in Driving Scene Reconstruction and Rendering. (CVPR2025)

为重建提供物理真实性：MaterialSeg3D

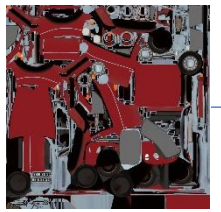
中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

- 基于**重建**的方法往往将**光源、纹理、材质**的影响**烘焙为整体**，因此不具备在新的光源下正确渲染的能力，也不具备仿真模拟的能力

缺乏物理真实性

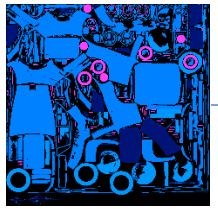
- **视觉逼真的渲染**需要依赖金属度、粗糙度等**PBR物理属性**
- **物理真实的仿真**需要摩擦系数、密度等真实的**物理材质属性**

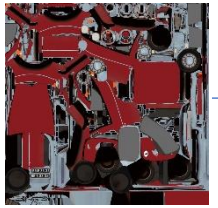

网格


颜色贴图


渲染效果


网格


材质贴图


颜色贴图


渲染效果

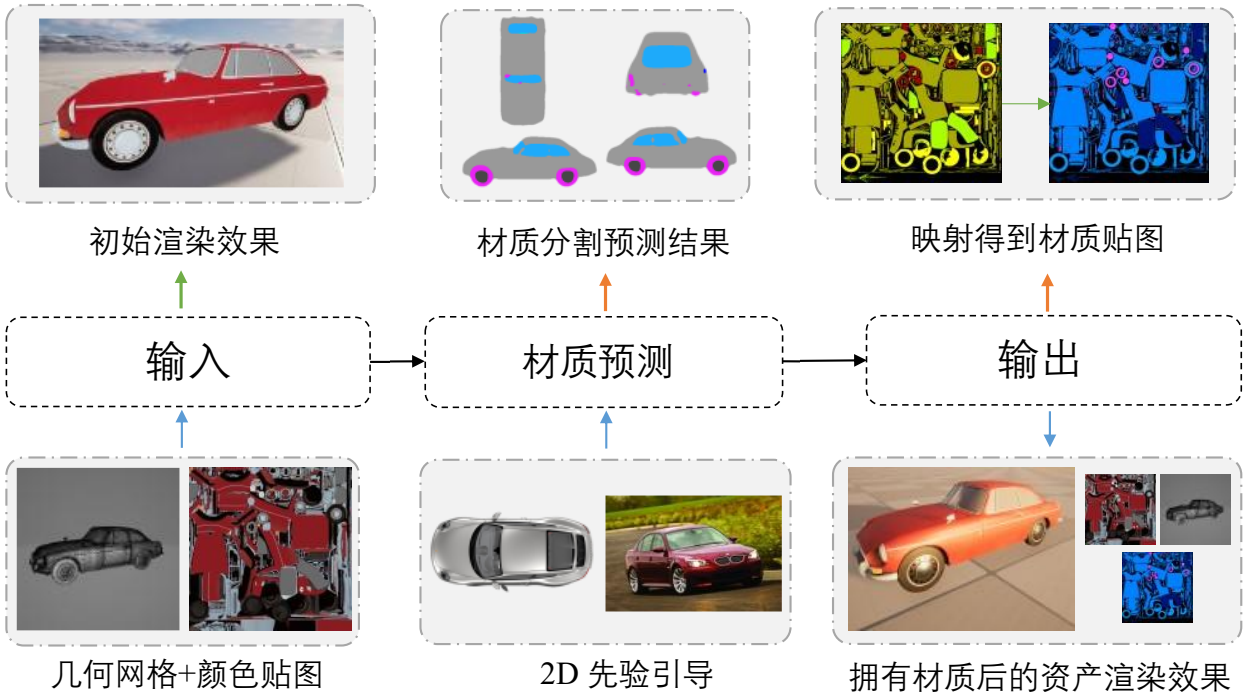
MaterialSeg3D: Segmenting Dense Materials from 2D Priors for 3D Assets (ACM MM 2024 Oral)

为重建提供物理真实性：MaterialSeg3D

中国科学院自动化研究所 模式识别实验室 New Laboratory of Pattern Recognition

将二维的材质认知提升到三维物体

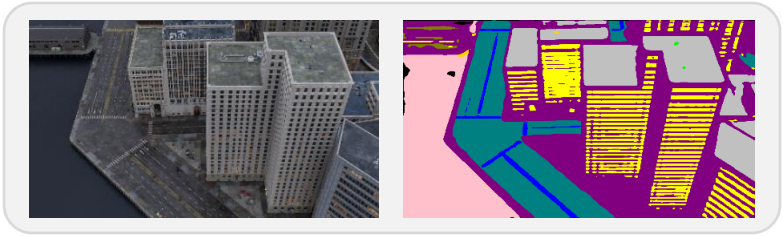
- ❖ 对**海量二维Web图片**中材质进行标注构建大规模材质数据库MIO。
- ❖ 利用模型学习**材质知识**进行多视角（41个视角）的预测。
- ❖ 在三维展开的表面积空间（**UV空间**）进行合并。



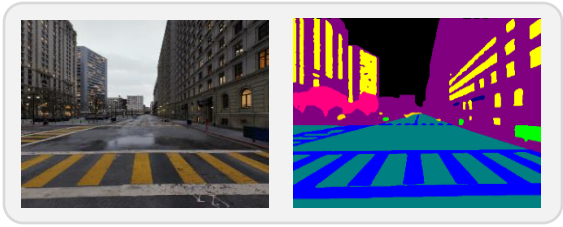
从三维物体扩展至三维场景

- ❖ 构建了面向场景(航拍/街景)的材质分割数据集。
- ❖ 对多视角的场景材质预测在表面展开空间进行聚合。

航拍



街景



- ❖ **物理材质分割而非通用语义分割**
  - 混凝土
  - 沥青
  - 粗糙玻璃
  - 油漆
  - 光滑玻璃
  - 粗糙木头

MaterialSeg3D: Segmenting Dense Materials from 2D Priors for 3D Assets (ACM MM 2024 Oral)

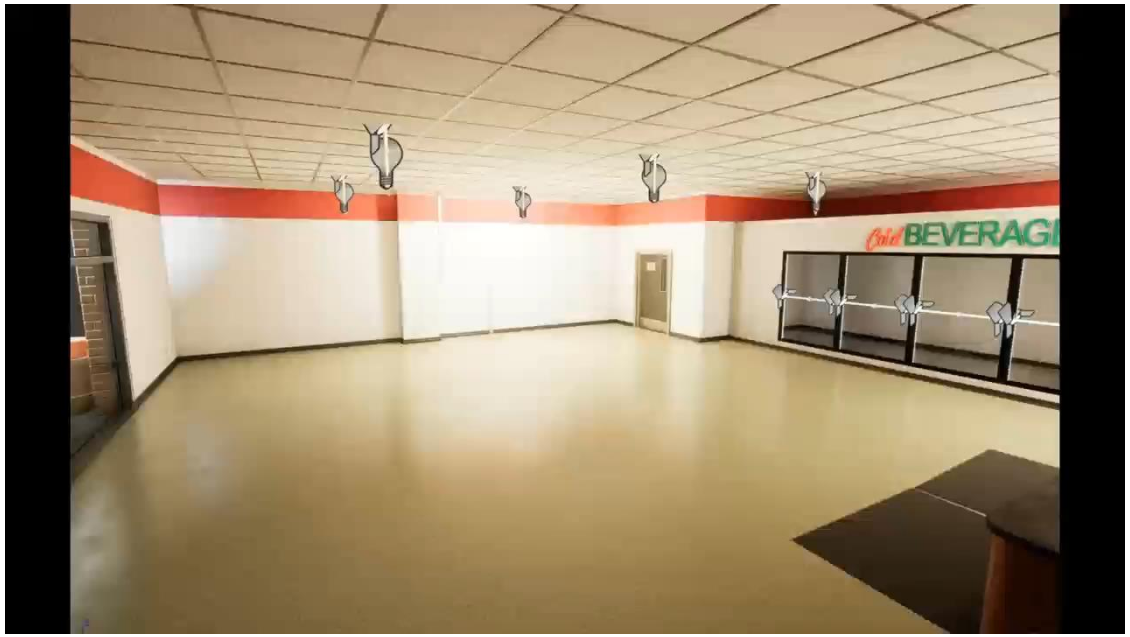
基于程序化生成的世界场景构建

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

❑ 基于视觉认知的重建方式在复杂场景的细节构建中仍有困难

- ○ 视频生成缺乏**精准三维结构**和**物理交互性**

基于已有机理驱动的程序化生成能力，利用LLM驱动程序化生成，实现具有复杂环境的可交互世界场景构建



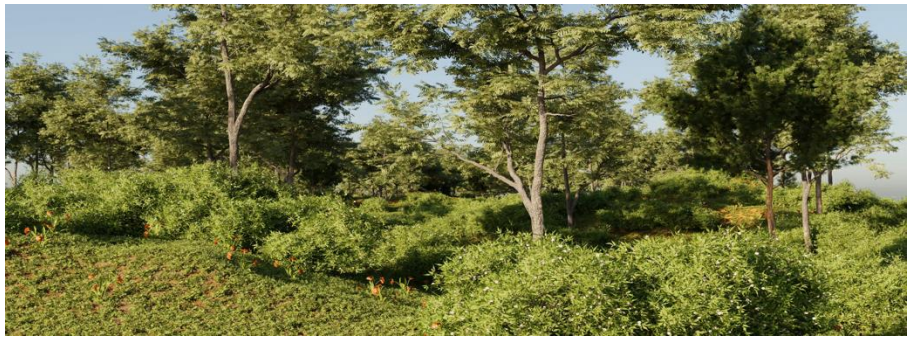
基于程序化生成的世界场景构建

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

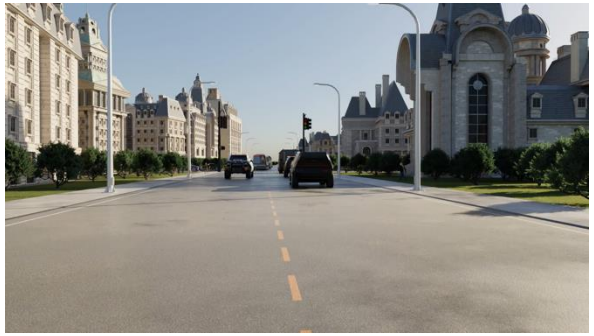
多模态驱动

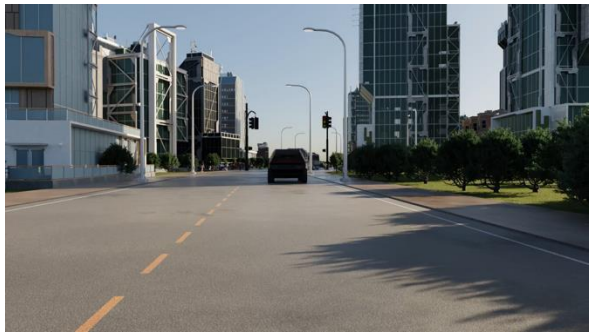
文本 | 草图 | 语义分割图 | 结构化数据 | 模态融合数据 | …

自然场景

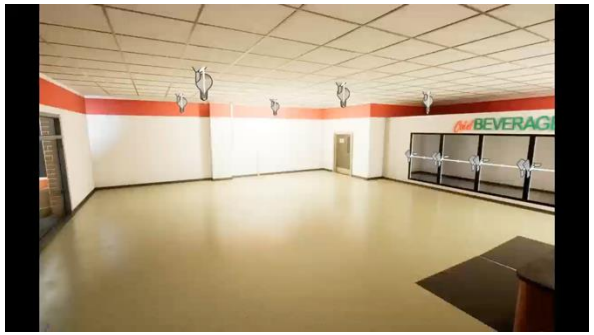


城市场景





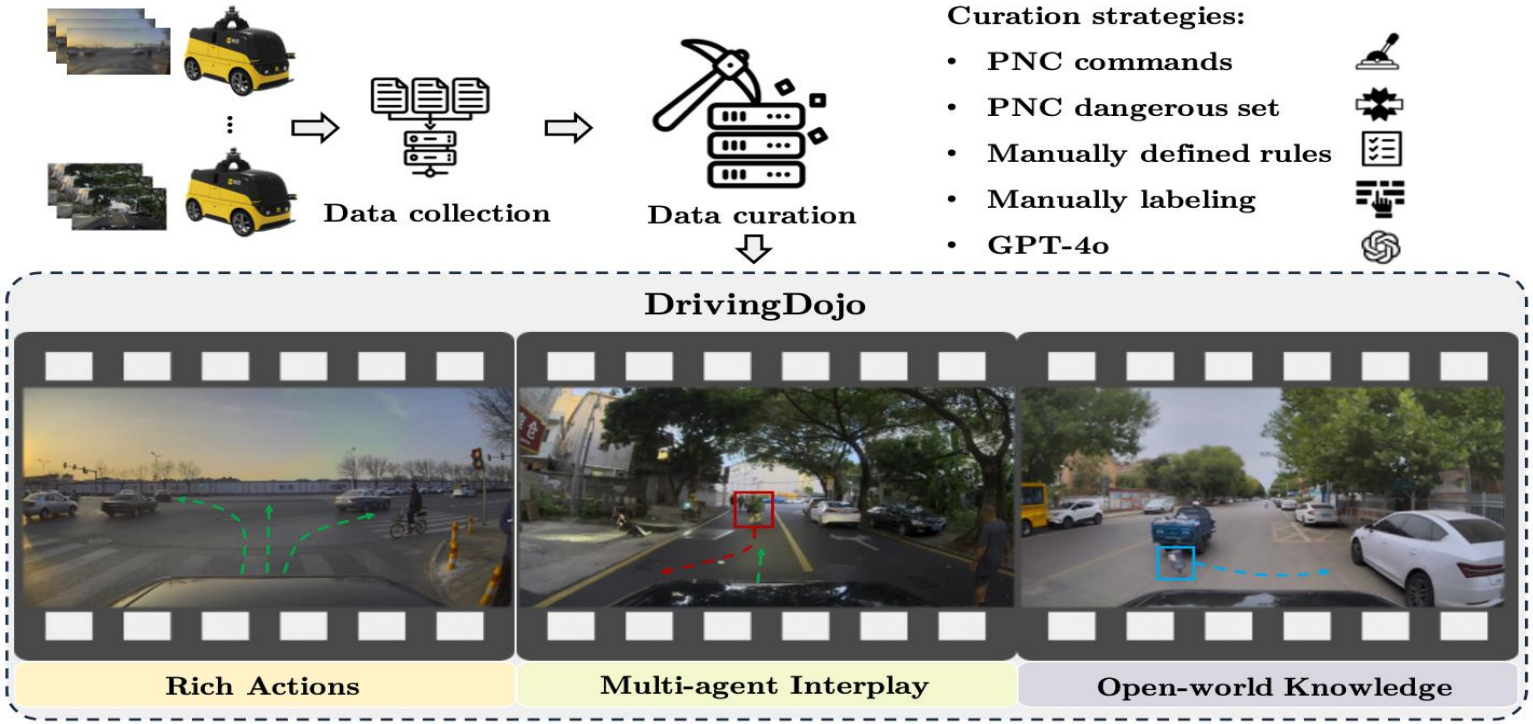
室内场景

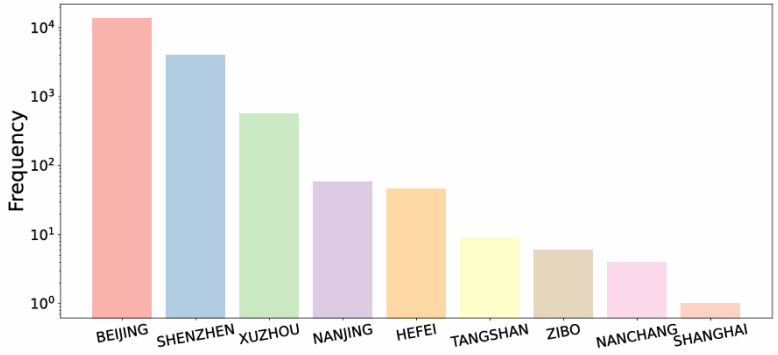


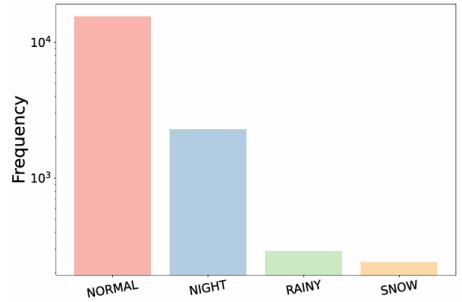
SceneX: Procedural Controllable Large-scale Scene Generation via Large-language Models (AAAI2025)

大规模世界模型推演数据集

- 联合美团无人车队，构建高质量的驾驶数据集，用于世界模型的研究
  - 数据采集：**8**个城市，**500**辆车，**2000小时/天**，多样外部环境的原始数据
  - 数据挖掘：**90万**案例库 -> **DrivingDojo数据集 (18K 视频)**







DrivingDojo Dataset: Advancing Interactive and Knowledge-Enriched Driving World Model (NeurIPS2024)

大规模世界模型推演数据集

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

❑ 运动指令跟随



❑ 运动指令的泛化

运动行为泛化

数据集泛化（倒车）

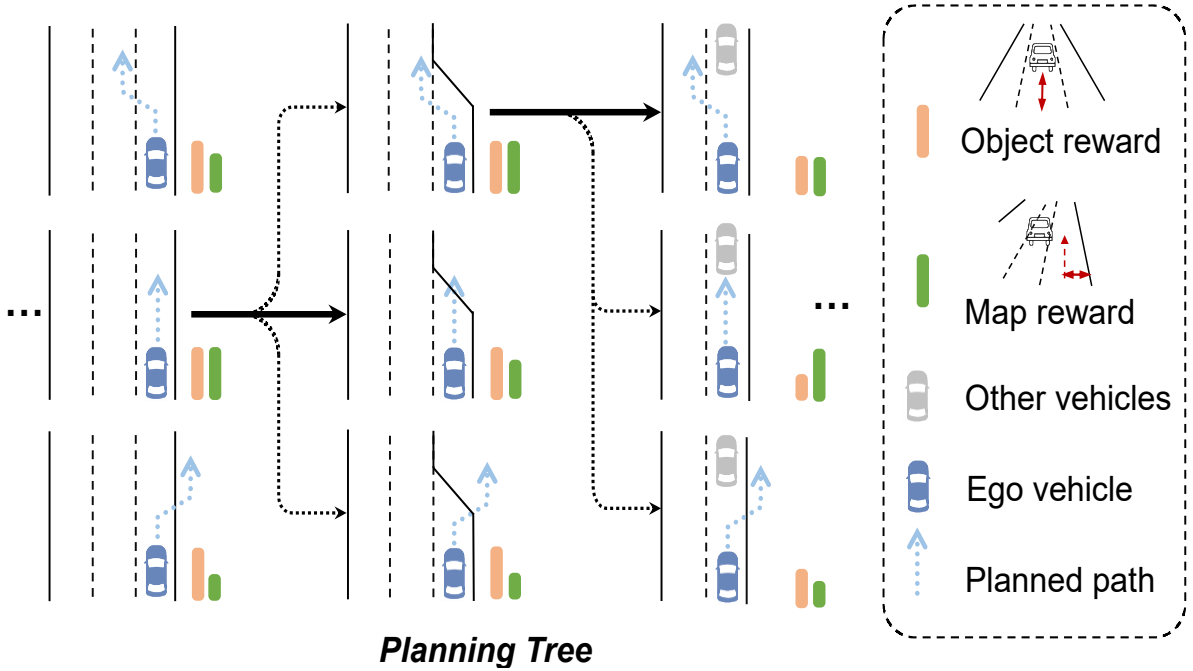
DrivingDojo Dataset: Advancing Interactive and Knowledge-Enriched Driving World Model (NeurIPS2024)

世界模型的推演与决策

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

世界模型根据
不同的动作输入生成未来情景

感知模型给出Reward，
形成决策树



Planning Tree

Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving. (CVPR2024)

世界模型的构建蓝图

真实场景重建

- 保证尺度规模
- 细节精细刻画
- 提供产业化支持

多元环境生成

- 扩充世界多样性
- 满足特化需求
- 增强泛化能力

推演未来状态

- 预测演化机理
- 模拟复杂场景
- 推动虚实融合

观点 **世界模型需要建立在对多样化世界场景的精准复刻与扩充之上，而世界模型的有效运作则离不开合理有效的推演进程。**

感谢！

张兆翔
中国科学院自动化研究所
zhaoxiang.zhang@ia.ac.cn
https://zhaoxiangzhang.net/

2025年4月12日，北京