3D Controllable Image Synthesis

Yiyi Liao

December 24, 2020

University of Tübingen
MPI for Intelligent Systems
Autonomous Vision Group

Collaborators

Katja Schwarz

Lars Mescheder

Michael Niemeyer

Jun Xie

Andreas Geiger


CARLA
Introducing Town10
A vibrant city with new assets to improve visual quality

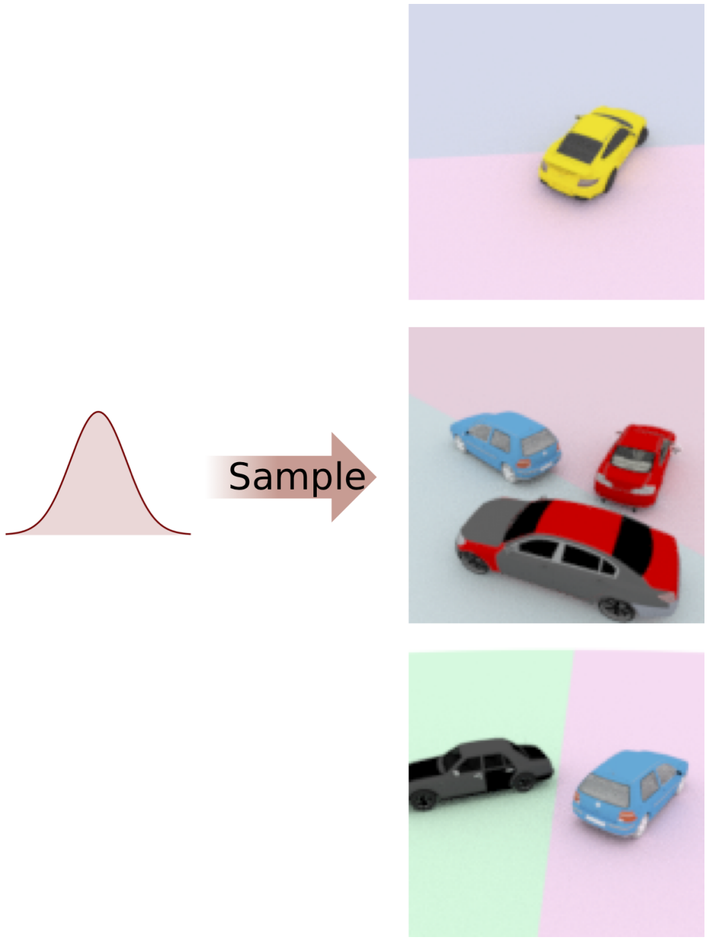
Can we learn a simulation tool from 2D images?

Liao*, Schwarz*, Mescheder, Geiger: Towards Unsupervised Learning of Generative Models for 3D Controllable Image Synthesis. CVPR, 2020

Goal

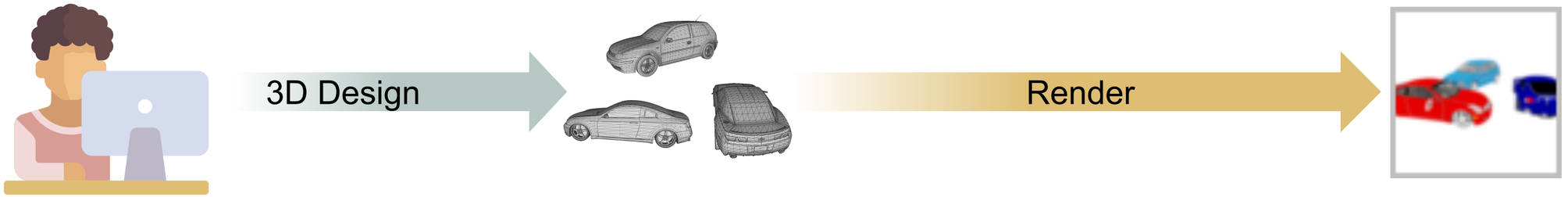
3D Controllable Image Synthesis

- Learning a **generative image model**
- With **controllable 3D** factors:
    - Object shape
    - Object appearance
    - Object pose
    - Camera viewpoint
- Learn from **2D observations** (unposed images)
- **No supervision** (segmentation, bounding box, depth)



Generative Models

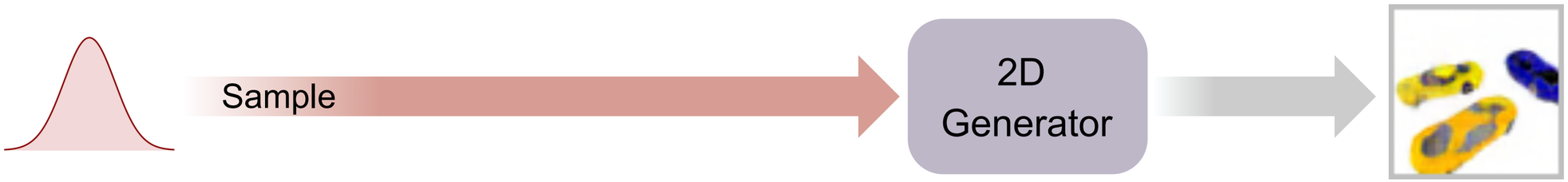
Classical Rendering Pipeline



- 3D factors can be controlled
- Expensive 3D content creation

Generative Models

2D Generative Models

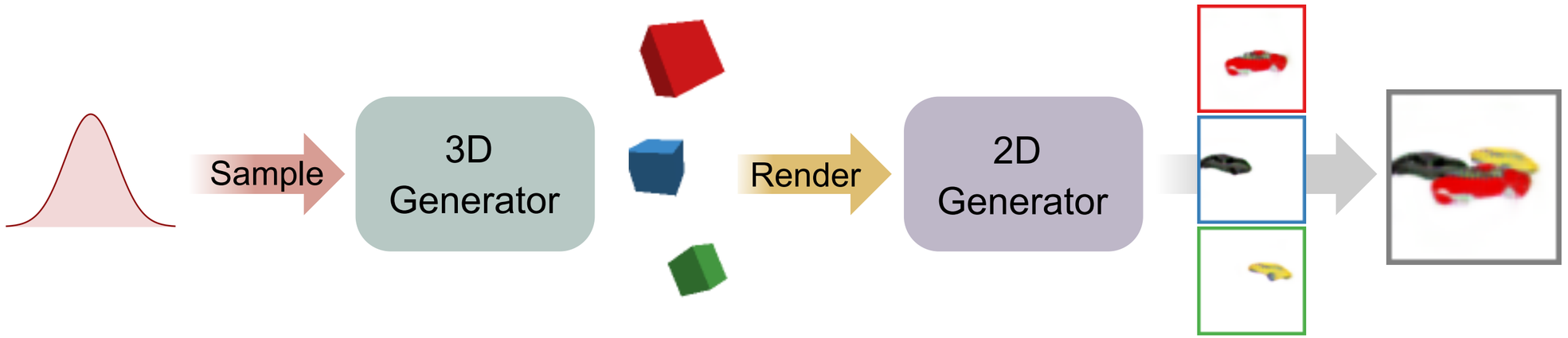


+ Unsupervised learning from 2D images

− 3D factors cannot be controlled

Generative Models

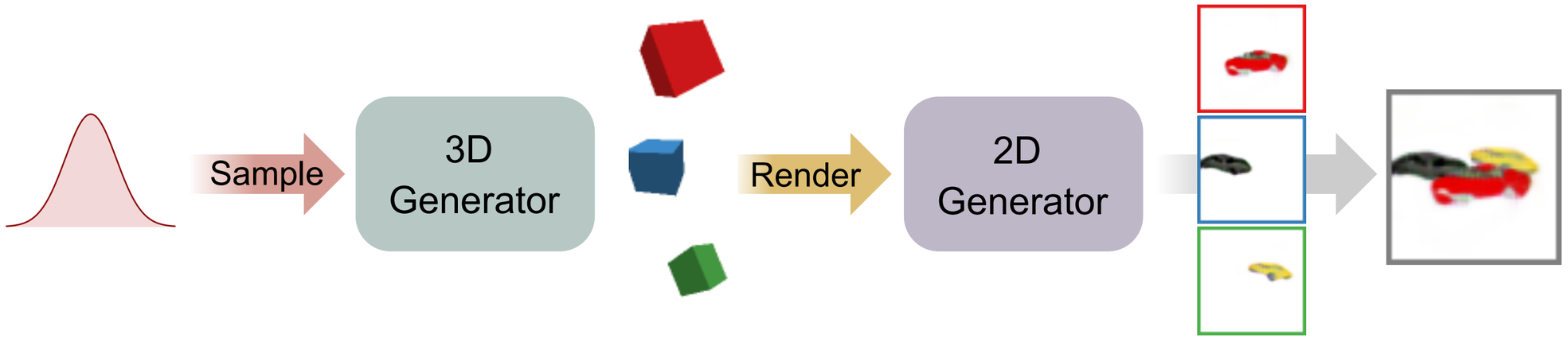
Our approach



- 3D factors can be controlled
- Unsupervised learning from 2D images

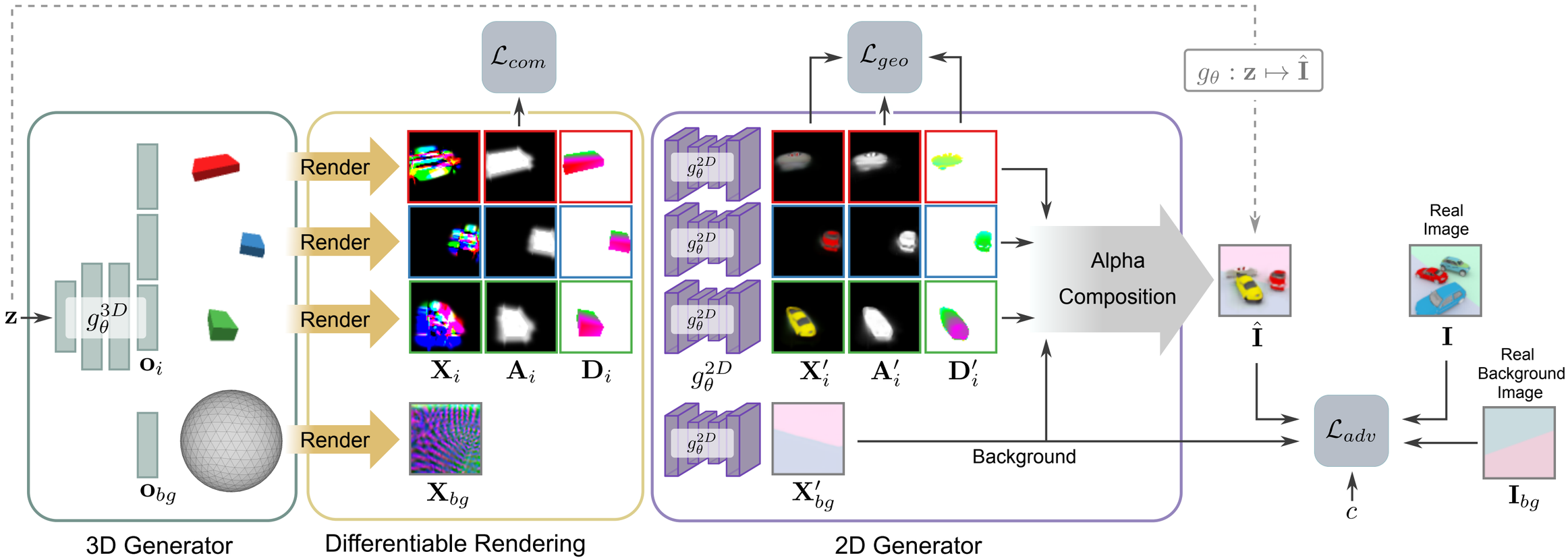
Generative Models

Our approach

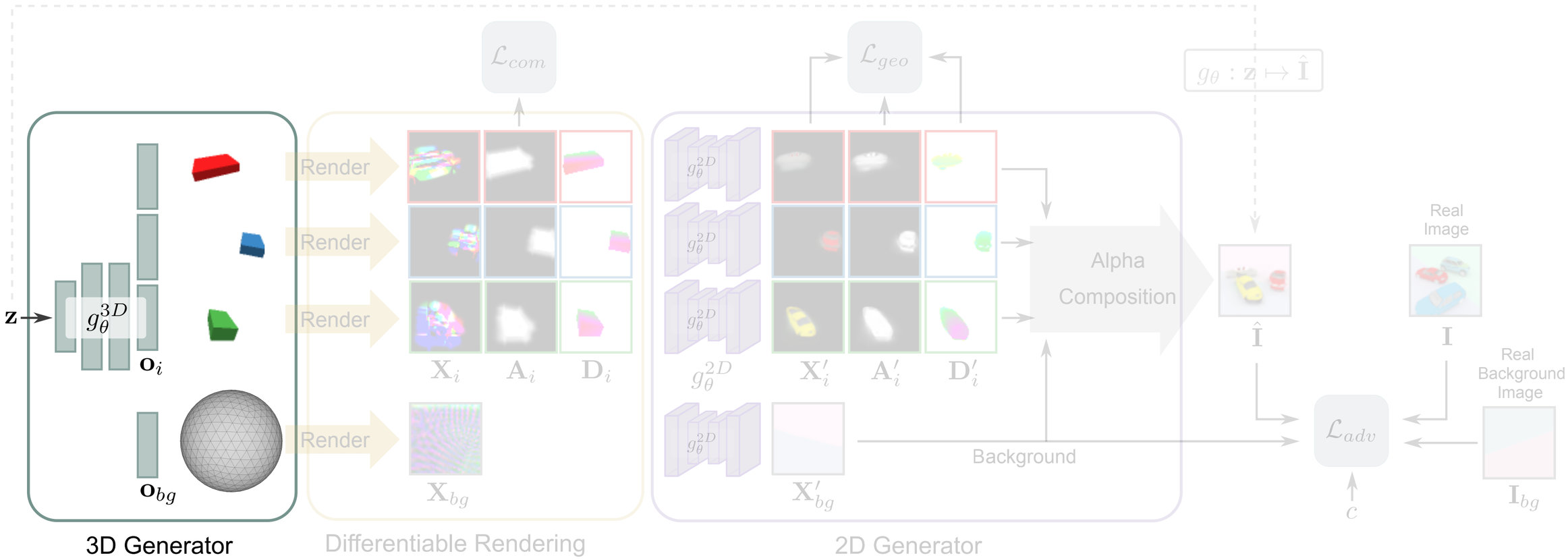


Idea: Learning the image generation process jointly in 3D and 2D space

Method

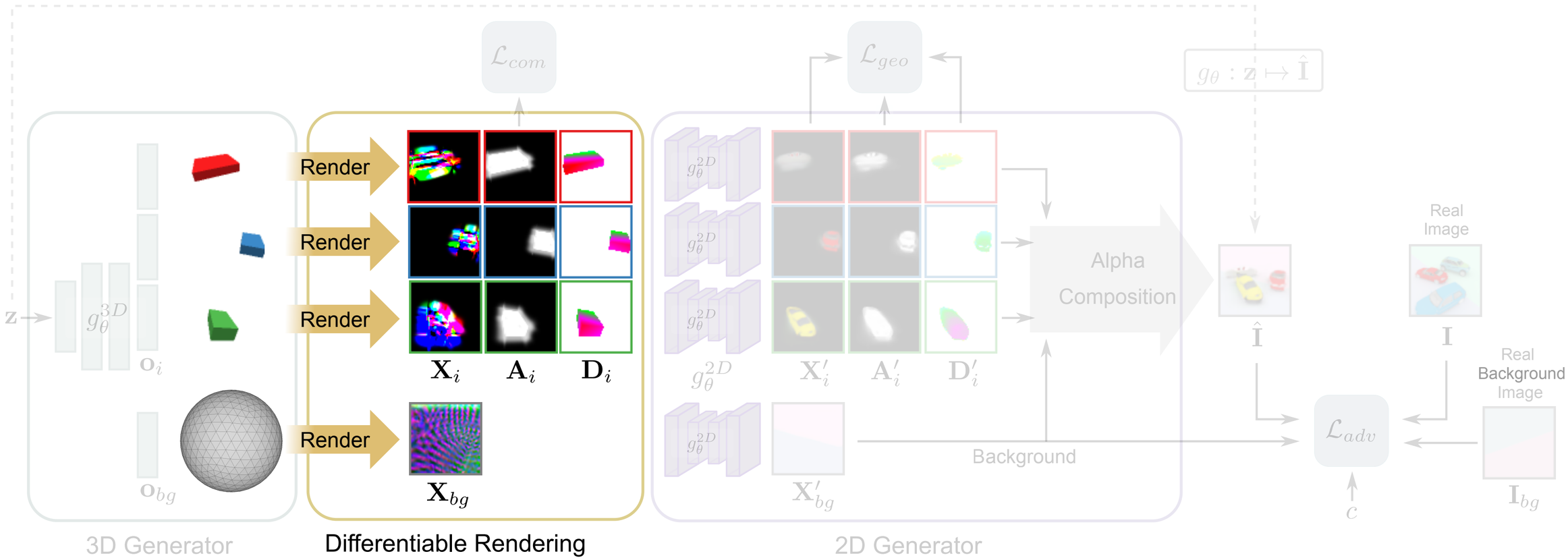

$\mathcal{L}_{com}$
$\mathcal{L}_{geo}$
$g_\theta : \mathbf{z} \mapsto \hat{\mathbf{I}}$
$\mathbf{z}$
g_θ^{3D}
$\mathbf{o}_i$
$\mathbf{o}_{bg}$
Render
Render
Render
Render
$\mathbf{X}_i$
$\mathbf{A}_i$
$\mathbf{D}_i$
$\mathbf{X}_{bg}$
g_θ^{2D}
$\mathbf{X}'_i$
$\mathbf{A}'_i$
$\mathbf{D}'_i$
$\mathbf{X}'_{bg}$
Alpha Composition
Background
$\hat{\mathbf{I}}$
Real Image
$\mathbf{I}$
Real Background Image
$\mathbf{I}_{bg}$
$\mathcal{L}_{adv}$
c
3D Generator
Differentiable Rendering
2D Generator

Method



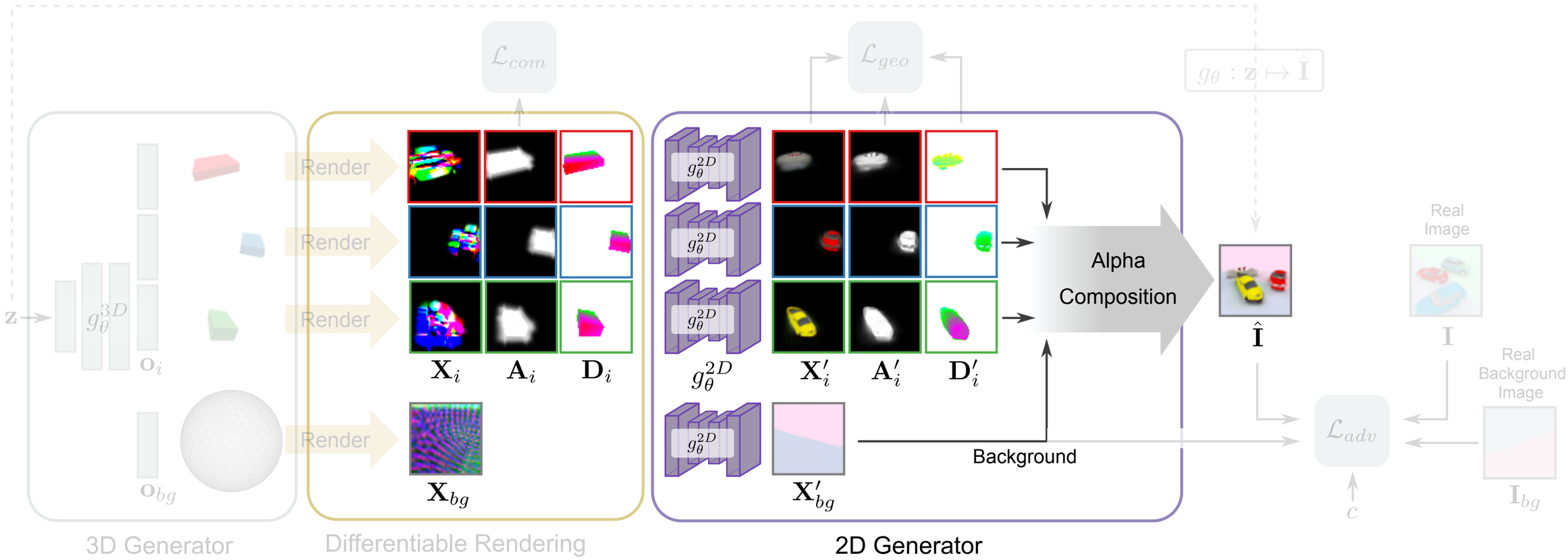
- Foreground/background primitives: $\{\mathbf{o}_1, \ldots, \mathbf{o}_N, \mathbf{o}_{bg}\}$, $\mathbf{o}_i = (\mathbf{R}_i, \mathbf{t}_i, \mathbf{s}_i, \phi_i)$
- Primitive type: point cloud, sphere, cuboid

Method



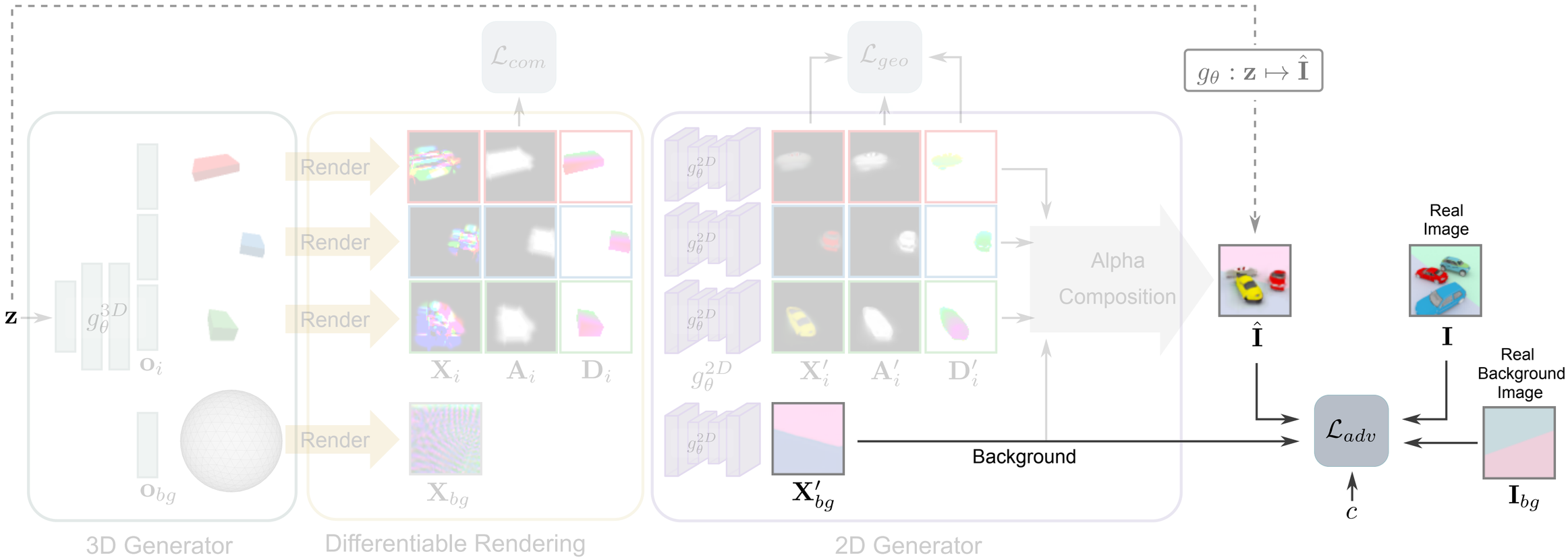
- Sample camera viewpoint, render each primitive individually
- Obtain feature map $\mathbf{X}$, alpha map $\mathbf{A}$ and depth map $\mathbf{D}$

Method



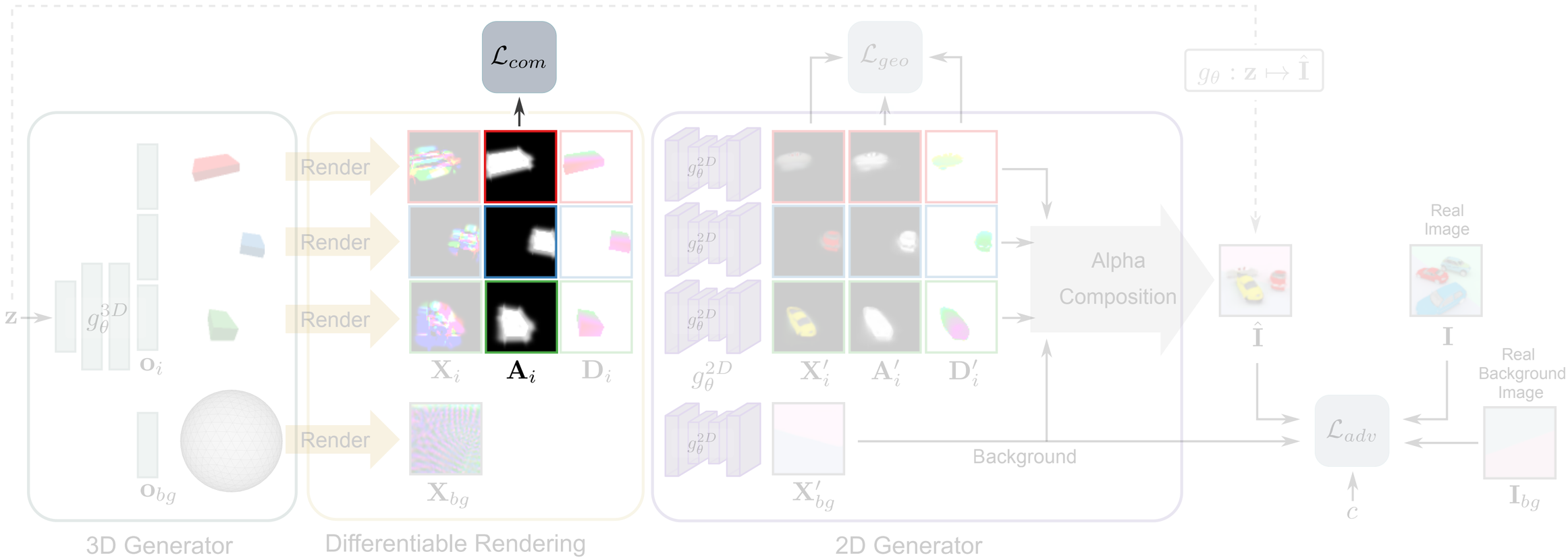
- Convert features to RGB pixel values
- Render to image via alpha composition (based on depth ordering)

Loss Functions



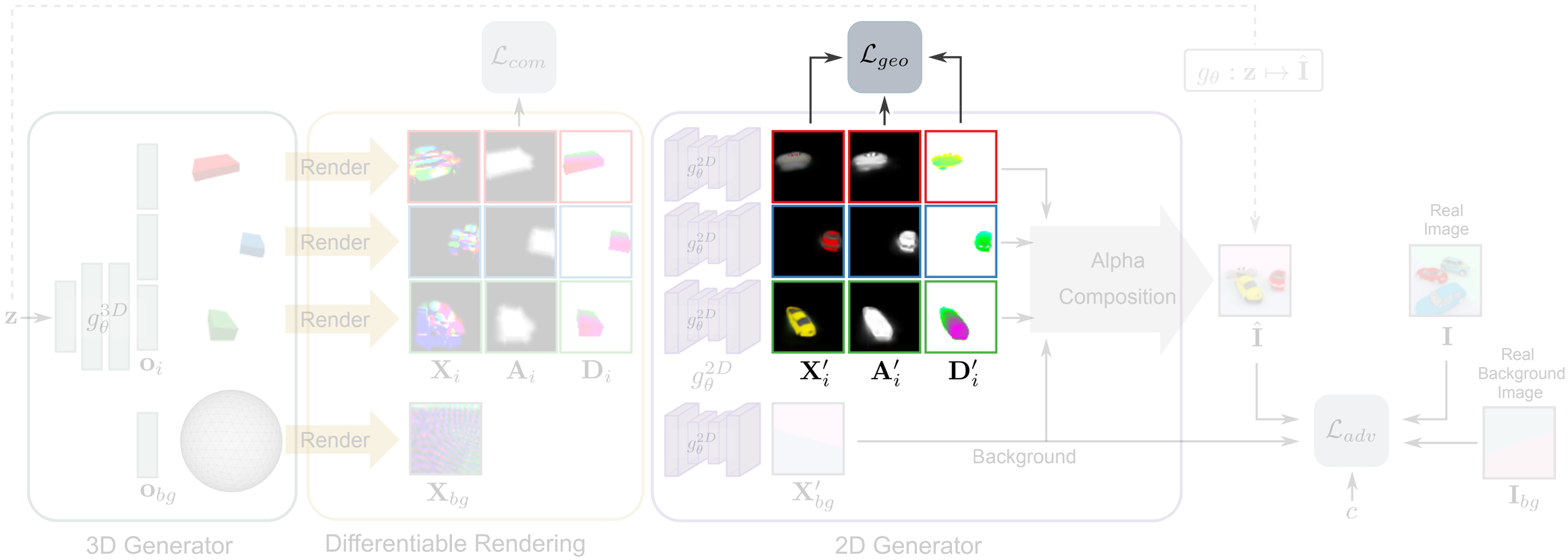
$$\mathcal{L}_{adversarial}(\theta, \psi, c) = \mathbb{E}_{p(\mathbf{z})}[f(d_\psi(g_\theta(\mathbf{z}, c), c))] + \mathbb{E}_{p_\mathcal{D}(\mathbf{I}|c)}[f(-d_\psi(\mathbf{I}, c))]$$

Loss Functions



$$\mathcal{L}_{compactness}(\theta) = \mathbb{E}_{p(\mathbf{z})}\left[\sum_{i=1}^{N} \max\left(\tau, \frac{\|\mathbf{A}_i\|_1}{H \times W}\right)\right]$$

Loss Functions



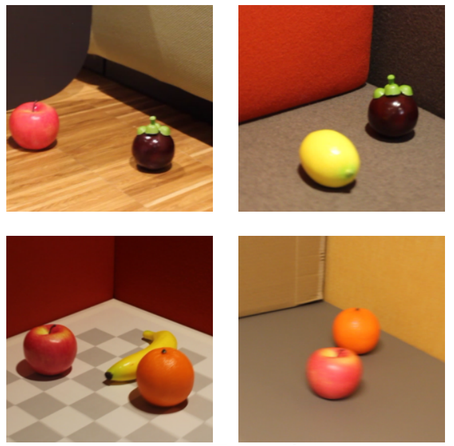
$$\mathcal{L}_{geometric}(\theta) = \mathbb{E}_{p(\mathbf{z})}\left[\sum_{i=1}^{N} \|\mathbf{A}'_i \odot (\mathbf{X}'_i - \tilde{\mathbf{X}}'_i)\|_1\right] + \mathbb{E}_{p(\mathbf{z})}\left[\sum_{i=1}^{N} \|\mathbf{A}'_i \odot (\mathbf{D}'_i - \tilde{\mathbf{D}}'_i)\|_1\right]$$

Datasets

Car w/o BG

Car with BG

Indoor



Fruit

Experiments

Baselines

Layout2Im

[Zhao et al., CVPR 2019]

- Only 2D translation control, requires 2D bounding box supervision
- Fails to disentangle object identity and pose

Experiments

Baselines

Ours 2D

Replace 3D primitives with 2D primitives

Experiments

Baselines

Ours 2D

Replace 3D primitives with 2D primitives

- Only 2D translation control

Experiments

Baselines

Ours 2D

Replace 3D primitives with 2D primitives

- Only 2D translation control
- Fails to disentangle rotation and translation

Experiments

Ours: Object Translation

Experiments

Ours: Object Rotation

Experiments

Ours: Camera Rotation

Experiments

Ours: Camera Rotation

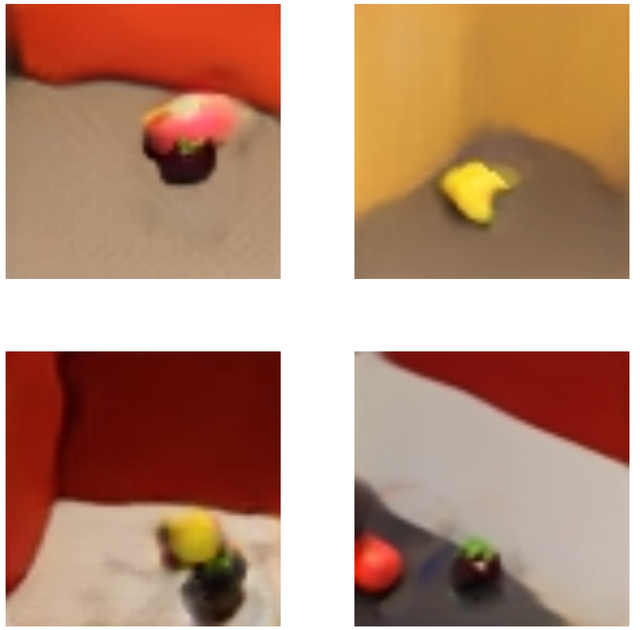
Experiments

Ours: Camera Rotation




Experiments

Ours: Object Translation

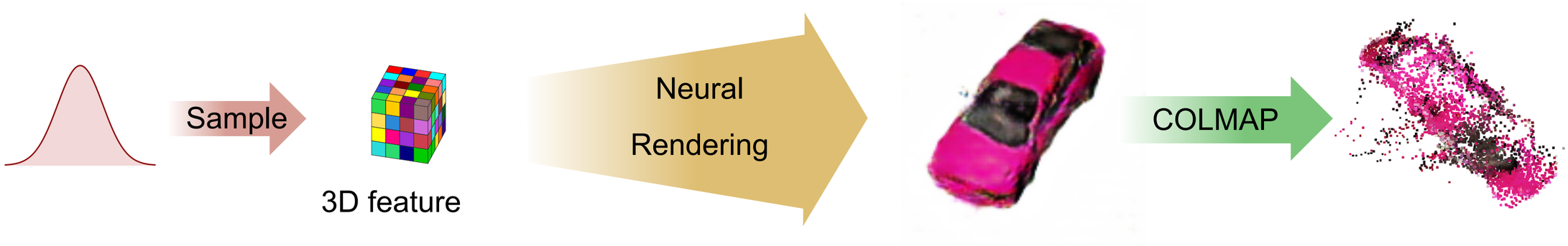


Is there a better 3D representation for
3D controllable image synthesis?

Schwarz*, Liao*, Niemeyer, Geiger: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. NeurIPS, 2020

3D Representations

3D Latent Feature with Learnable Projection

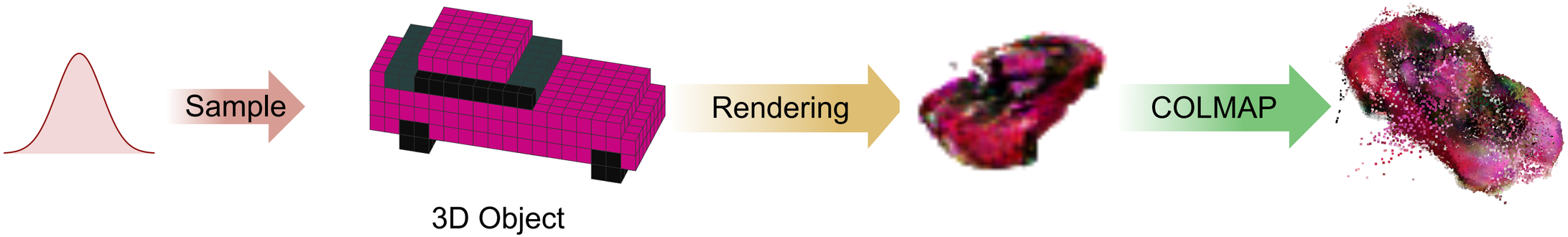


HoloGAN [Nguyen et al., ICCV 2019]

- \+ High image fidelity
- — Object identity may vary with viewpoint due to learnable projection

3D Representations

3D Shape with Volumetric Rendering

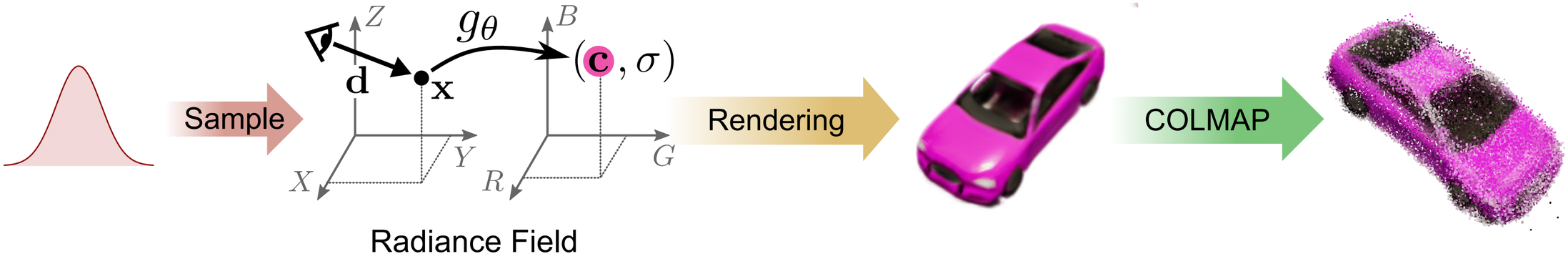


PlatonicGAN [Henzler et al., ICCV 2019]

- \+ Multi-view consistent
- — Low image fidelity, high memory consumption

3D Representations

Generative Radiance Fields



- ✚ Continuous representation, multi-view consistent
- ✚ High image fidelity, low memory consumption

Experiments

Comparison to Baselines

PlatonicGAN 

HoloGAN 

Ours 

64×64

Experiments

Scalability to High Resolution

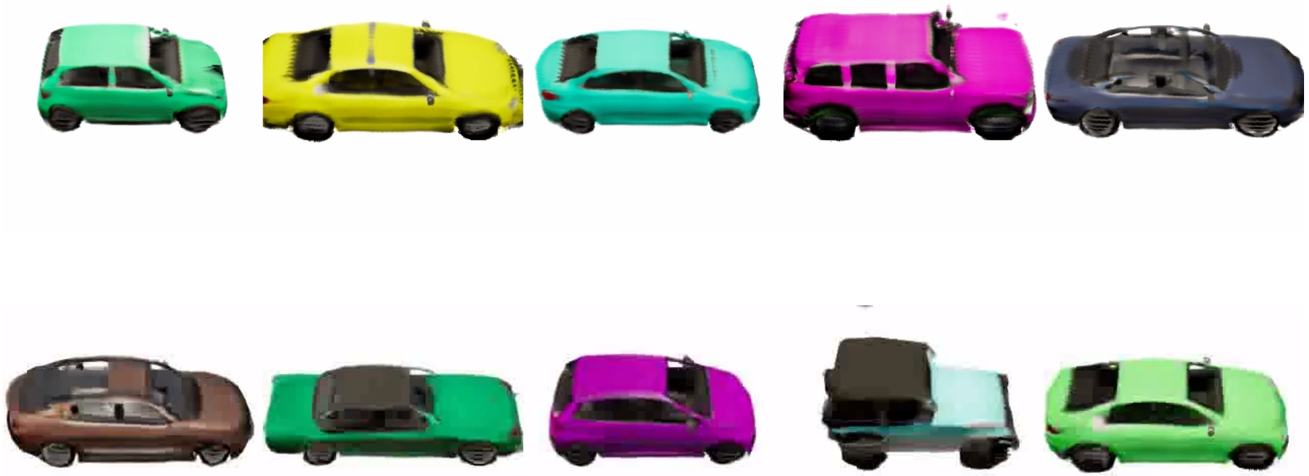
HoloGAN

Ours

256×256

Experiments

Disentangling Shape & Appearance



256×256

Experiments

Real-World Datasets

Training Images

64 × 64

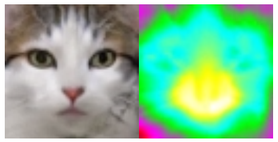
Training Images

256 × 256

Failure Cases

White artifacts

Inward facing depth

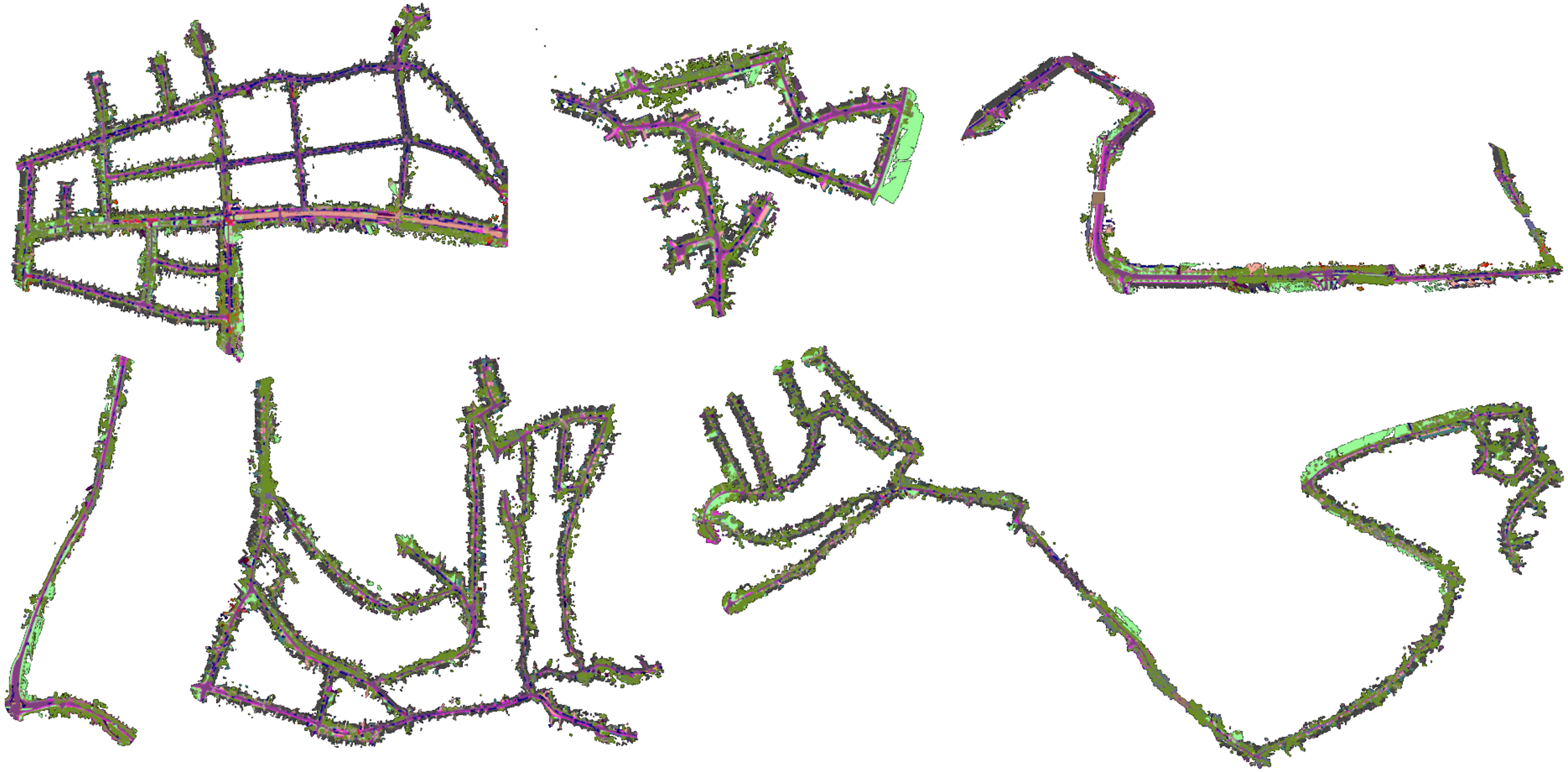


More complex real world data

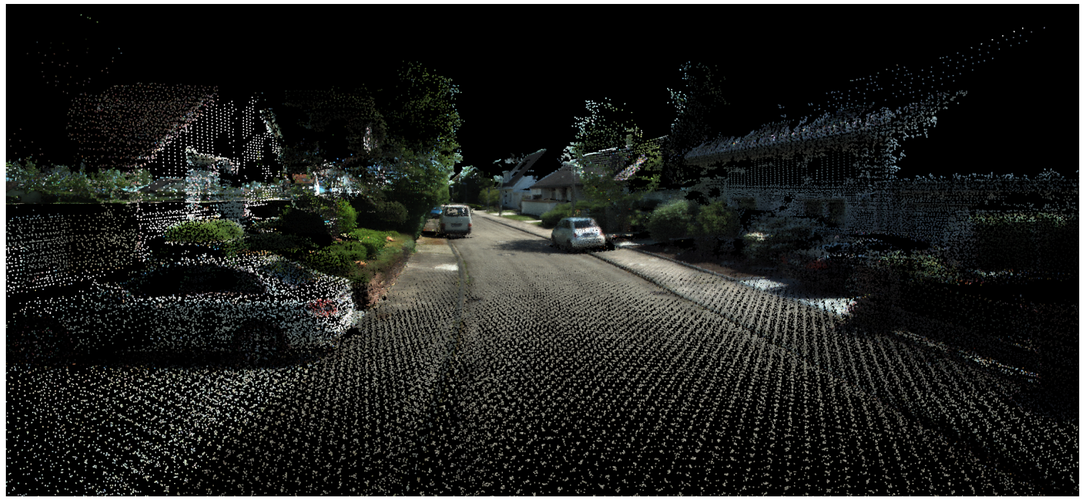
- Incorporate more supervision
- Object disentanglement

What's next?

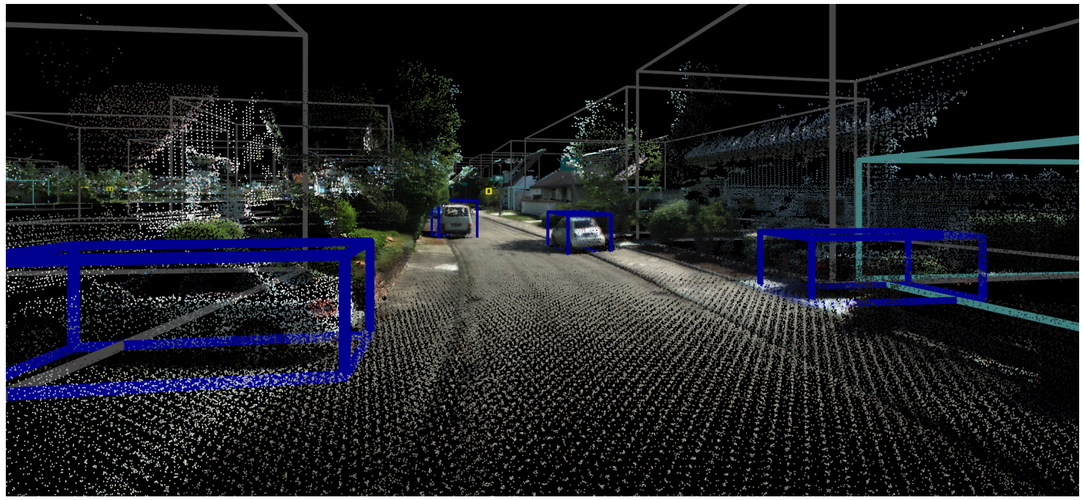
KITTI-360



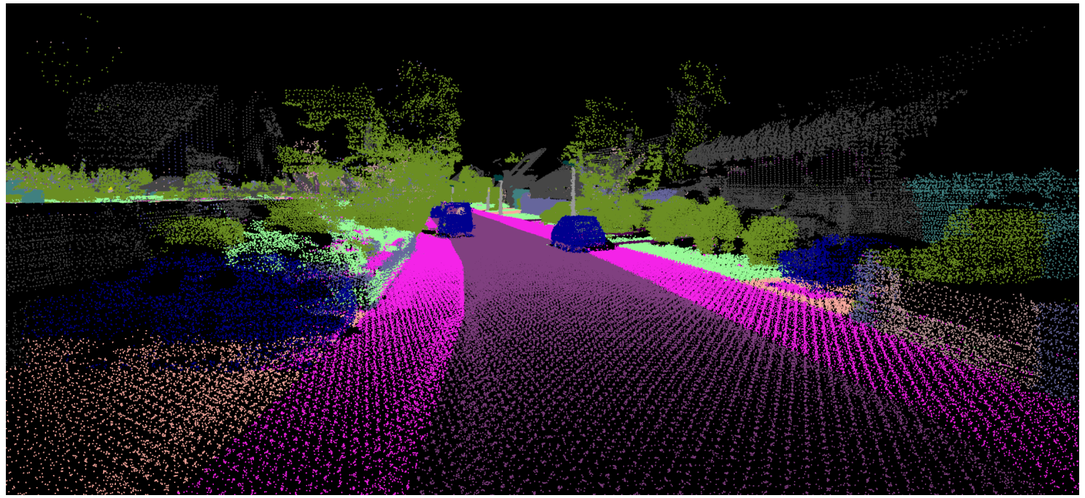
KITTI-360



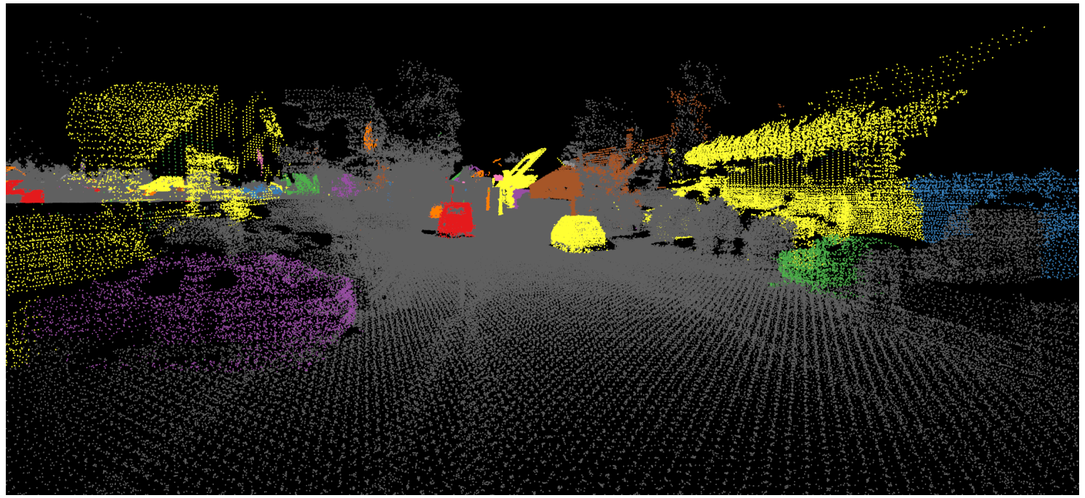
RGB



Bounding Box



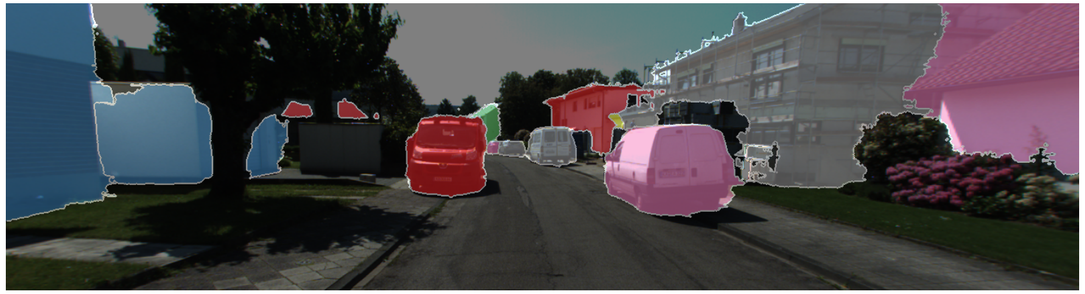
Semantic



Instance

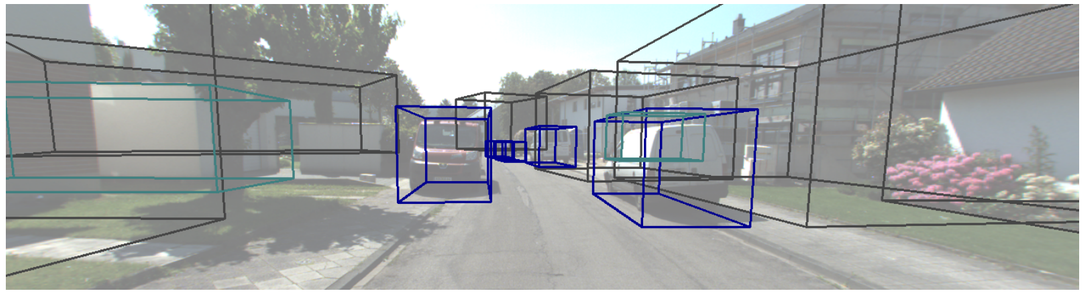
KITTI-360

Semantic

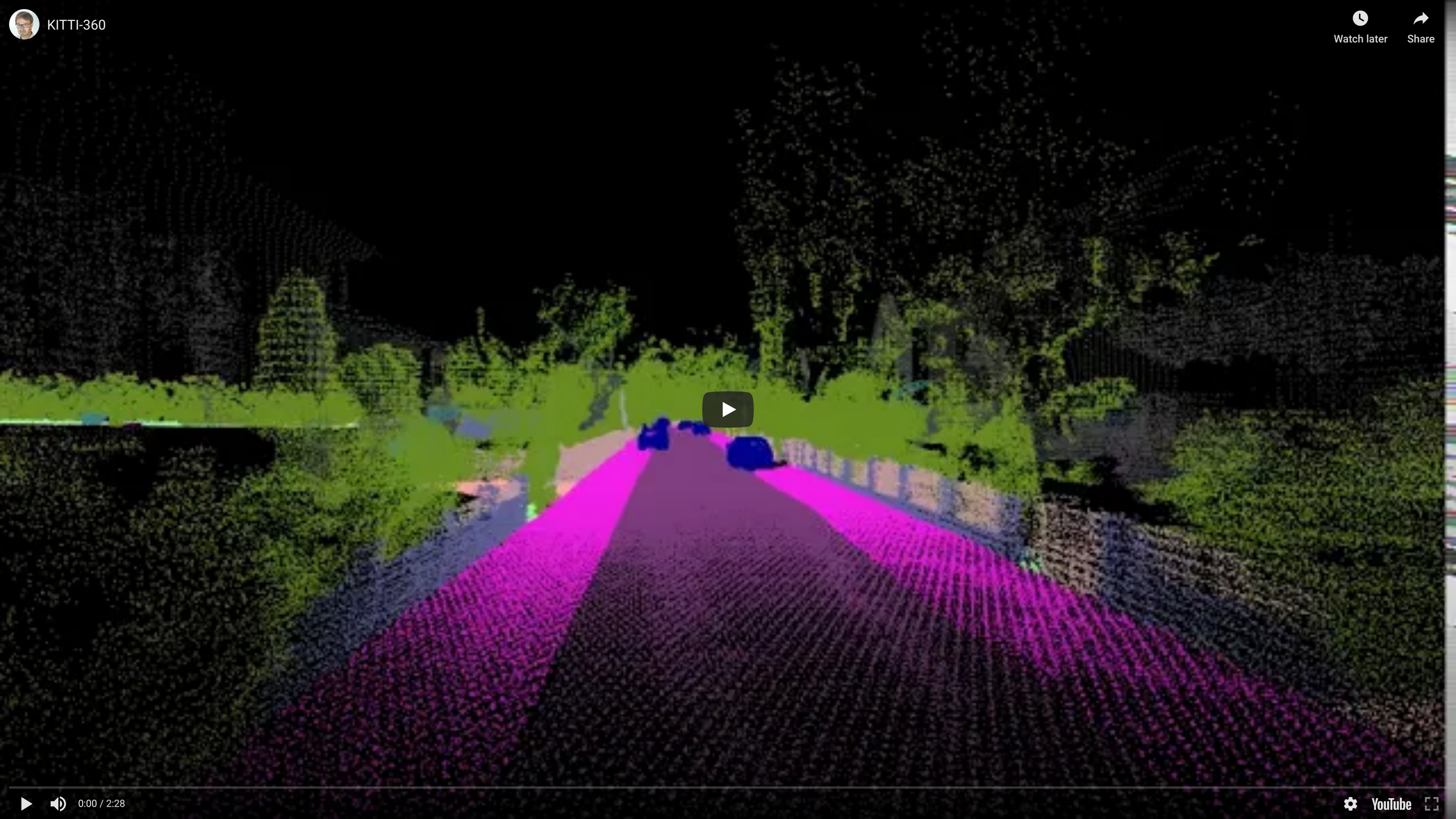


Instance

Confidence



Bounding Box


KITTI-360
Watch later
Share
0:00 / 2:28
YouTube

KITTI-360

- **Simulation**
  - 3D Controllable image synthesis
  - Novel view & semantic synthesis
- **Perception**
  - Semantic & instance segmentation in 2D & 3D
  - Holistic scene understanding
- **Robotics**
  - Semantic SLAM

Thank you!

Live slides with videos: yiyiliao.github.io/20201224_GAMES