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

Yun Chen*, Frieda Rong*, Shivam Duggal*, Shenlong Wang, Xinchen Yan,
Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, Raquel Urtasun

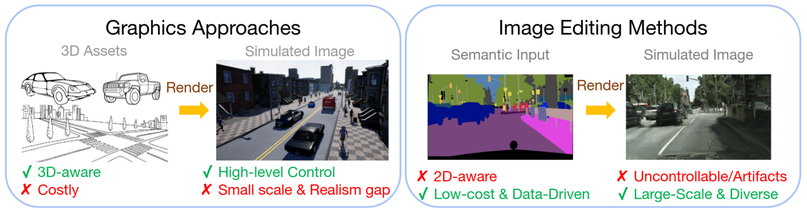
Uber ATG · University of Toronto · MIT · Stanford University

CVPR 2021 Best Paper Candidate

Introduction

Scalable camera simulation is critical for validating self-driving systems because it is costly and risky to test them in the real world.

Existing methods can be divided into 2 paradigms. But they either lack scalability, realism or 3D awareness.

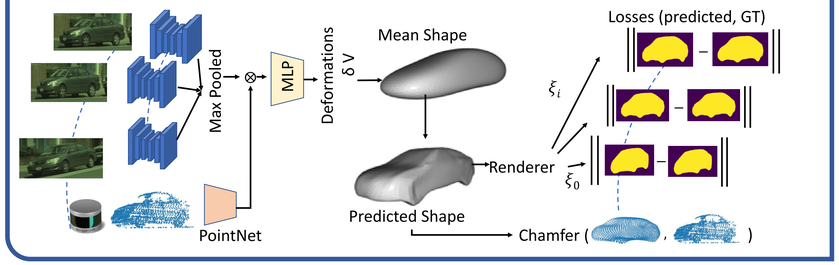


GeoSim combines best of both the worlds: graphics and image-editing methods. GeoSim first **reconstructs a large bank of 3D assets** and then leverages the 3D asset bank to **geometrically simulate new objects** into existing videos.

Asset Creation

GeoSim reconstructs a large asset bank of 3D vehicles from real-world data in the wild without any groud-truth 3D supervision. Each asset is associated with accurate pose, shape and texture. Key components of reconstruction pipeline:

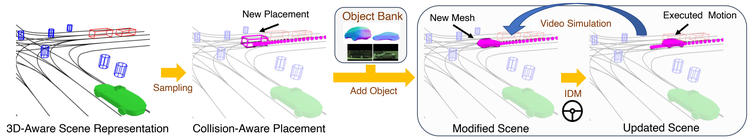
1. **Multi-sensor feature extraction** from images and partial LiDAR sweeps.
2. **3D mesh reconstruction** in form of per vertex defomrations on top of a learnable mean shape.
3. **Differentiable rendering** to render 2D silhouettes maps and **self supervised learning** through silhouette loss/ 3D chamfer loss (+regularizers).



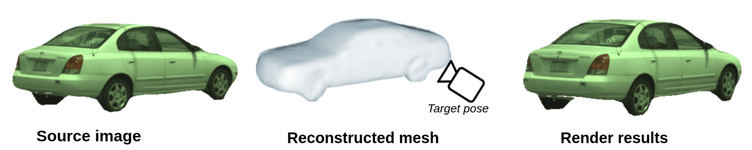
Simulation Pipeline

With the reconstructed asset bank, GeoSim can automatically simulate new objects in videos. The approach is scalable and the results are realistic and geometrically-consistent.

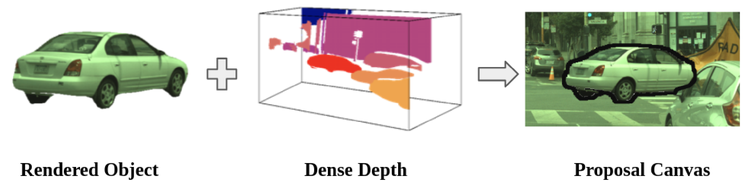
1. Scenario Generation: (1) 3D-aware placement using HD Map, (2) 3D asset selection for object insertion, (3) 3D object trajectory simulation for video generation.



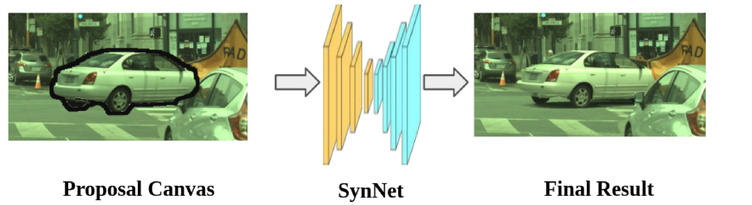
2. The selected 3D object is rendered at novel target viewpoint with object appearance generated through inverse-warping. Corresponding shadows are also rendered onto the backgound scene image.



3. Dense depth map is utilized to determine occlusions between the rendered object (and corresponding shadow) and the background scene.

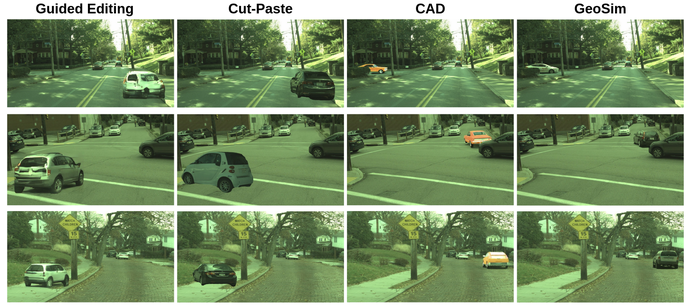


4. Post-composition refinement is applied to refine the simuated image by correcting inconsistent illuminations/ discrepancies in object appearance.



Experiments

- Qualitative comparison of image simulation approaches on UrbanData.



- Quantitative comparison (left) and ablation on rendering approaches (right). *HS*: human score (% of participants who prefer our GeoSim results over baseline)

Method	HS(%)	FID ↓
Guided Editing	94.3	20.3
Cut-Paste	98.5	22.1
CAD	94.3	17.3
GeoSim	-	**14.3**

Approach	Shadow	HS (%)	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

- Qualitative on Argoverse and video results on UrbanData.

Video results in 4K
tinyurl.com/cvpr2021

- **Sim2Real:** Augmenting labeled real data with GeoSim results leads to consistent improvement in segmentation.