GeoSim: Realistic Video Simulation via Geometry-Aware Composition for Self-Driving

Yun Chen[1*] Frieda Rong[1,3*] Shivam Duggal[1*] Shenlong Wang[1,2] Xinchen Yan[1]
Sivabalan Manivasagam[1,2] Shangjie Xue[1,4] Ersin Yumer[1] Raquel Urtasun[1,2]
[1]Uber Advanced Technologies Group [2]University of Toronto
[3] Stanford University [4]Massachusetts Institute of Techonology

Presented by Annie Wang & Jun Myeong Choi

Introduction

Self-Driving Car

- Safety(Stability) testing is crucial stage
- Costly and risky to test them in the real world

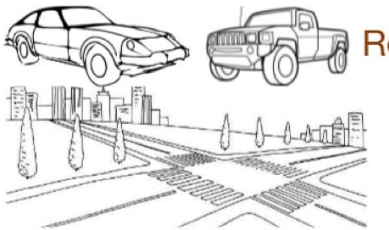
Video Simulation

- Easy to validate self-driving systems

Existing Methods

Graphics Approach

3D Assets



Render

Simulated Image

Pros: 3D-aware. High-level Control

Cons: Costly. Realism gap

Image Editing Approach

Semantic Input

Render

Simulated Image

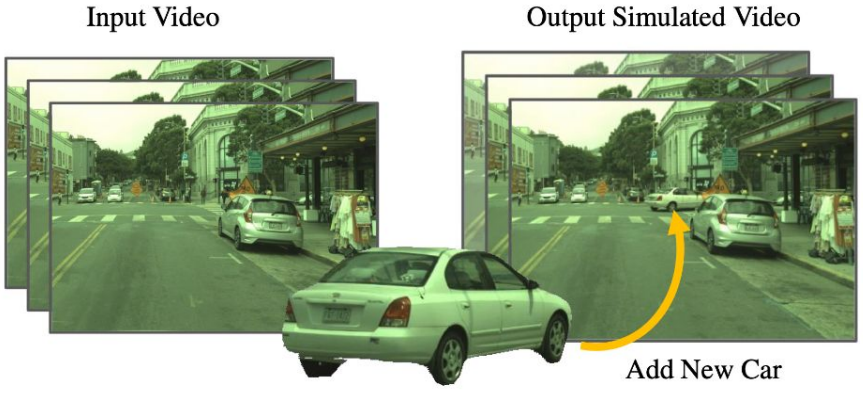
Pros: Low-cost. Large-Scale. Diverse

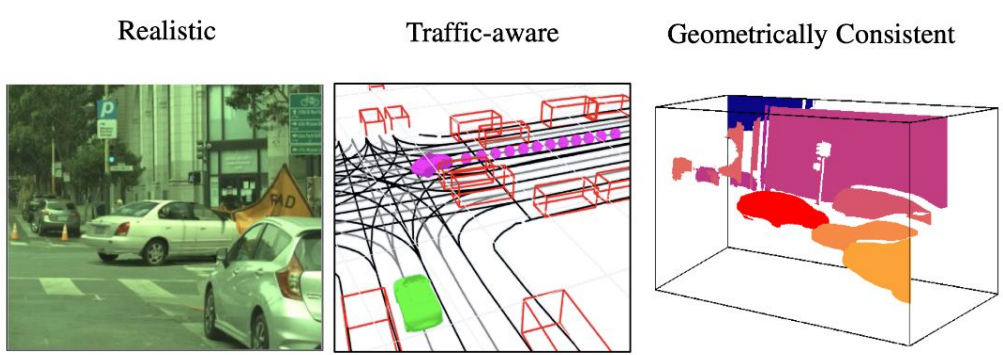
Cons: 2D-aware. Uncontrollable. Artifacts

Key Ideas

Graphics Approach + Image Editing Approach

- Reconstruct a large bank of 3D assets
- Leverage the 3D asset bank to geometrically simulate new objects into existing videos





Overview of methods

2 main parts:

- 3D reconstruction of assets
  - Leverage data captured by self-driving vehicles to reconstruct the objects around us
  - Self-supervised learning method
- Geometry-aware image simulation
  - places novel objects into an existing 3D scene and generates a high-quality video sequence of the composition

Multi-sensor 3D asset reconstruction

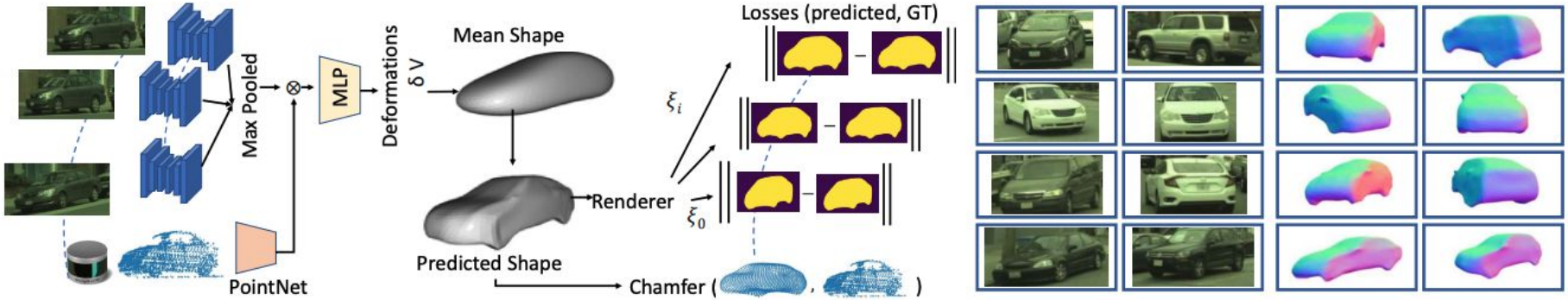
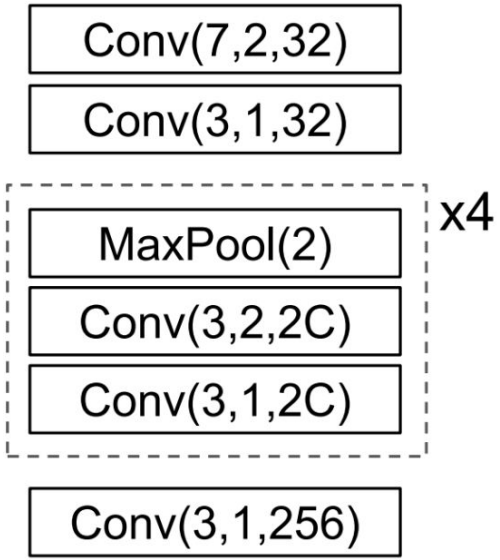


Figure 2: **Realistic 3D assets creation.** Left: multi-view multi-sensor reconstruction network; Right: 3D asset samples. For each sample we show one of the source images and the 3D mesh.

3D reconstruction network architecture



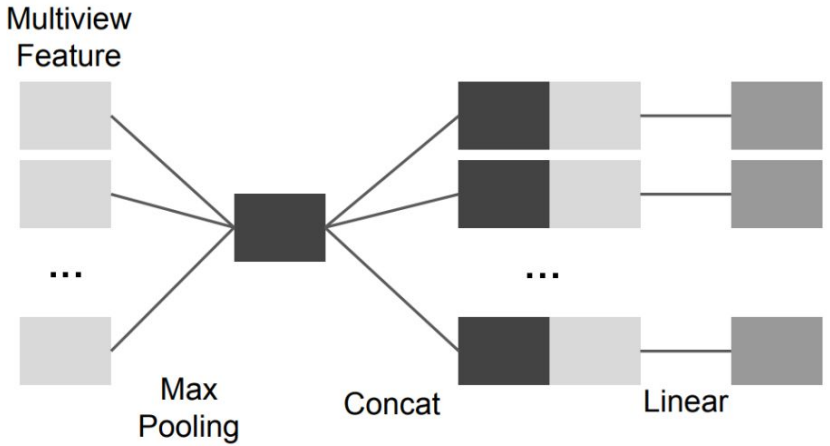


Figure 8: **3D reconstruction network architecture.** Left: Image feature extraction backbone; Right: Multi-view image fusion block.

How does it learn?

- No shape supervision
  - Solution - train end-to-end with self-supervision
- Guided by how the 3D shape agrees with the camera and LiDAR observations

$$\ell_{\text{total}} = \sum_i \{\ell_{\text{sil}}(\mathbf{M}_i; \mathbf{P}_i, \mathbf{I}_i) + \ell_{\text{lidar}}(\mathbf{M}_i; \mathbf{X}_i) + \ell_{\text{reg}}(\mathbf{M}_i)\}$$

Silhouette loss

$$\ell_{\text{sil}}(\mathbf{M}_i; \mathbf{P}_i, \mathbf{I}_i) = \sum_j \|\mathbf{S}_{i,j} - \tau(\mathbf{M_{i,j}}, \mathbf{P_{i,j}})\|_2^2$$

- $\mathbf{S_{i,j}} \in \mathbb{R}^{H \times W}$ is 2D silhouette
- $\tau(\mathbf{M}, \mathbf{P})$ is a differentiable neural rendering operator that renders a differentiable mask on the camera image given a projection matrix P

LiDAR loss

$$\ell_{\text{lidar}}(\mathbf{M_i}, \mathbf{X_i}) = \sum_{\mathbf{x} \in \mathbf{X}_i} \min_{\mathbf{v} \in \mathbf{V}_i} \|\mathbf{x} - \mathbf{v}\|_2^2$$

- represents the consistency between the accumulated LiDAR point cloud and the mesh vertices
- $\mathbf{X}_i$ is aggregated set of LiDAR points for i-th object

Multi-sensor 3D asset reconstruction

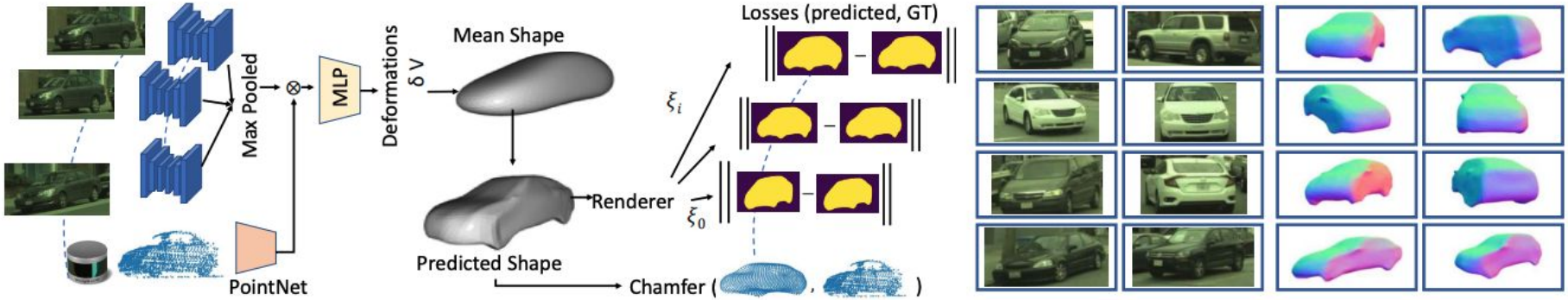


Figure 2: **Realistic 3D assets creation.** Left: multi-view multi-sensor reconstruction network; Right: 3D asset samples. For each sample we show one of the source images and the 3D mesh.

Geometry-aware image simulation

Overview:

- Places novel objects into an existing 3D scene
- Input: camera video footage, LiDAR point clouds, and an HD map in the form of a lane graph
- Output: video with novel objects inserted into the scene

Summary of simulation steps

- Generate scenario
- Use novel-view rendering with 3D occlusion reasoning to create new image
- Use neural network to fill in the boundary of the inserted objects, create any missing texture and handle inconsistent lighting.

Generating the scenario

What do we want to add? Where?

Summary of approach:

- Infer the location of all objects in the scene by performing 3D object detection and tracking
- For each new object to be inserted, select where to place it as well as which asset to use based on the HD map and the existing detected traffic
- Use an intelligent traffic model for newly placed object such that its motion is realistic, takes into account the interactions with other actors and avoids collision

Selecting where to place the object

3D sampling procedure

- exploit HD maps that contain the location of the lanes in bird's eye view (BEV), and parameterize the object placement as a tuple (x, y, θ) defining the object center and orientation in BEV
- randomly sample a placement (x, y) from the lane regions lying within the camera's field of view and retrieve the orientation from the lane
- reject all samples that result in collision with other actors or background objects

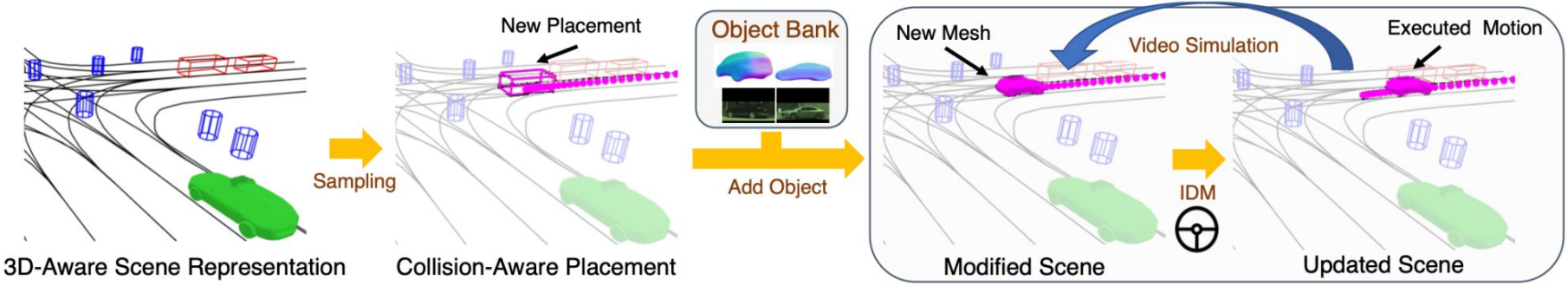


Figure 3: **3D-aware object placement, segment retrieval, and temporal simulation.**

Which object to place there?

- Use objects from the asset bank that were viewed with similar viewpoints and distance to the camera in the original footage
- Objects are sampled according to a categorical distribution weighted by their inverse score
- Perform collision checking again with specific sampled shape to make sure it is valid

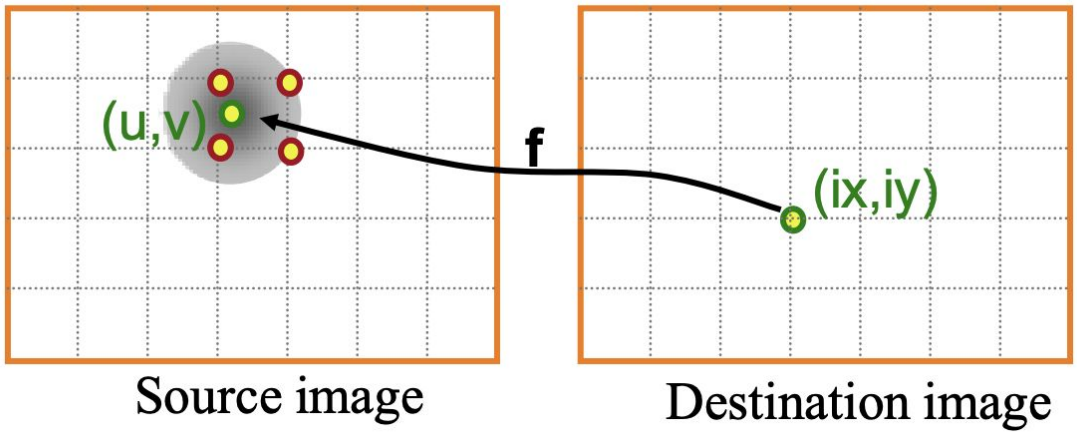
Rendering the object into the scene

Summary of approach:

1. Novel view warping
2. Add shadows
3. Occlusion reasoning
4. Post-composition synthesis

1. Novel view warping

Inverse warping operation



1. Novel view warping

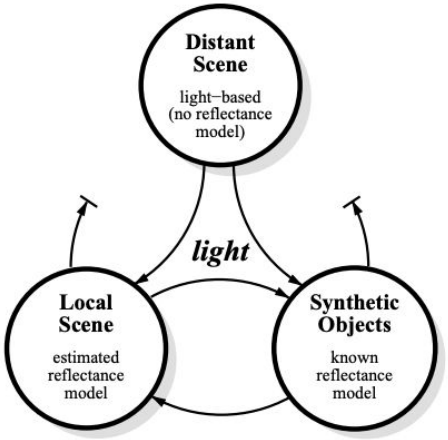
$$\mathbf{I}_t = \mathbf{I}_s(\pi(\pi^{-1}(\mathbf{D}_t, \mathbf{P}_t), \mathbf{P}_s)) \text{ , where } \quad \mathbf{D}_t = \psi(\mathbf{M}, \mathbf{P}_t)$$

$\mathbf{M}$	is the object's 3D mesh
$\mathbf{I}_s$	is the source object's camera image
$\mathbf{P}_s/\mathbf{P}_t$	are the source/target camera matrices
ψ	is a differentiable neural renderer
$\mathbf{D}_t$	is the target depth map

2. Generate shadows

Render shadows with graphics engine

- Construct virtual scene with inserted object
- Blend image intensities of scene with and without object
- Cloudy lighting used with good results



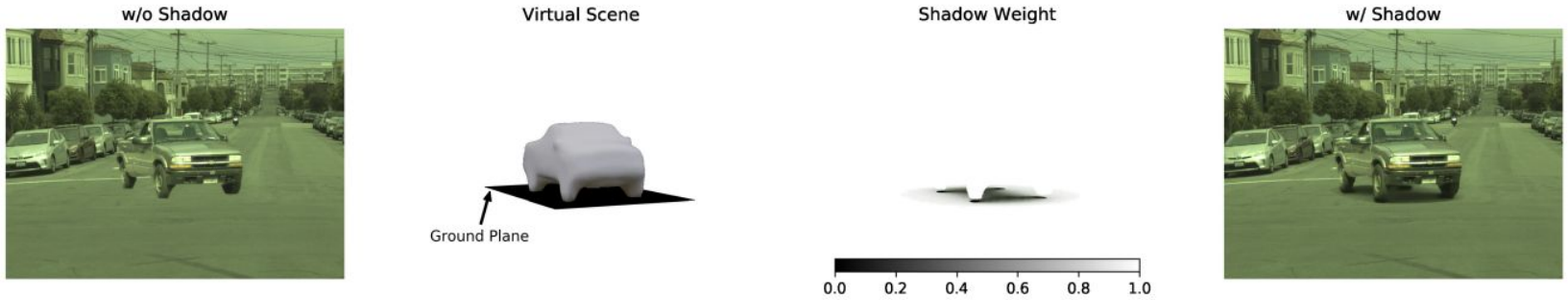


Figure 9: **Schematics of shadow generation**. (left to right): result without shadow, schematics of virtual scene, shadow weight (ratio of intensity between rendered image with inserted object and without inserted object), result with shadow

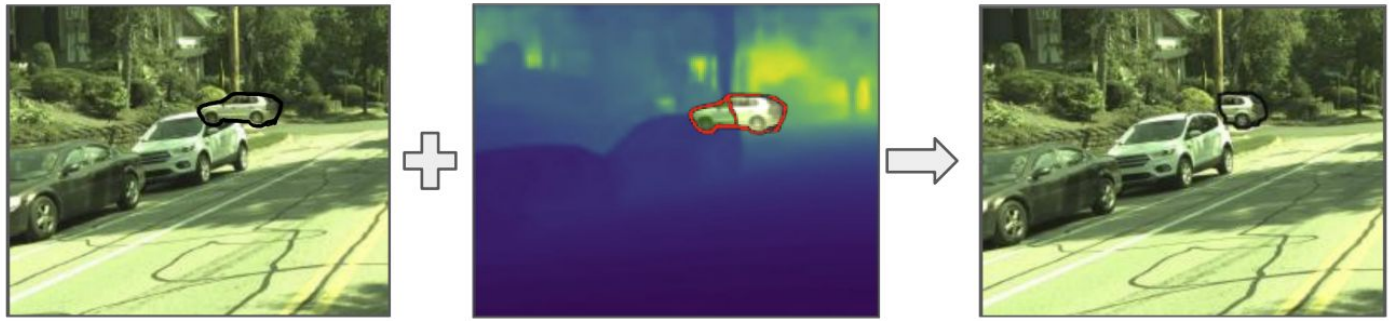
3. Occlusion reasoning

Simple approach: compare depth maps

Compute target image's dense depth map through depth completion network

- Input: RGB image and sparse depth map
- Output: dense depth map

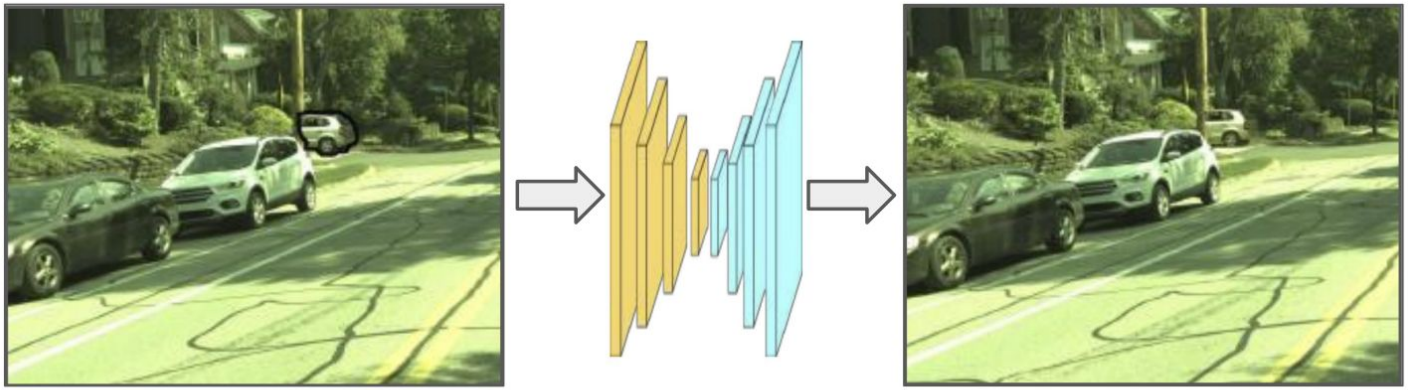
Compute occlusion mask through per-pixel comparison



4. Post-composition synthesis

Use image synthesis network to blend the source segment to target scene

- Input: background image, rendered target object, and object binary mask
- Output: final blended image that looks realistic



Data preprocessing for synthesis network training

Network inputs:

- object segment
- target scene with 512x512 region around object center cropped
- mask region

Figure 10: **Input data preparation for training the synthesis network**. From left to right: scene image I, object segment S and mask M and three random data augmentation including color-jitter, segment boundaries erosion-expansion and random mask in the boundary.

Synthesis network objective

- Inpainting network architecture
- Loss functions:
  - Perceptual loss

$$L_G^{\text{perc}} = \sum \|F_v(I) - F_v(G(I \cdot (1 - M^A), M^A, S^A))\|_1$$

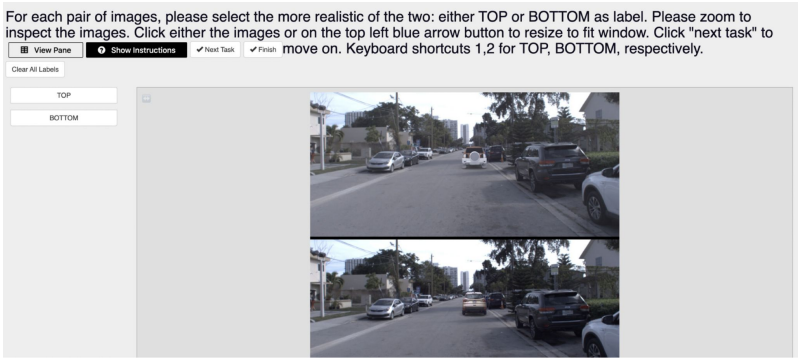
  - GAN loss

$$L_G^{\text{gan}} = -\mathbb{E}_{z \sim P_z(z)}[D(G(I \cdot (1 - M^A), M^A, S^A))]$$

Evaluation

Dataset

- UrbanData
- Argoverse



Metrics

- Human study (user study)
- Perceptual quality score (FID)

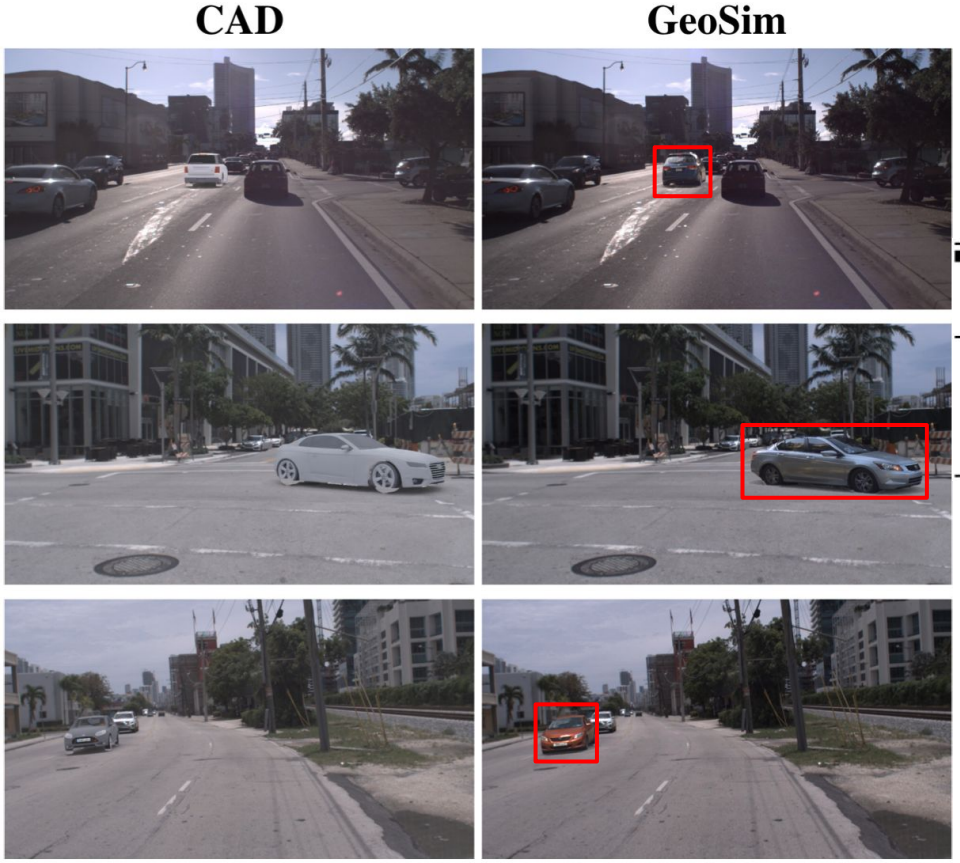
$$FID = d^2 = \|\mu_1 - \mu_2\|_2^2 - Tr(\Sigma_1 + \Sigma_2 - 2\Sigma_1\Sigma_2)$$

Comparison of image simulation approach

Comparison of image simulation approach

Method	Human Score (%)	FID
SPADE [58]	99.3	43.2
Guided Editing [29]	94.3	20.3
Cut-Paste [20]	98.5	22.1
CAD [2]	94.3	17.3
GeoSim	-	**14.3**

Comparison of image simulation approach (Argoverse)



Method	Human Score (%)	FID
CAD	84.0	28.3
GeoSim	-	**24.5**

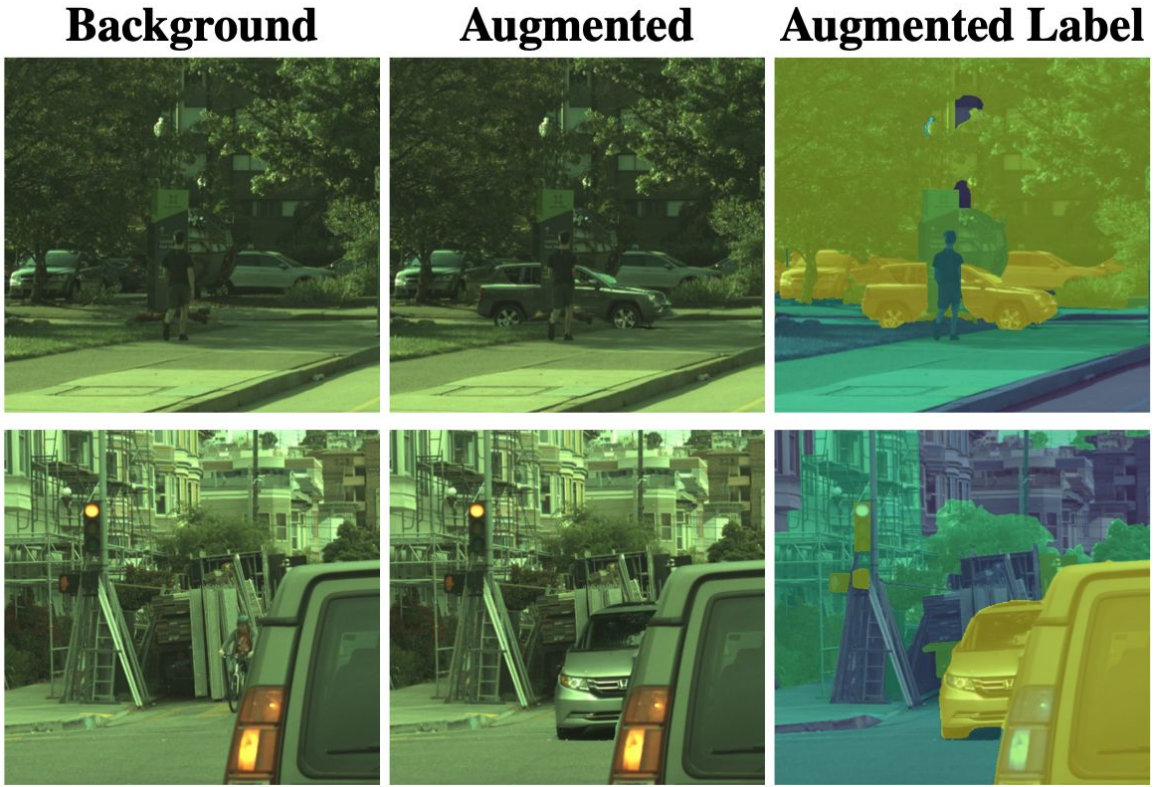
Ablation on Rendering options

Approach	Shadow	Human Score (%)	FID
Physics	Yes	94.2	17.3
2D Synthesis	-	75.7	**13.7**
Geo Synthesis	No	71.9	**13.7**
Geo Synthesis	Yes	-	14.3

FID : With Shadow > without Shadow

-> a gap between perceptual measurements and humans' criteria.

Performance boost using data augmentation



Method	PSPNet [83]		DeepLabv3 [10]	
	mIOU	carIOU	mIOU	carIOU
Real	93.5	87.8	94.0	88.7
Real+GeoSim	**95.3**	**91.2**	**94.2**	**89.2**

Failure cases

- Incorrect occlusion relationships
- Inaccurate object pose
- Irregular reconstructed mesh
- Illumination failure

Thank you