# 面向自动驾驶的写实可控视觉仿真

廖依伊

June 8, 2024

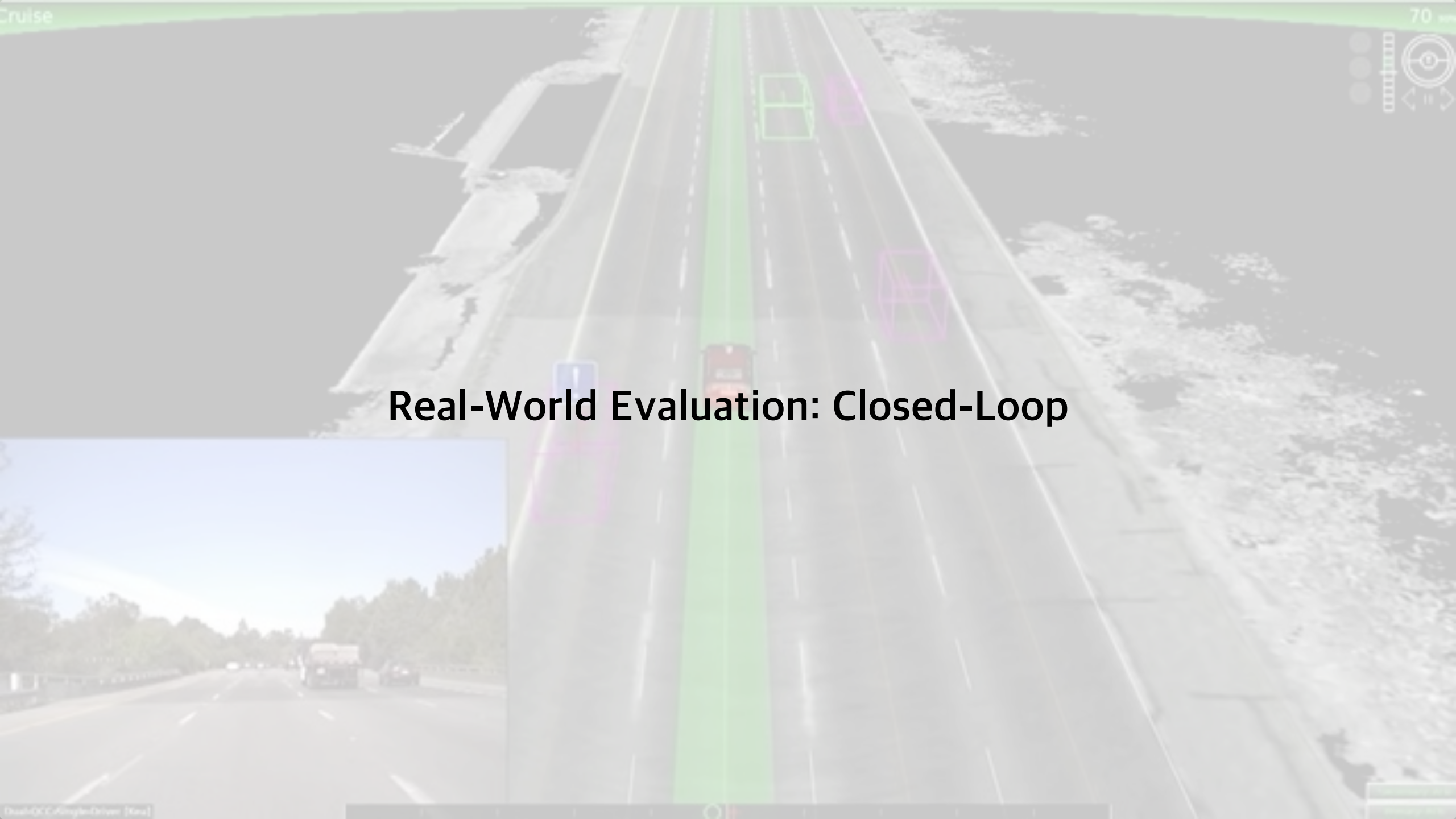

浙江大学
ZHEJIANG UNIVERSITY

# Autunomous Driving Datasets





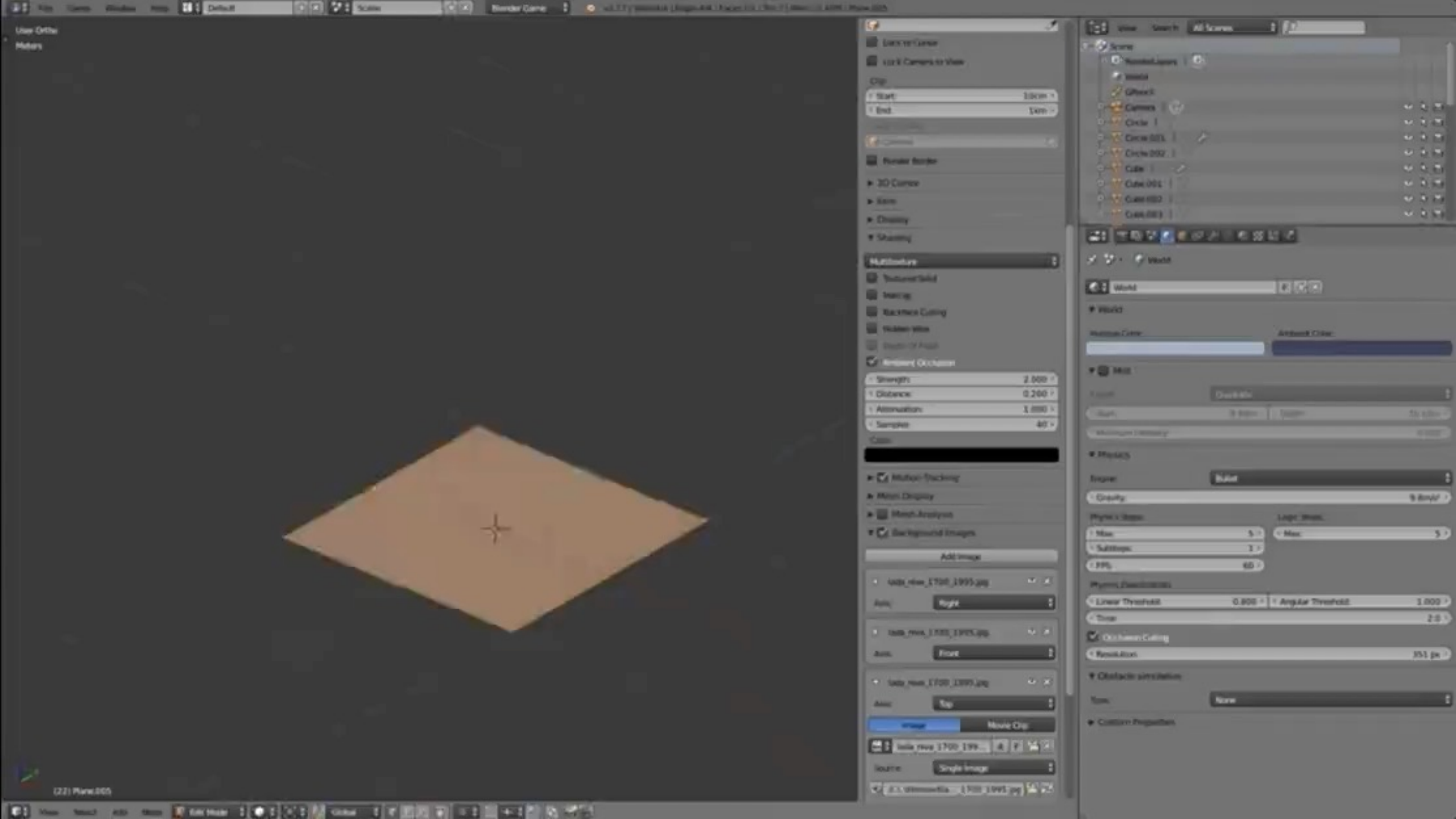
**Open-Loop Evaluation**

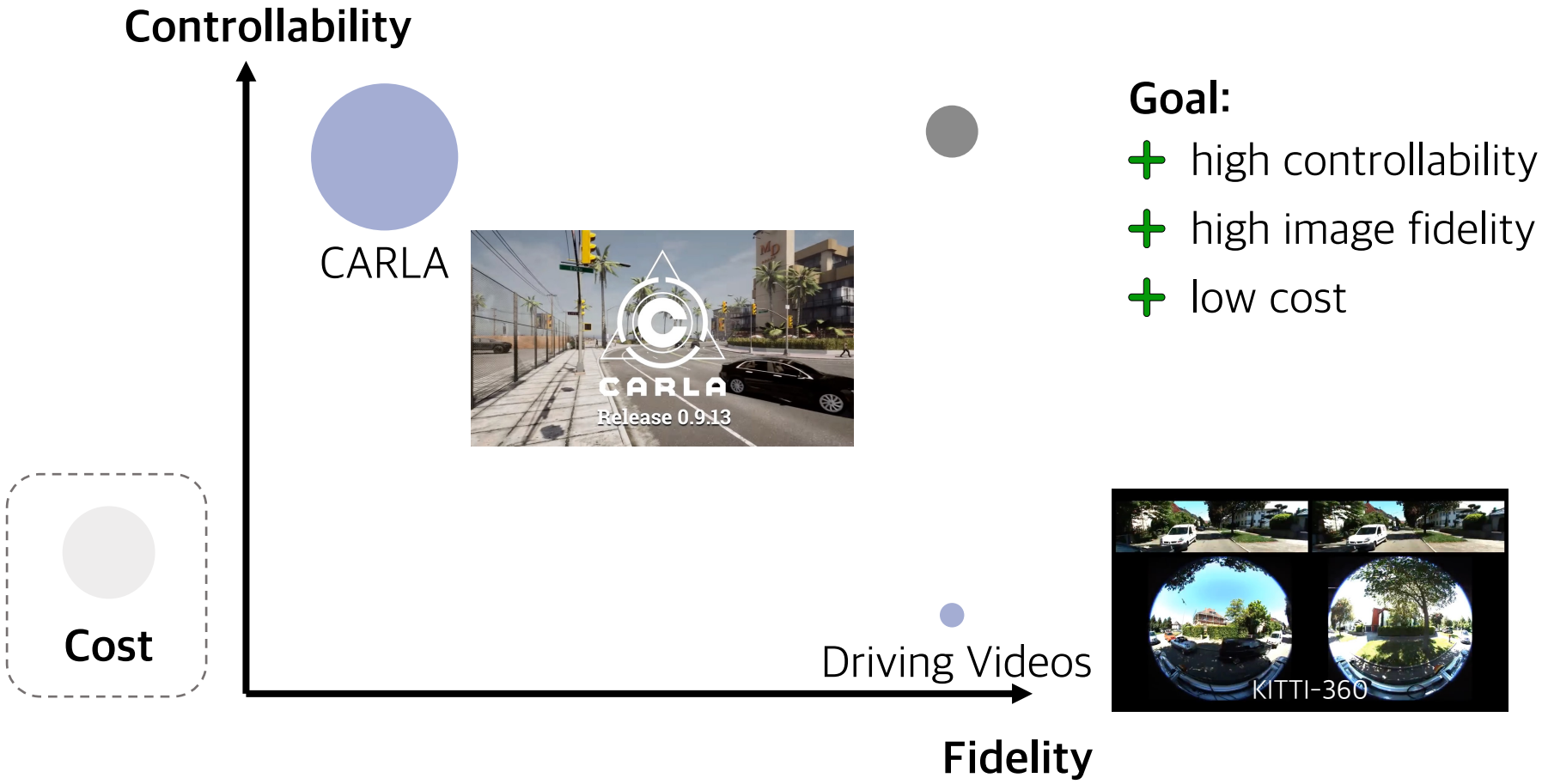

# Real-World Evaluation: Closed-Loop


# CARLA

## Release 0.9.13



# Towards Autonomous Driving Simulator

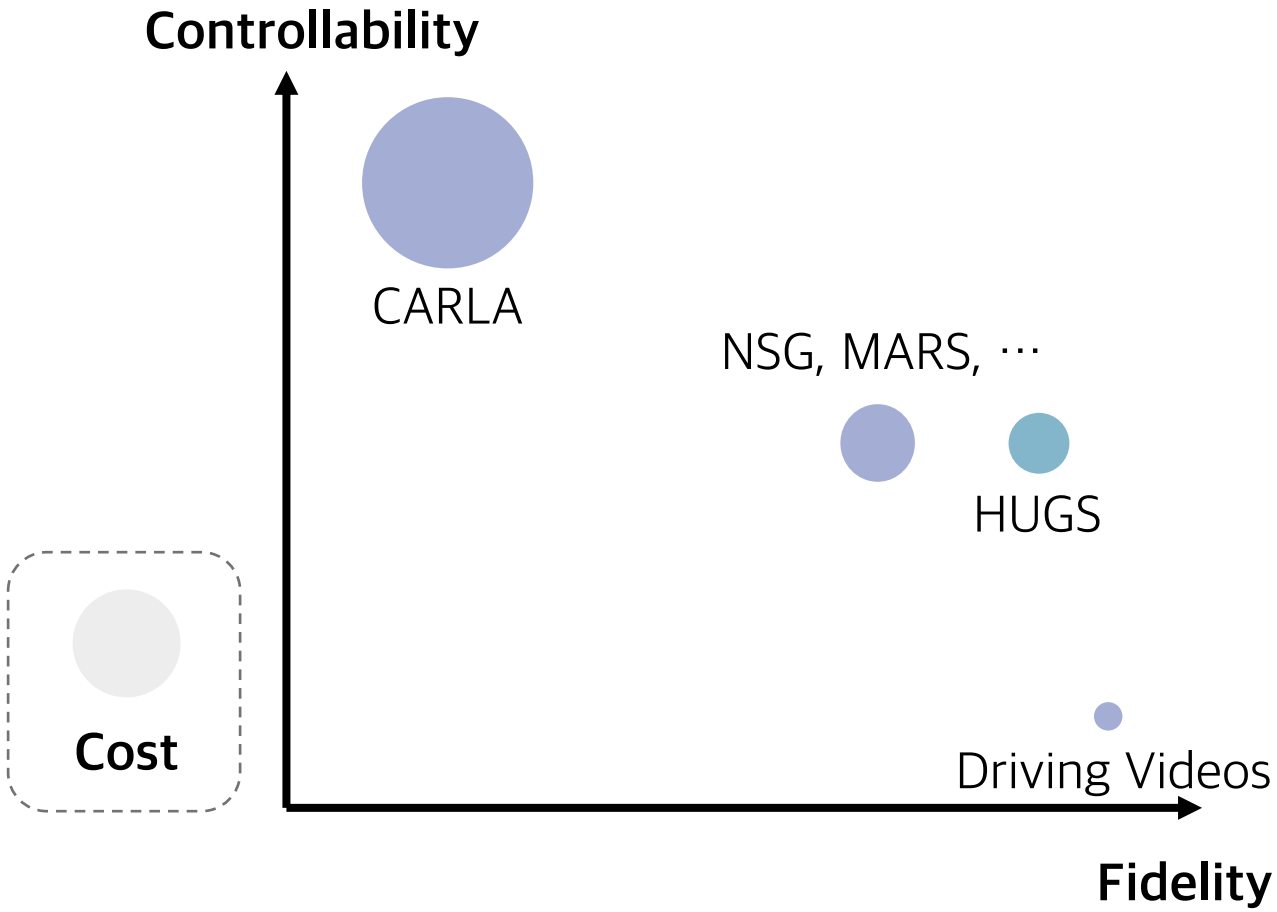



**Goal:**

+ high controllability
+ high image fidelity
+ low cost

Liao, Xie, Geiger: KITTI-360: A Novel Dataset and Benchmarks for Semantic Scene Understanding in 2D and 3D. TPAMI, 2022

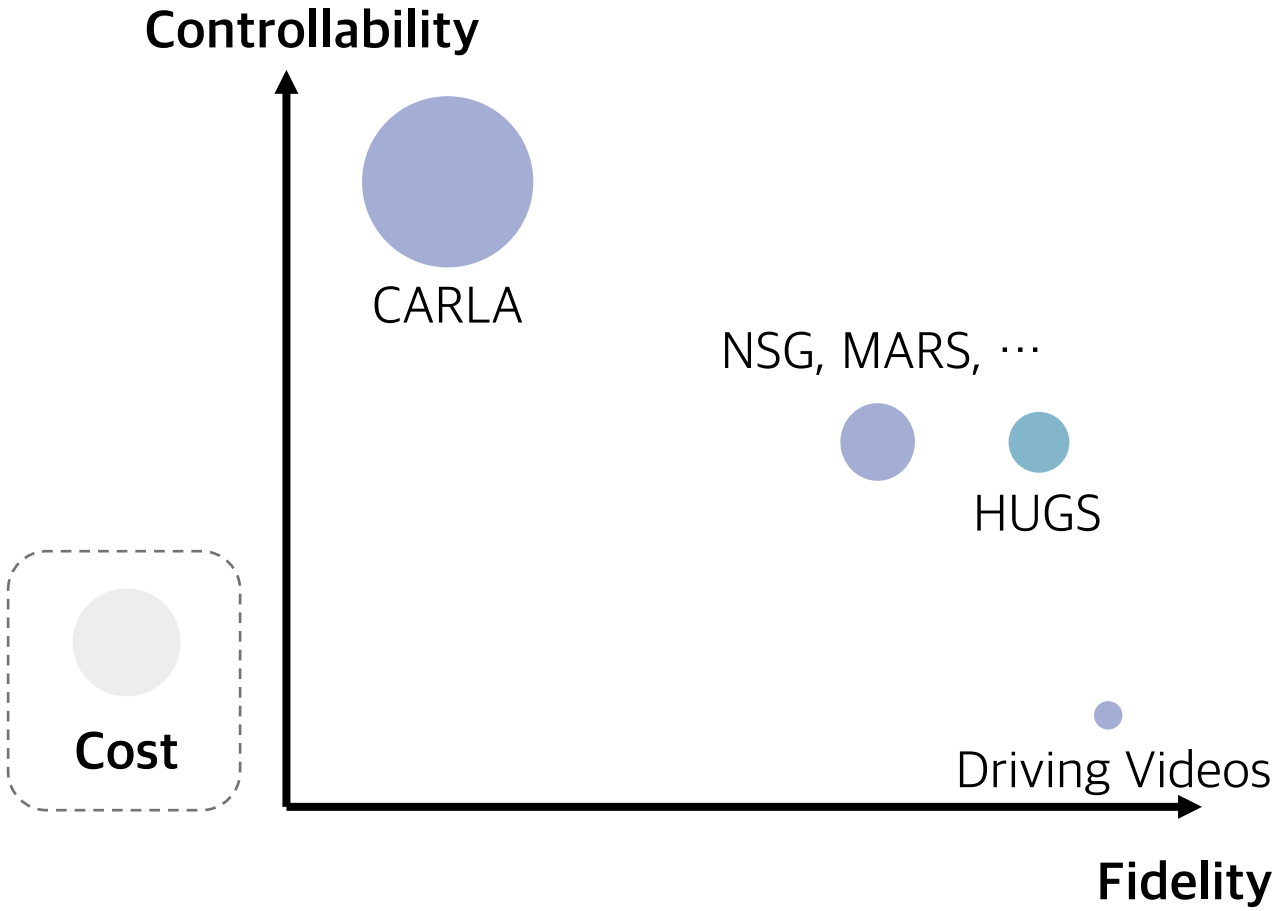
# Towards Autonomous Driving Simulator




**Reconstruction:**

- \+ control over viewpoints
- \+ control over dynamic objects
- − image fidelity
- − ground truth 3D annotations

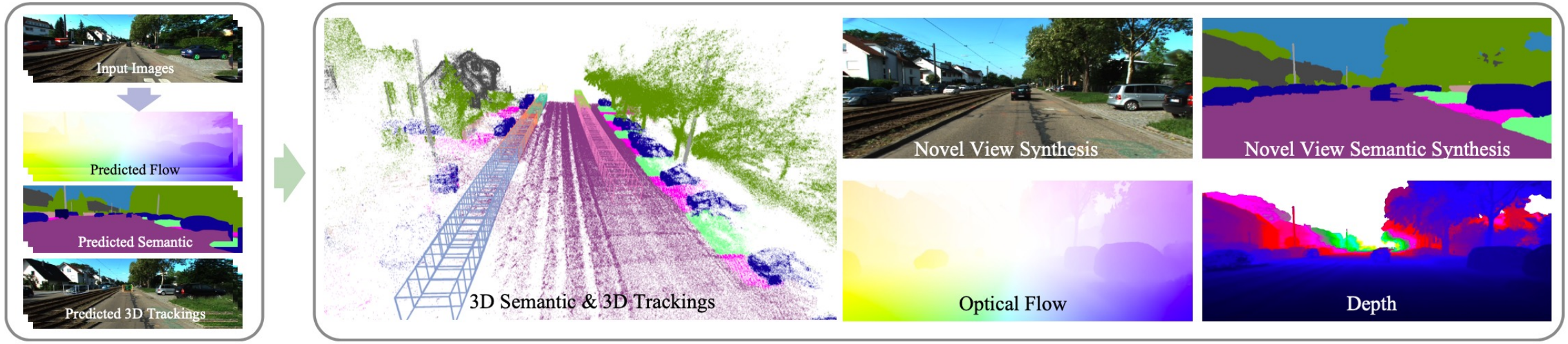
# Towards Autonomous Driving Simulator




**Reconstruction:**

- \+ control over viewpoints
- \+ control over dynamic objects
- \+ image fidelity
- \+ no ground truth 3D annotations

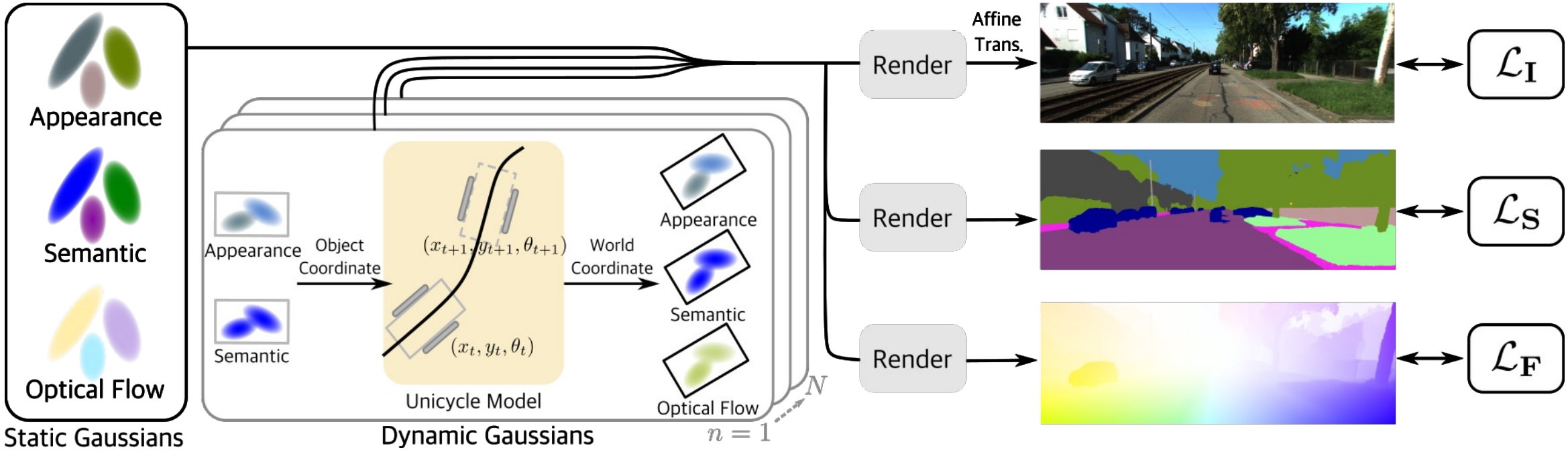
# HUGS: Holistic Urban Scene Understanding via 3D Gaussian Splatting




- Input: Posed RGB images, noisy 2D & 3D predictions
- Remove the dependency of **3D GT bboxes of static and dynamic objects**
- Enable hostistic scene understanding with **fast rendering** at ~100FPS

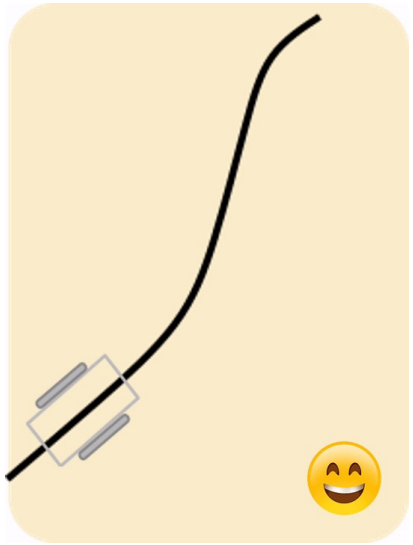
Zhou, Shao, Xu, Bai, Qiu, Liu, Wang, Geiger, Liao. HUGS: Holistic Urban 3D Scene Understanding via Gaussian Splatting. CVPR 2024

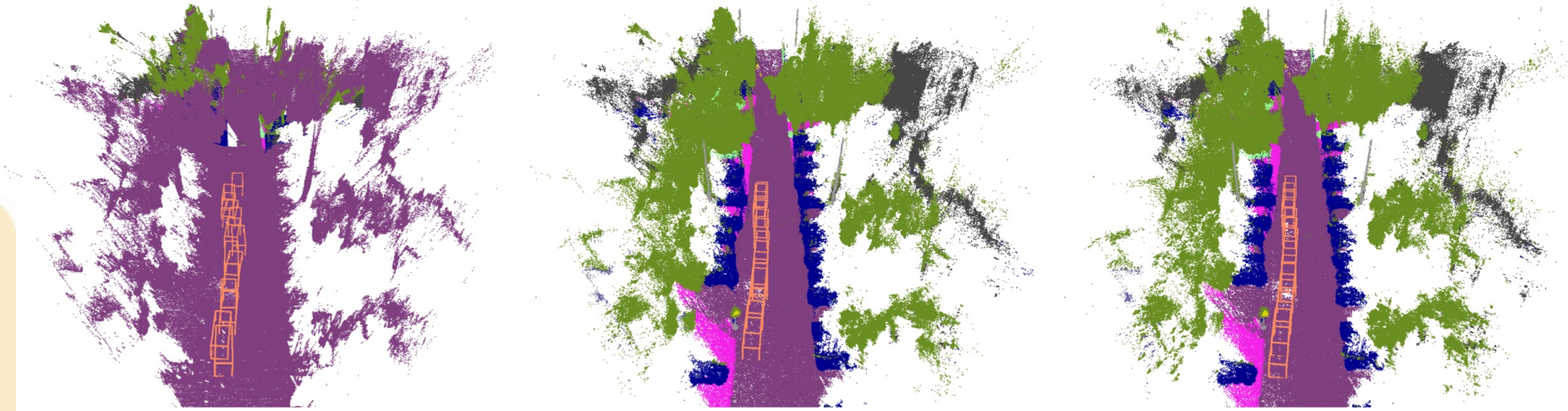
# HUGS: Method Overview




- Extend 3DGS to model **camera exposure**, **semantics**, **optical flow**
- Decomposing scenes into **static** regions and multiple rigidly **moving objects**
- Add **physical constraints** to dynamic object pose optimization via unicycle model

# Holistic Understanding: 3D Trajectory





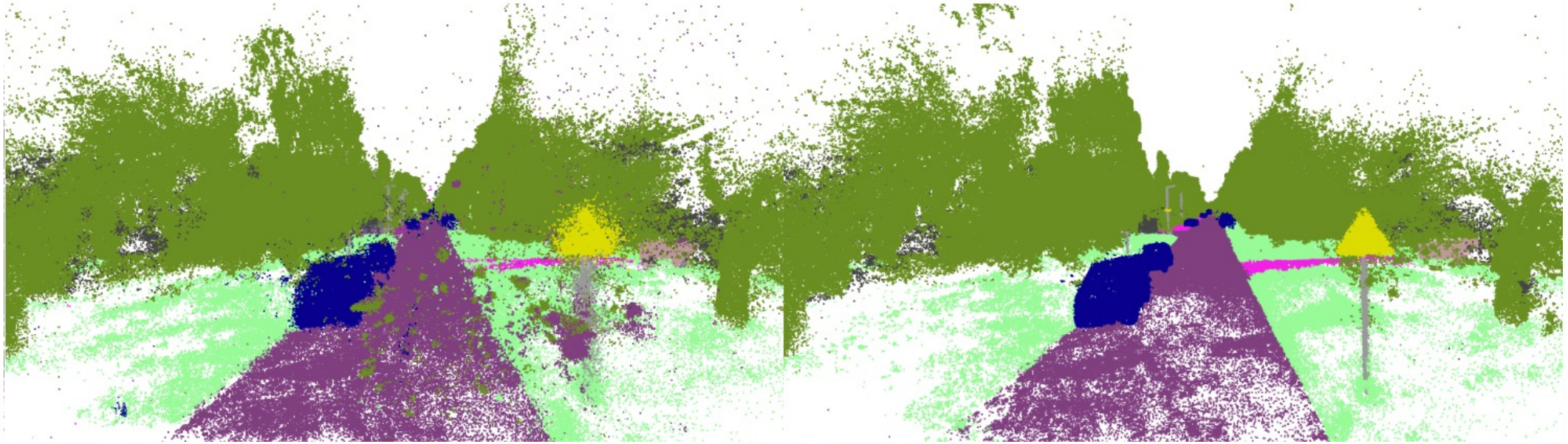
10 steps | 2000 steps | 5000 steps

w/o opt., w/o uni. | w/ opt., w/o uni. | Ours

➤ Optimize 3D bboxes independently w/o unicycle model leads to artifacts

Zhou, Shao, Xu, Bai, Qiu, Liu, Wang, Geiger, Liao. HUGS: Holistic Urban 3D Scene Understanding via Gaussian Splatting. CVPR 2024

# Holistic Understanding: 3D Semantic Reconstruction



Ours w/ $\mathbf{S}_{\text{2D\_norm}}$

Ours w/ $\mathbf{S}_{\text{3D\_norm}}$

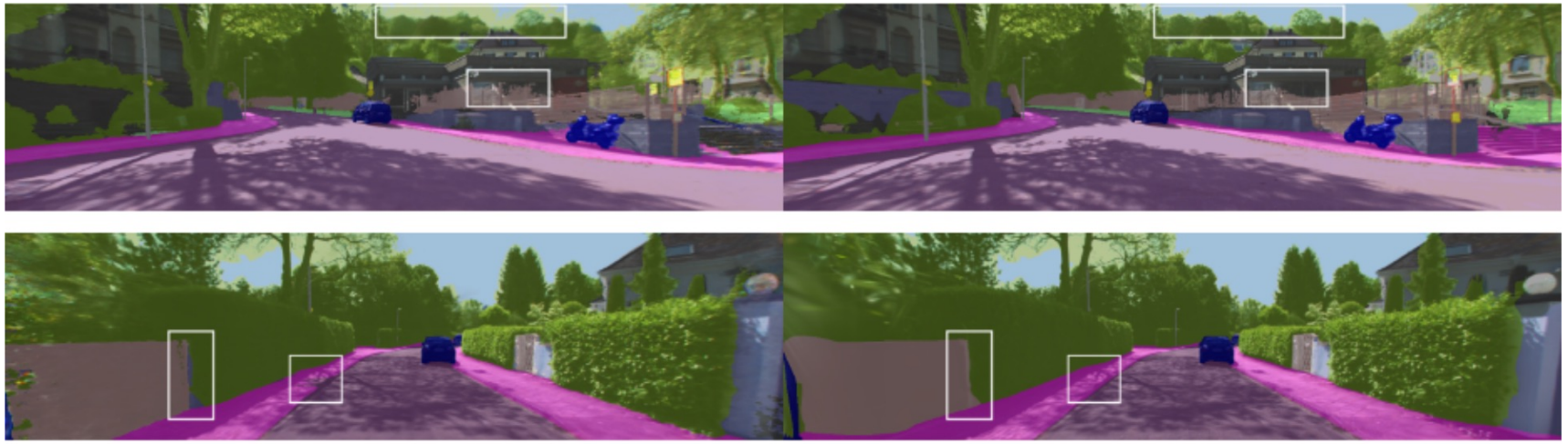
$$\mathbf{S}_{\text{2D\_norm}} = \text{softmax}\left(\sum_{i\in\mathcal{N}} \mathbf{s}_i \alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_j)\right)$$

$$\mathbf{S}_{\text{3D\_norm}} = \sum_{i\in\mathcal{N}} \text{softmax}(\mathbf{s}_i)\alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_j)$$

➤ Incorporating semantic improves geometry when **applying softmax in 3D**

Zhou, Shao, Xu, Bai, Qiu, Liu, Wang, Geiger, Liao. HUGS: Holistic Urban 3D Scene Understanding via Gaussian Splatting. CVPR 2024

# Holistic Understanding: 2D Semantic Reconstruction

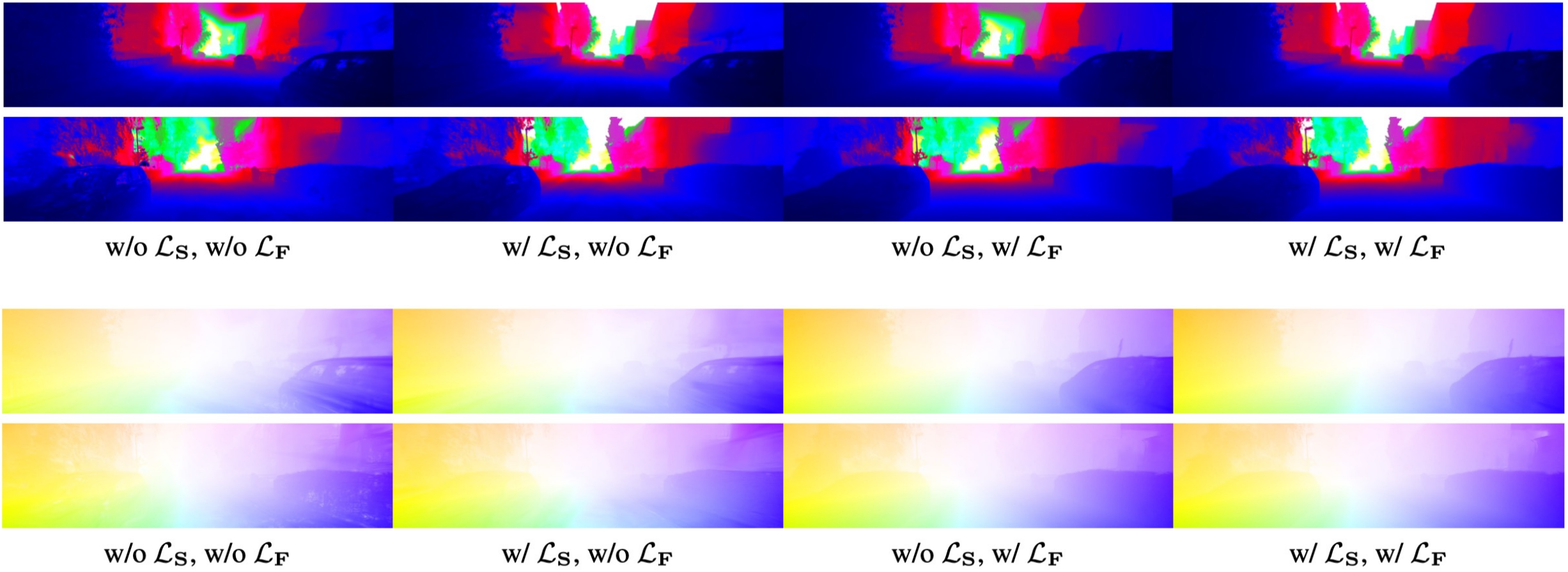


PNF Ours

| | PSNR↑ | SSIM↑ | LPIPS↓ | $mIoU_{cls}$ ↑ | $mIoU_{cat}$ ↑ |
|---|---|---|---|---|---|
| mip-NeRF [2] | 21.54 | 0.778 | 0.365 | 48.25 | 67.47 |
| PNF [19] | 22.07 | 0.820 | 0.221 | **73.06** | 84.97 |
| MARS [40] | 23.09 | 0.857 | 0.174 | - | - |
| Ours | **23.38** | **0.870** | **0.121** | 72.65 | **85.64** |

- Allow for rendering high-quality 2D semantic labels

# Holistic Understanding: Optical Flow




➤ Flow supervision further improves geometry, despite not enhancing appearance

# Viewpoint Extrapolation

- Ranking #1 on KITTI-360 Novel View Synthesis Leaderboard

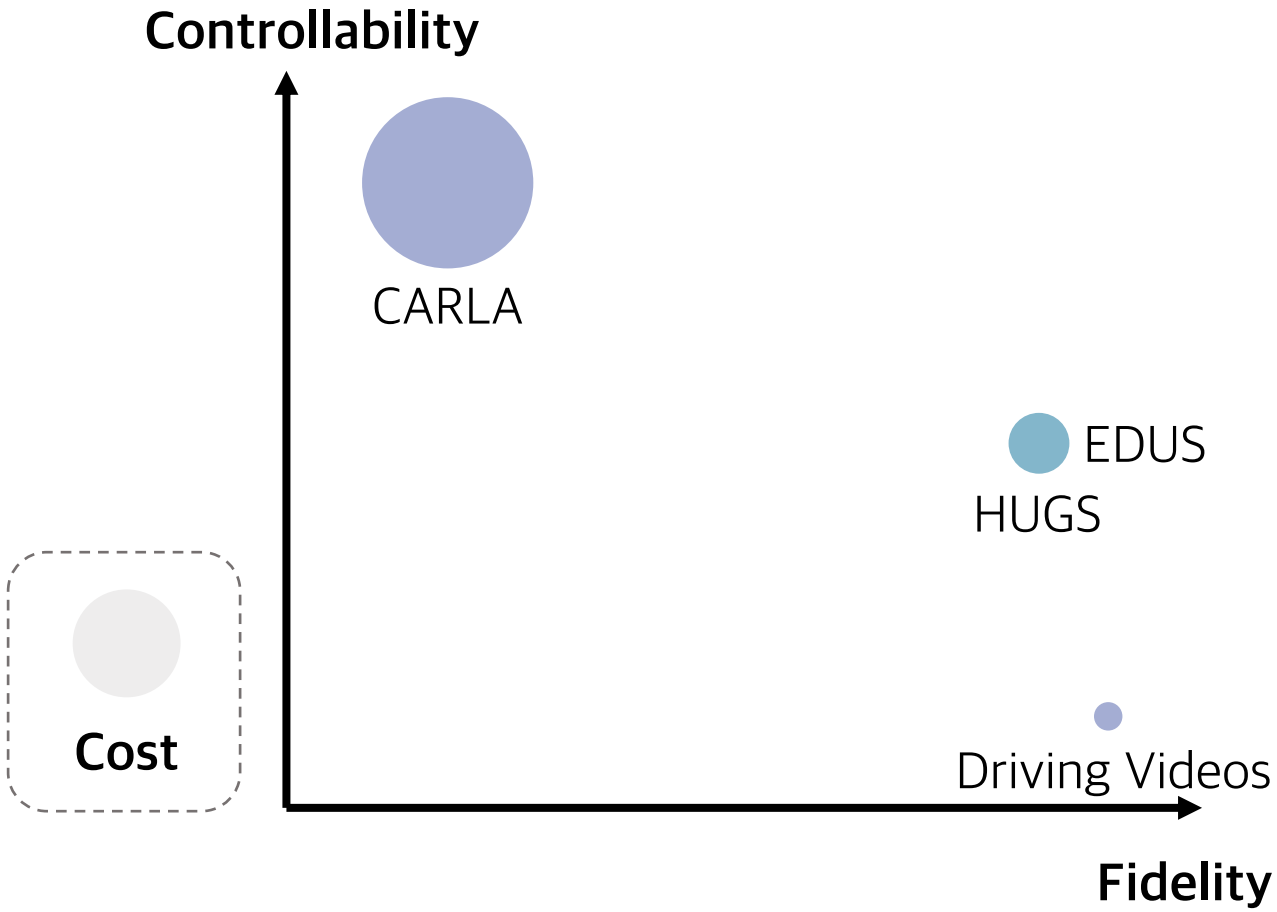
# Scene Editing

# Generalization on Other Datasets

Waymo

Waymo

# Towards Autonomous Driving Simulator




**Reconstruction:**

- \+ control over viewpoints
- \+ control over dynamic objects
- \+ image fidelity
- \+ no ground truth 3D annotations
- − relatively long training time
- − struggle with sparse views

# EDUS: Efficient Depth-Guided Urban View Synthesis

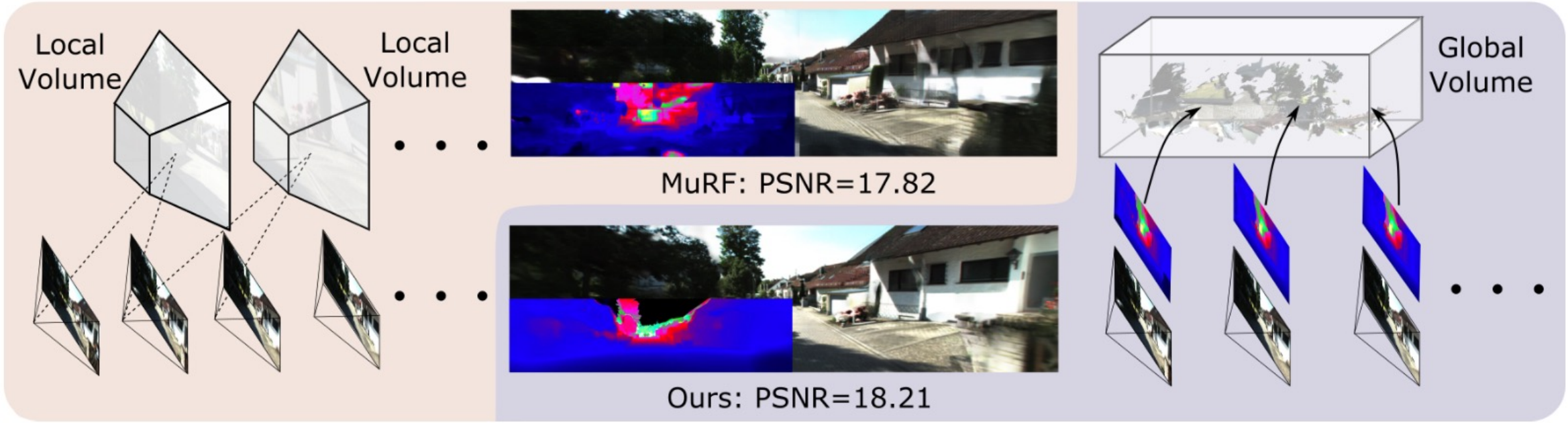
- Input: Posed RGB images of static scenes
- Enable efficient urban reconstruction in **2 seconds** via **feed-forward inference**

# EDUS: Efficient Depth-Guided Urban View Synthesis

- Input: Posed RGB images of static scenes
- Enable efficient urban reconstruction in **2 seconds** via **feed-forward inference**
- Per-scene finetuning converges in **5 minutes**

Miao*, Huang*, Bai, Qiu, Liu, Geiger, Liao. Efficient Depth-Guided Urban View Synthesis. 2024

# EDUS: Key Idea




- **Key idea**: Use **depth priors** for generalizable urban scene reconstruction
- Existing generalizable NeRF approaches learn **local volume**
- Learning in **global volume**, avoid overfitting to specific camera settings

# Comparison with Generalizable Methods

| Methods | Setting | KITTI360$_{drop50}$ | | | KITTI360$_{drop80}$ | | | Waymo$_{drop50}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| IBR-Net | No per-scene opt. | 19.99 | 0.624 | 0.217 | 15.96 | 0.469 | 0.354 | 21.28 | 0.777 | 0.199 |
| MVSNeRF | | 17.73 | 0.618 | 0.328 | 16.50 | 0.577 | 0.365 | 19.58 | 0.662 | 0.278 |
| Neo360 | | 13.73 | 0.394 | 0.624 | 12.98 | 0.357 | 0.659 | 14.07 | 0.541 | 0.708 |
| MuRF | | **22.19** | 0.741 | 0.264 | 18.69 | 0.639 | 0.353 | 23.12 | 0.779 | 0.318 |
| **Ours** | | 21.93 | **0.745** | **0.178** | **19.63** | **0.668** | **0.244** | **23.16** | **0.761** | **0.189** |
| IBR-Net | Per-scene opt. | 21.17 | 0.657 | 0.199 | 17.98 | 0.529 | 0.279 | 23.39 | 0.825 | 0.163 |
| MVSNeRF | | 19.47 | 0.647 | 0.310 | 18.06 | 0.602 | 0.353 | 24.28 | 0.759 | 0.207 |
| Neo360 | | 17.92 | 0.489 | 0.566 | 17.51 | 0.445 | 0.581 | 22.59 | 0.670 | 0.522 |
| MuRF | | 23.71 | 0.762 | 0.233 | 19.70 | 0.666 | 0.321 | 28.30 | 0.846 | 0.175 |
| **Ours** | | **24.43** | **0.793** | **0.136** | **20.91** | **0.712** | **0.220** | **28.45** | **0.834** | **0.132** |

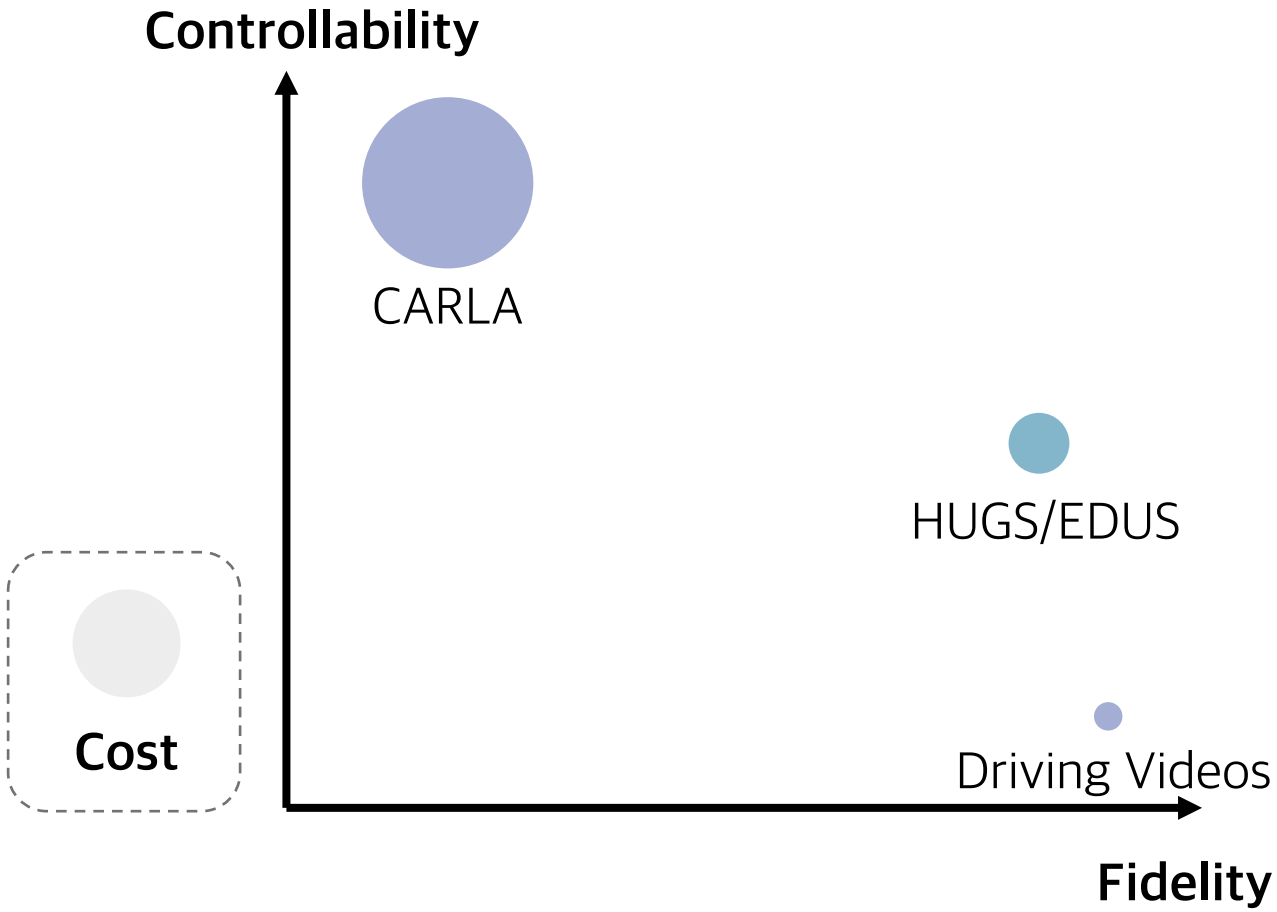
- Applicable to **various sparsity levels**
- Generalizes well to **Waymo** when trained on **KITTI-360**

Comparison with other generalizable baselines in feed-forwad inference (Drop80 setting)

# Zero-Shot Generalization

➤ **Zero-shot generalization** on **Waymo** using model trained on **KITTI-360**

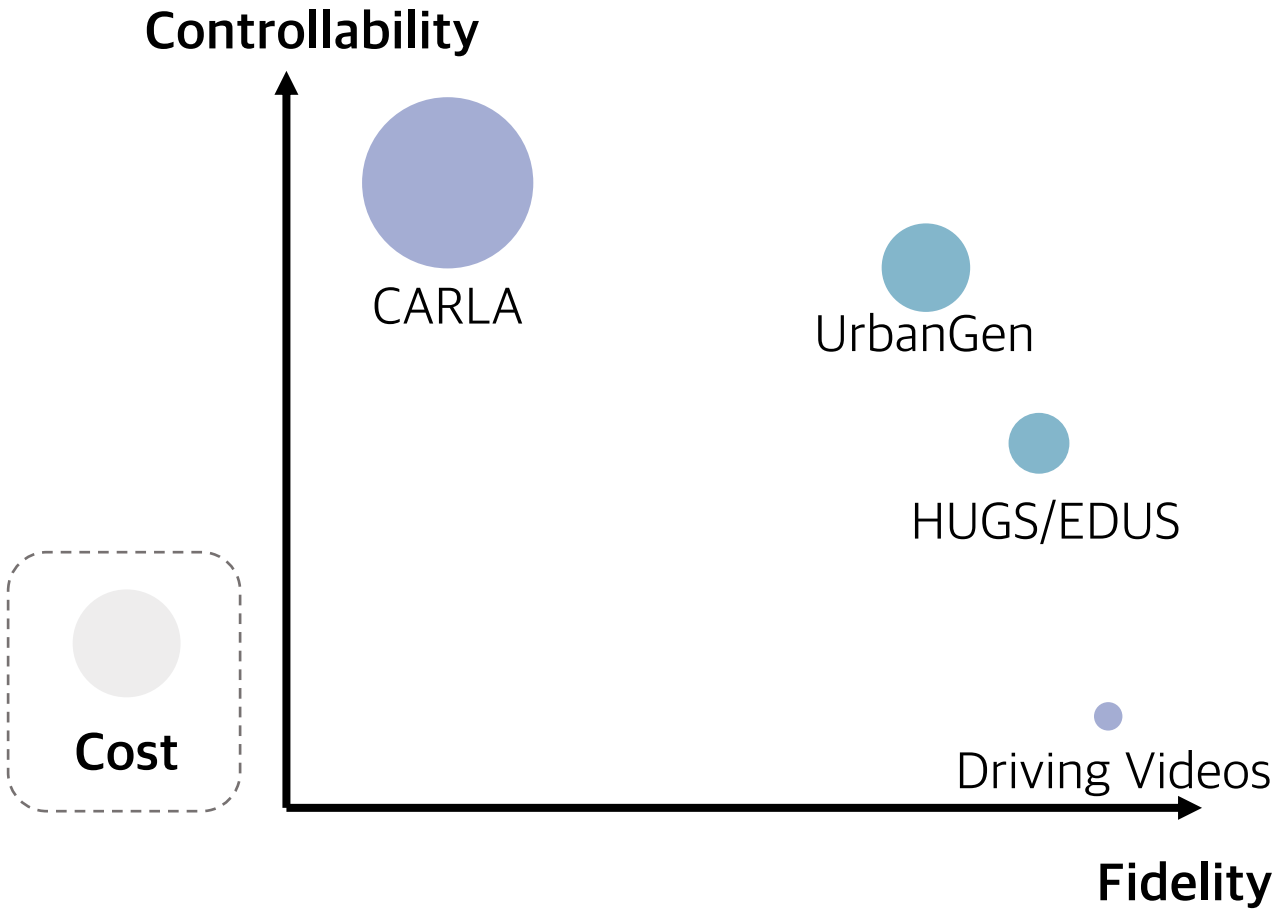
# Towards Autonomous Driving Simulator




**Reconstruction:**

- \+ control over viewpoints
- \+ control over dynamic objects
- \+ image fidelity
- \+ efficient reconstruction
- \+ handle sparse views
- − control over full scene

# Towards Autonomous Driving Simulator




**Reconstruction -> Generation**

- control over viewpoints
- control over dynamic objects
- image fidelity
- efficient reconstruction
- handle sparse views
- control over full scene

# Scene-Level 3D Generative Model

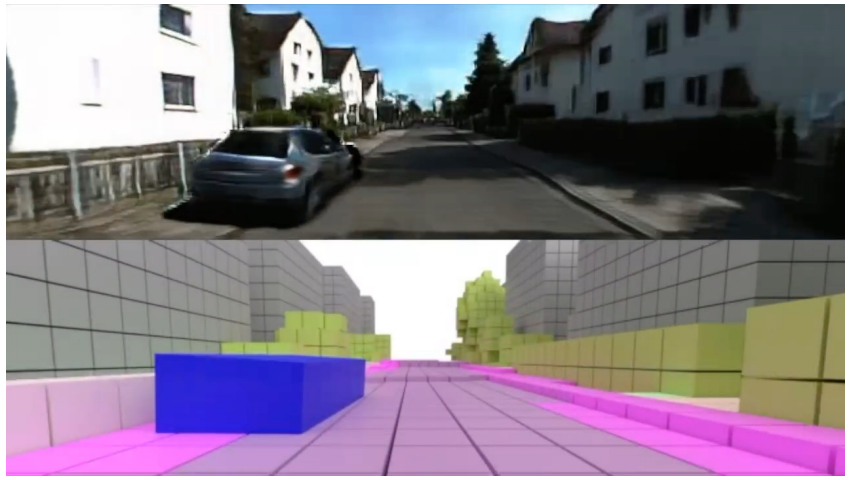
GIRAFFE

DiscoScene

+ **Compositional** modeling, canonical space for **foreground** objects based on 3D Bbox
– **Static background**, lack of control over **background** regions

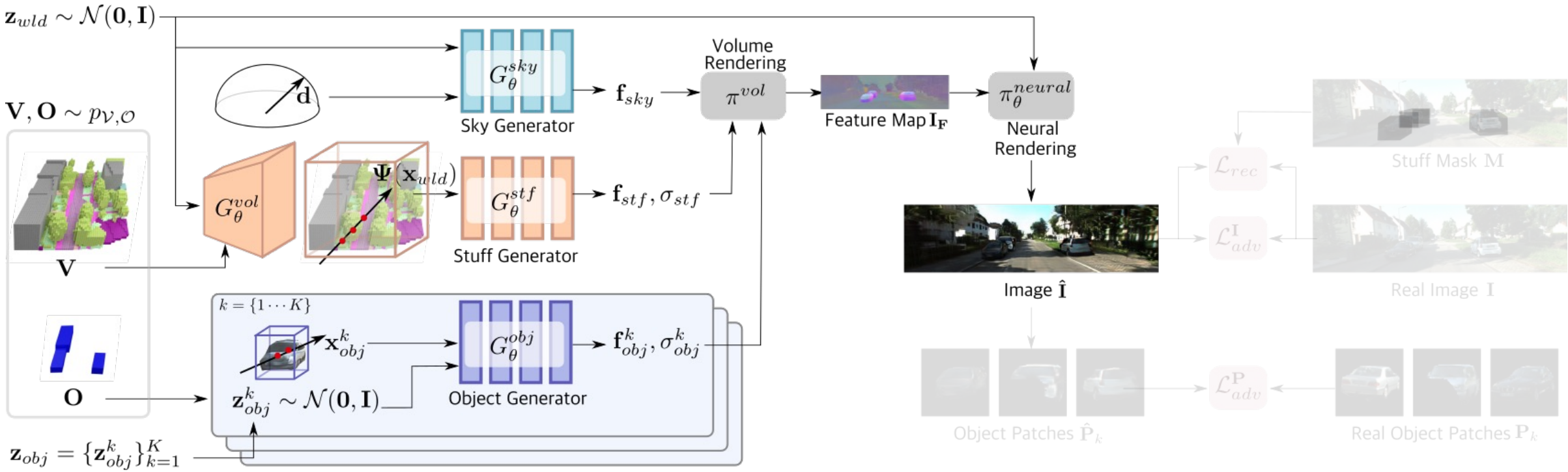
Niemeyer, Geiger: GIRAFFE Representing Scenes as Compositional Generative Neural Feature Fields. CVPR 2021
Xu, Chai, Shi, et al.: DiscoScene: Spatially Disentangled Generative Radiance Field for Controllable 3D-aware Scene Synthesis, CVPR 2023

# UrbanGIRAFFE

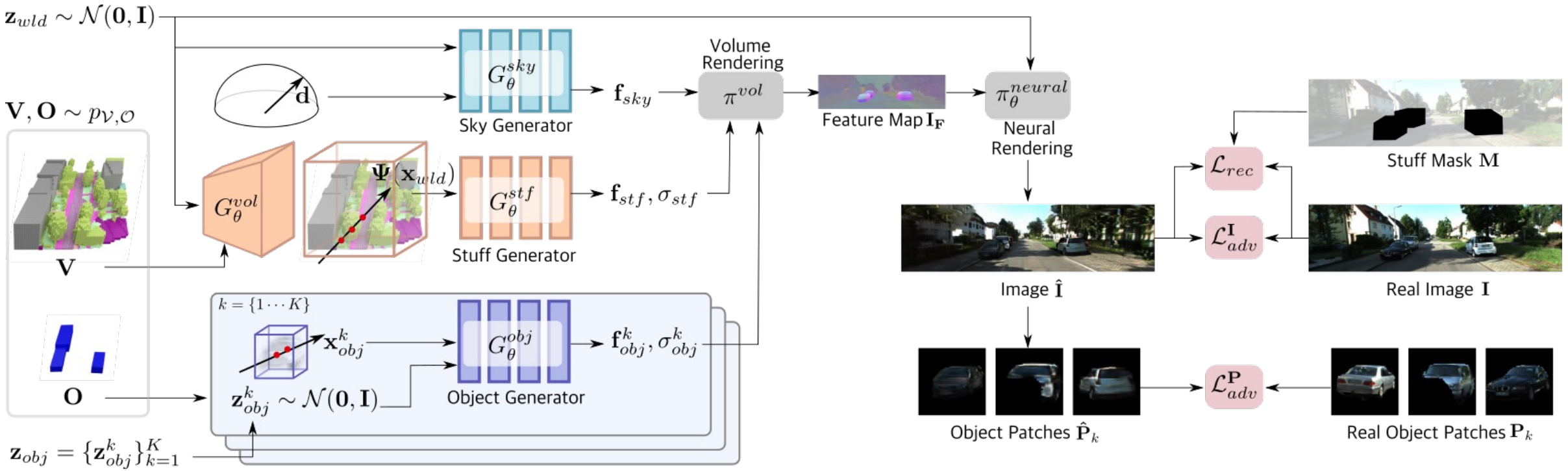
- **Goal**: **Compositional** and **controllable** synthesis of foreground and background
- **Key Idea**: introduce **panoptic prior** for coarse geometry and semantic guidance

# UrbanGIRAFFE



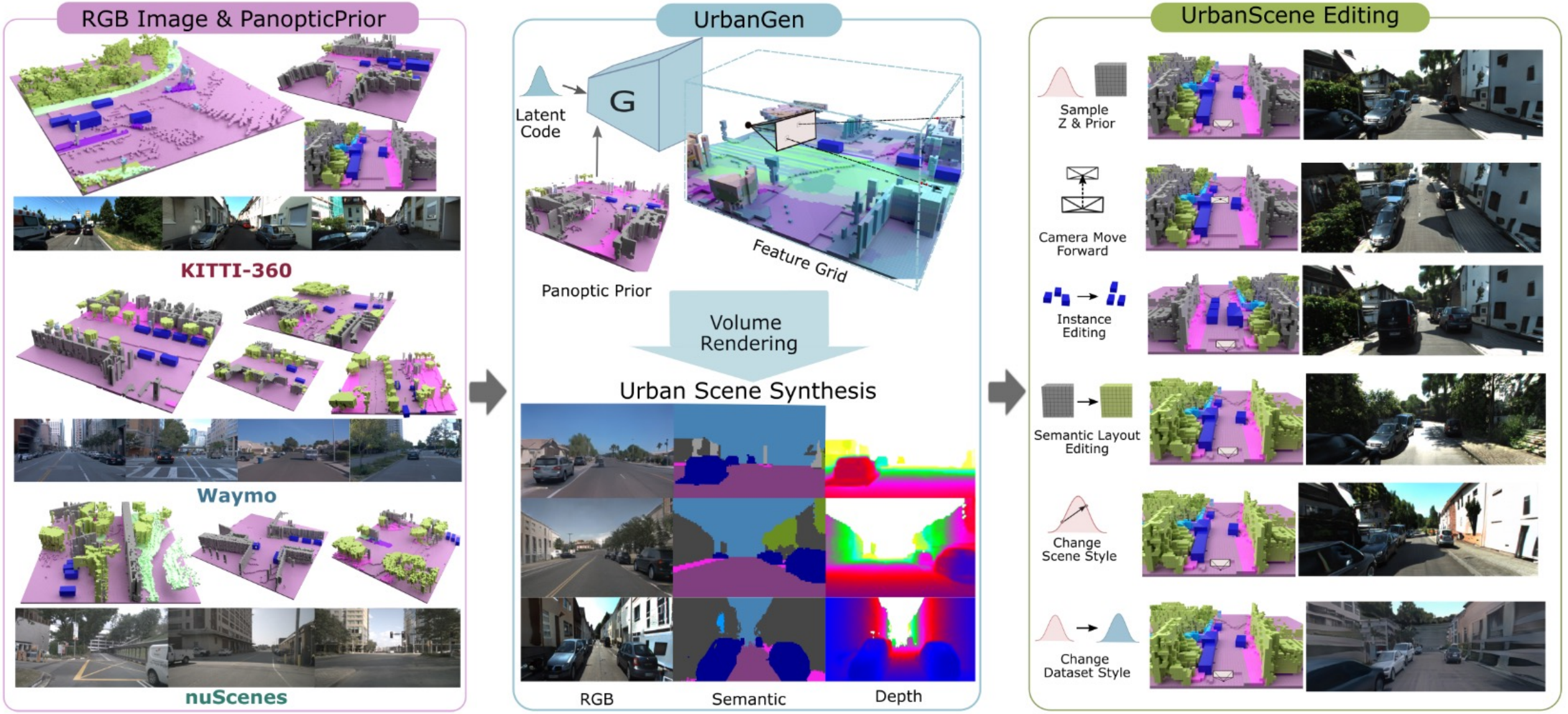

- **Panoptic Prior**: Semantic volume $\mathbf{V}$ and object layouts $\mathbf{O}$
- Semantic voxel-conditioned **stuff** generator, **object** generator, **sky** generator

# UrbanGIRAFFE




- **Adversarial loss** for full image and object patches
- **Reconstruction loss** for stuff regions

# UrbanGen




Yang, Yang, Guo, Xiong, Wang, Geiger, Liao: UrbanGen: Urban Generation with Compositional and Controllable Neural Fields, 2024

# Scene Generation

KITTI-360 Waymo nuScenes

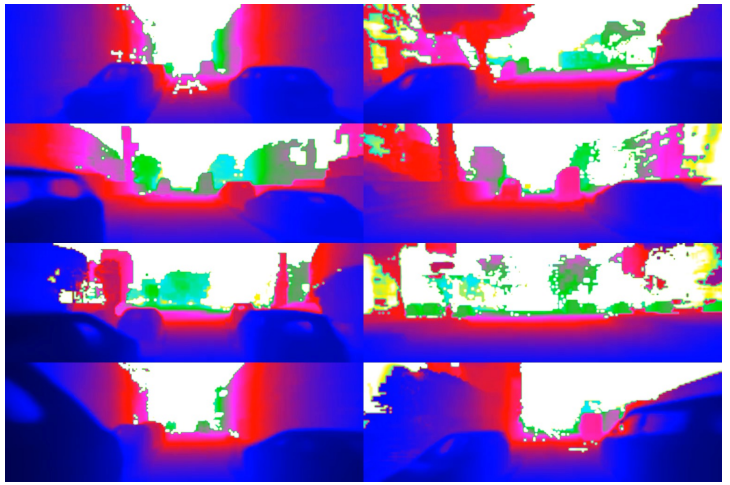
Yang, Yang, Guo, Xiong, Wang, Geiger, Liao: UrbanGen: Urban Generation with Compositional and Controllable Neural Fields, 2024

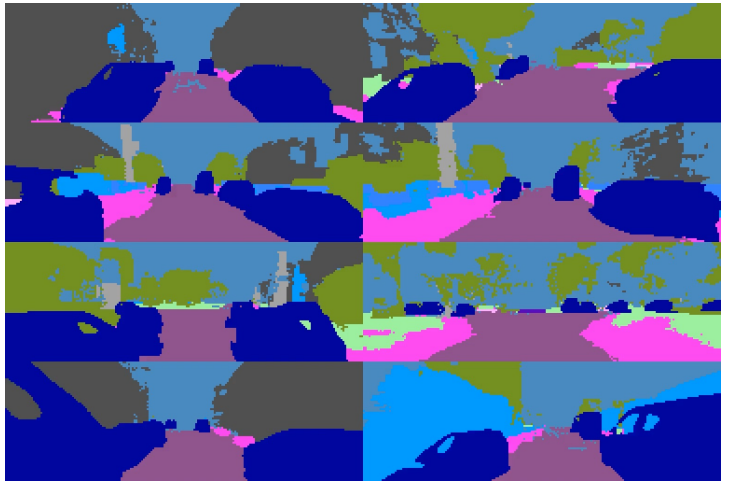
# Style Interpolation

KITTI-360 Waymo nuScenes

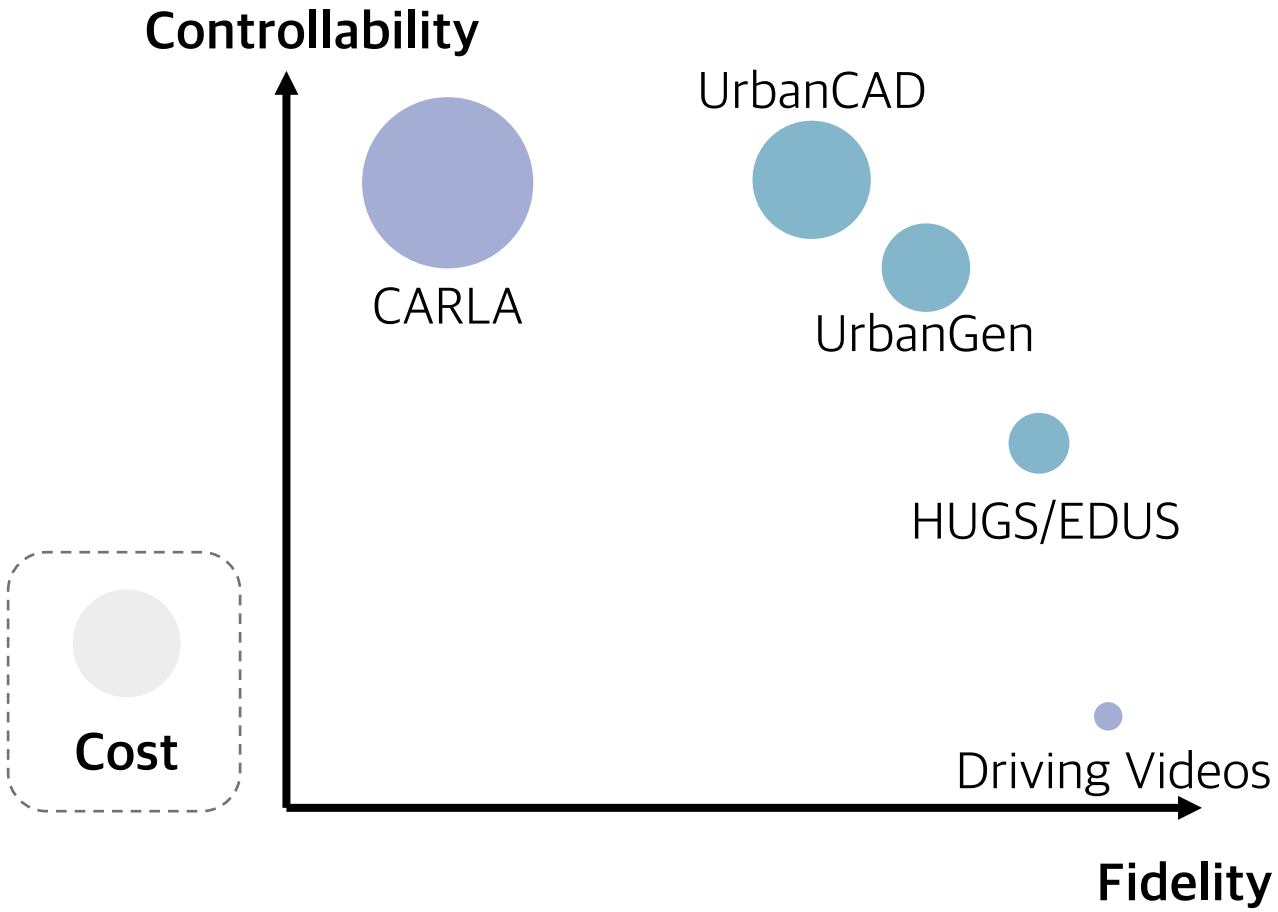
Yang, Yang, Guo, Xiong, Wang, Geiger, Liao: UrbanGen: Urban Generation with Compositional and Controllable Neural Fields, 2024

# Object Editing





# Towards Autonomous Driving Simulator

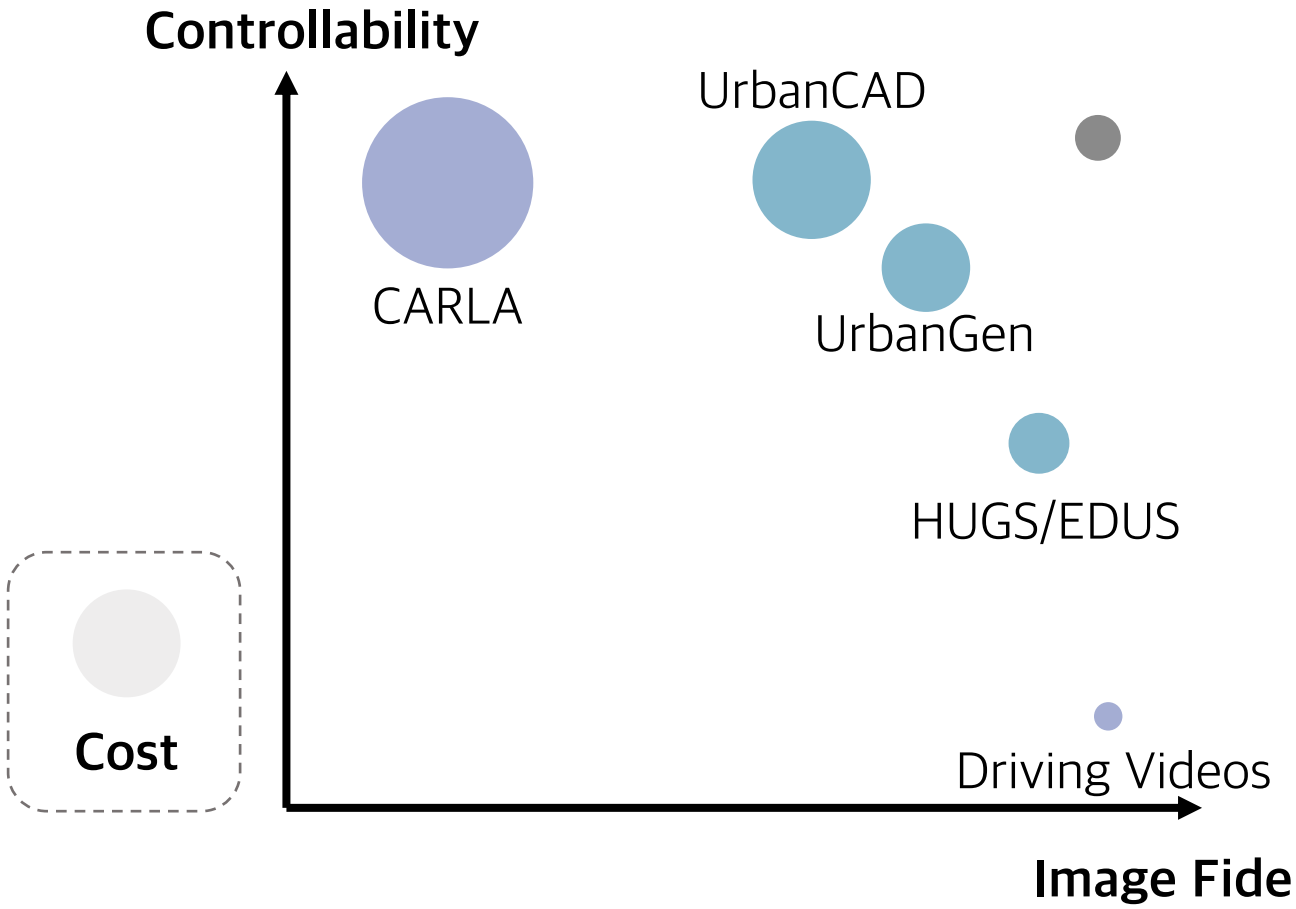



**How do we simulate corner cases?**

# UrbanCAD: Towards Fully Controllable and Photorealistic 3D Vehicles from a Single Urban Image

Taking single view urban image as input

# Towards Autonomous Driving Simulator

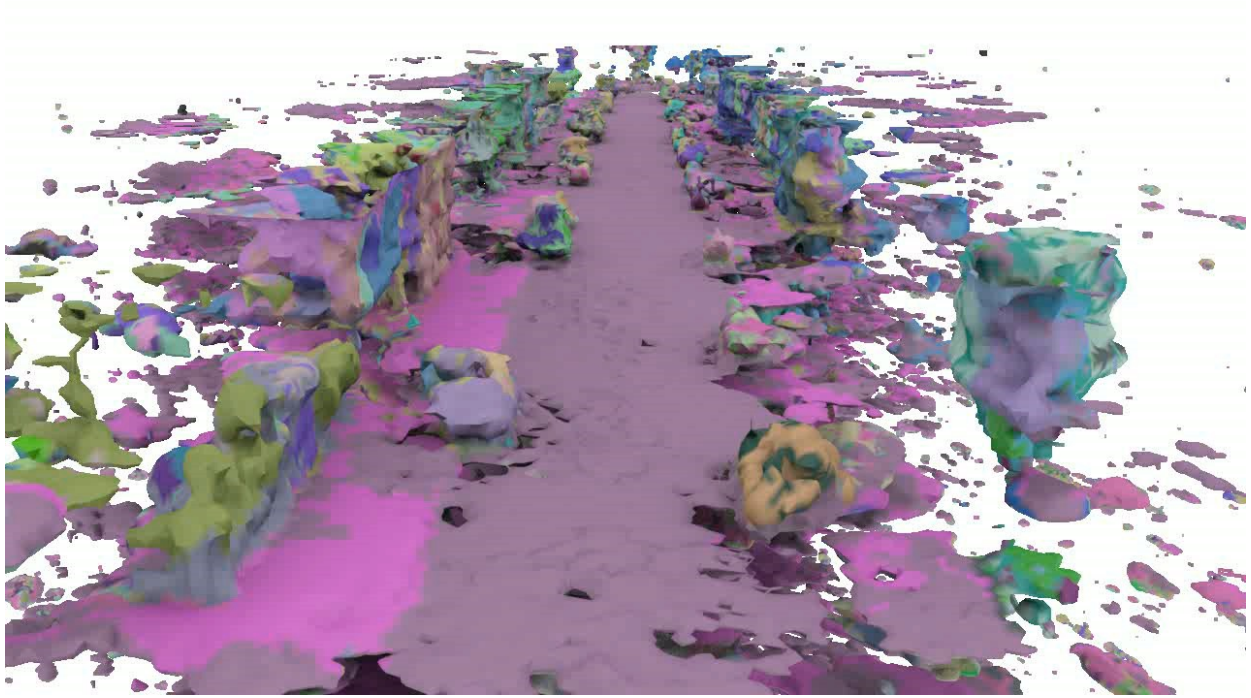



**Goal:**

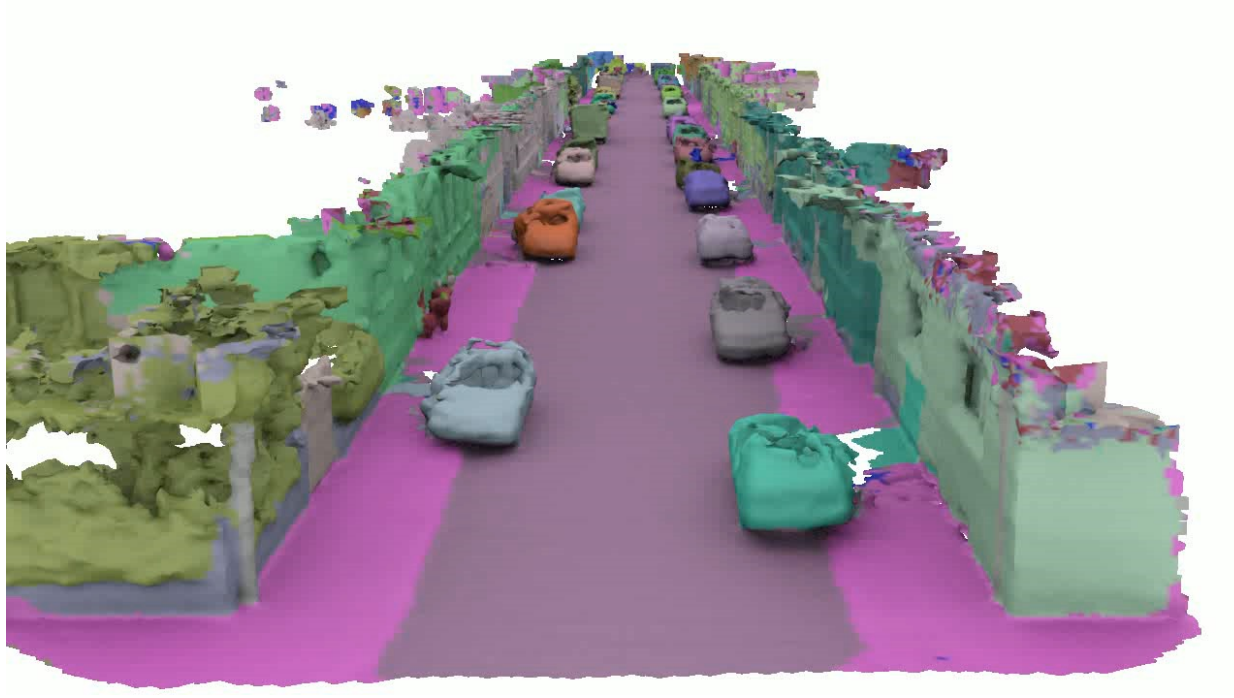
+ high controllability
+ high image fidelity
+ low cost

**The journey continues; more to uncover!**

# Towards Reducing Labeling Cost

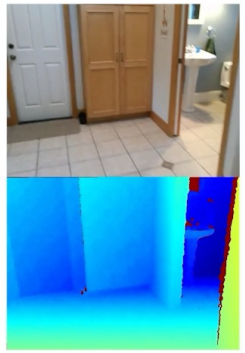


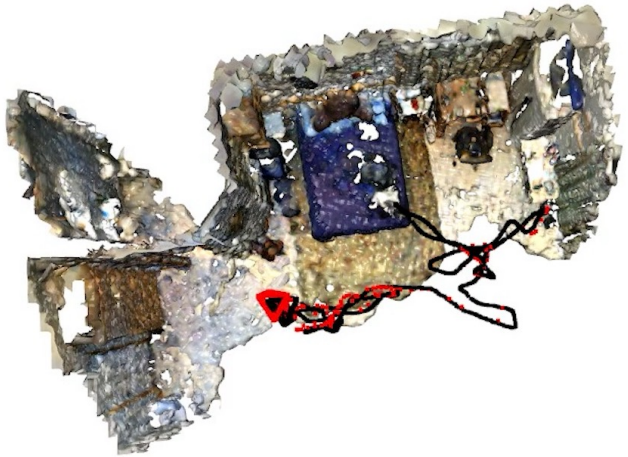
PVLFF [RA-Letter]

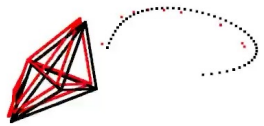


PanopticRecon (Ours)

# Towards Faster Reconstruction







NGEL-SLAM (Ours)

Mao, Yu, Wang, Wang, Xiong, Liao. NGEL-SLAM: Neural Implicit Representation-based Global Consistent Low-Latency SLAM System. **ICRA 2024 (Best Robot Vision Paper)**

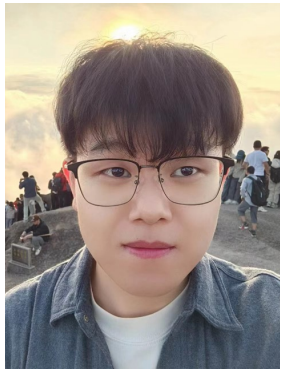
# Towards Higher Fidelity


SORA: https://openai.com/index/sora

# Collaborators

Hongyu Zhou

Sheng Miao



Jiaxin Huang

Yuanbo Yang

Yichong Lu

Yue Wang

Andreas Geiger

Thank you!

yiyiliao.github.io/