GenAssets: Generating in-the-wild 3D Assets in Latent Space

Ze Yang, Jingkang Wang, Haowei Zhang, Sivabalan Manivasagam, Yun Chen, Raquel Urtasun

https://waabi.ai/genassets

UNIVERSITY OF TORONTO

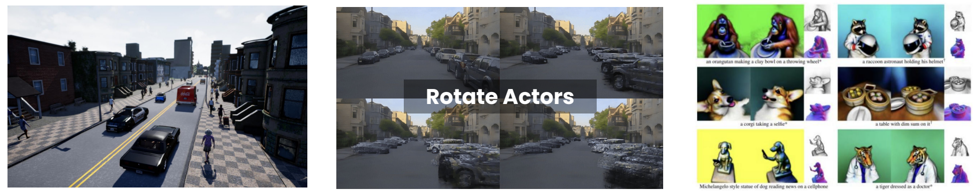
CVPR Nashville JUNE 11-15, 2025

Motivation

+ **Simulation is Important for Developing Safe Autonomous Systems**

World Modeling (Digital Twin Creation) Simulation

+ **Existing Approaches:**

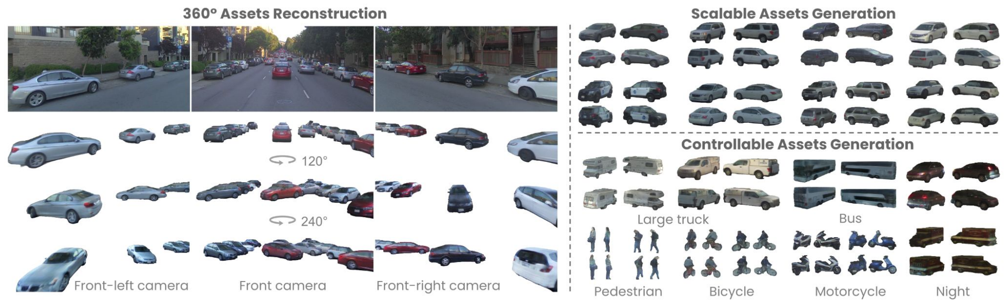


Artistic-created CAD assets
- Requires manual efforts
- Lacks diversity
- Lacks Realism

Per-scene Reconstruction
- Slow optimization
- Lacks diversity
- Lacks Completion

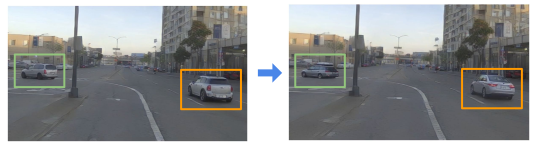
Generative Models
- Synthetic / Object-centric
- Lacks 3D supervision
- Domain Gap

+ **GenAssets: Learning Assets Recon/Generation from in-the-wild Data**



GenAssets for Simulation

+ **Replacing Existing Actors with Generated Counterparts**



	mAP↑	AP@1m↑	AP@2m↑
Real	27.08	8.58	26.99
Real + Sim	**29.32**	**9.78**	**29.18**

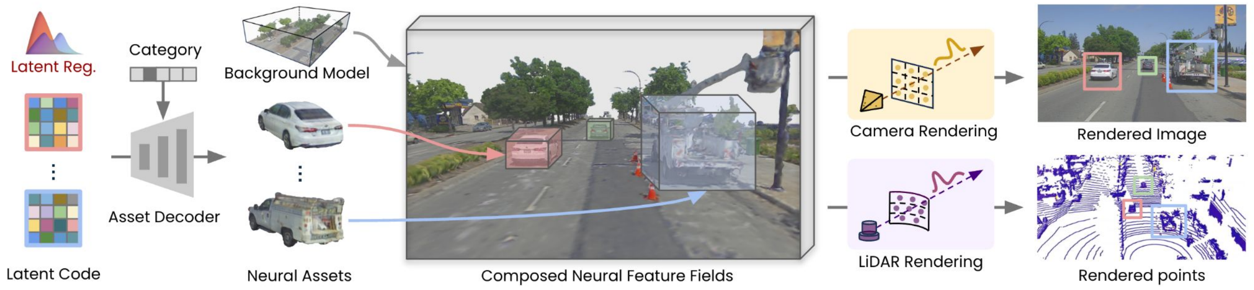
GenAssets augmentation improves 3D detection

+ **Generating Extreme Scenario Variations**

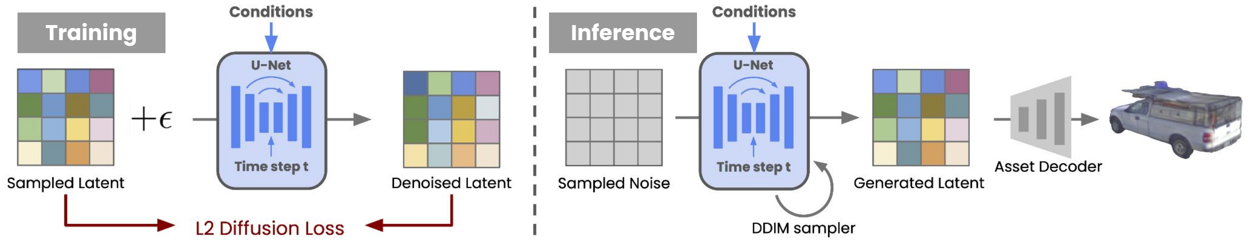
The actor ahead executes a U-turn

GenAssets Method

+ **Learning Latent Asset Representations via Occlusion-Aware Neural Rendering**
  - ***Motivation:*** Learning a compact and complete latent space for neural assets
  - ***Approach***
    - Each assets is represented as a low dimensional *latent code*
    - A shared *asset decoder* is trained to map the latent code into *neural assets*
    - The *neural assets* is composed with learnable background models to form a compositional neural scene representation
    - The scene representation is rendered to match real-world sensor observations
  - ***Benefit***
    - Trained across many scenes to learn asset priors
    - Compact space learning to reduce computation and memory for large datasets
    - Latent bottleneck encourages the model to learn asset priors
    - Infer occluded or unobserved regions from sparse observations.



+ **Learning Latent Asset Diffusion Model**
  - ***Motivation:*** Learning asset generation in the latent space with diffusion model
  - ***Approach***
    - Training a diffusion probabilistic model in the latent space
    - Sample from the learned diffusion priors using the DDIM solver
    - The *asset decoder* decodes the generated latent code to the neural asset
  - ***Benefit***
    - Focusing on essential contents of the data
    - Operating in a computationally efficient, compact space
    - Support both conditional or unconditional generation



+ **Learning Objective**
  - Reconstruction loss and regularization in latent space

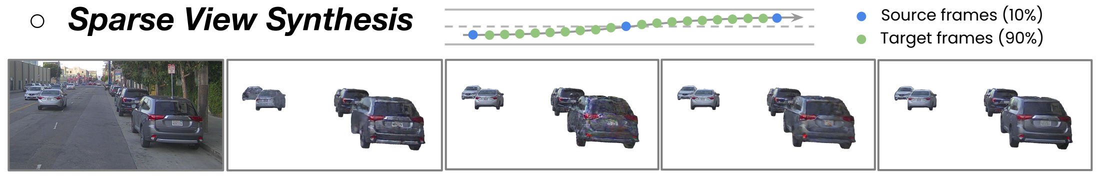
$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{perp}}\mathcal{L}_{\text{perp}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{lid}}\mathcal{L}_{\text{lid}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}$$

  - Latent diffusion loss

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{c},\epsilon,t}\left[\frac{1}{2}w^{(t)}\|f_{\text{diff}}(\mathbf{c}^{(t)},t) - \epsilon\|_2^2\right]$$

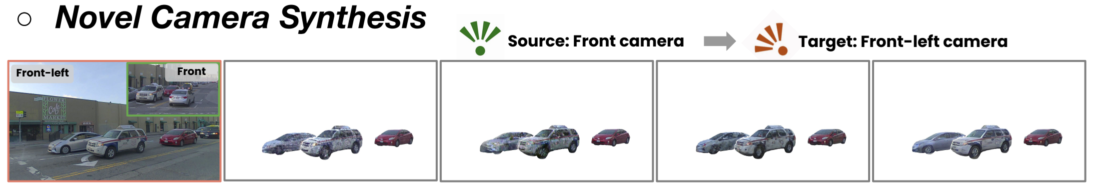
Results

+ **Asset Reconstruction**
  - ***Sparse View Synthesis***



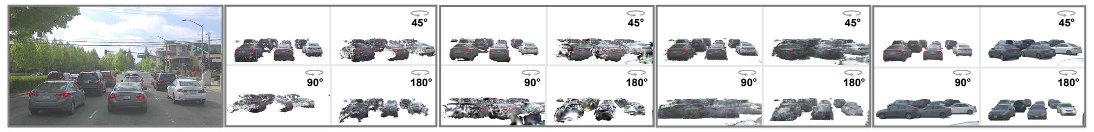
GT G3R Street Gaussian NeuRAD Ours

  - ***Novel Camera Synthesis***



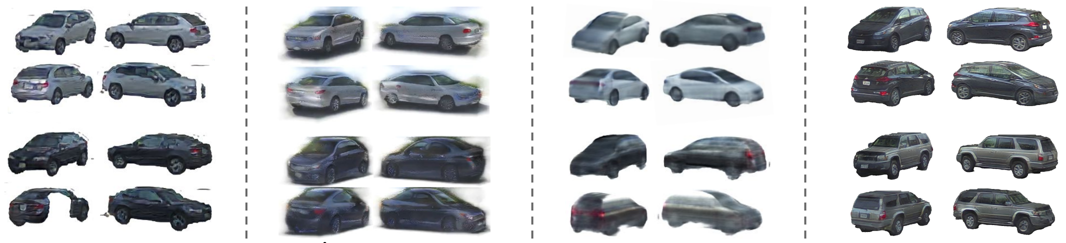
GT G3R Street Gaussian NeuRAD Ours

  - ***360° Synthesis***



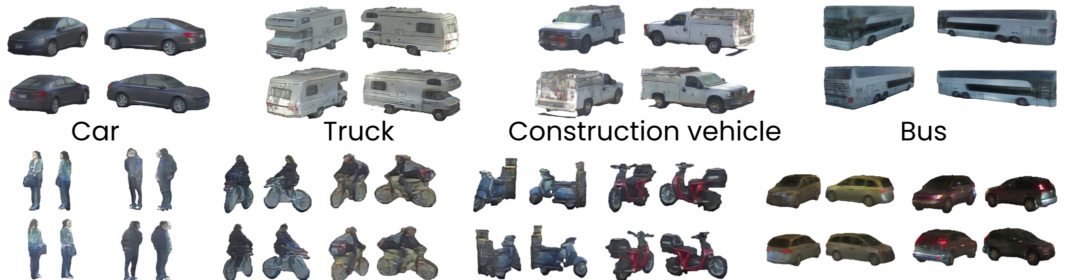
GT G3R Street Gaussian NeuRAD Ours

+ **Asset Generation**
  - ***Unconditional Generation***



EG3D DiscoScene SSDNeRF Ours

  - ***Conditional Generation***



  - ***Single Image to 3D Generation***