HUGS: Holistic Urban 3D Scene Understanding via Gaussian Splatting

Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, Yiyi Liao

Zhejiang University, Huawei Noah's Ark Lab, University of Tübingen, Tübingen AI Center

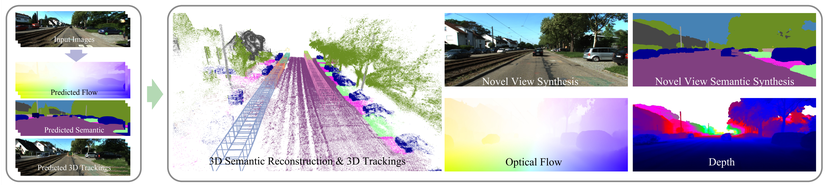
Highlights

Key idea: We reconstruct and understand 3D urban scenes from sparse input views by decomposing static and dynamic objects. Dynamic objects can be reconstructed even with noisy labels by jointly optimizing the 3D bounding boxes under physical constraints.

- Our method addresses the task of dynamic 3D urban scene understanding by extending Gaussian Splatting to model additional modalities as well as dynamic objects.
- We achieve the decomposition of static and multiple dynamic objects from sparse urban images and noisy labels by incorporating physical constraints through the unicycle model, omitting the requirement of ground truth 3D bounding boxes for reconstructing dynamic scenes.
- Our method achieves state-of-the-art performance on various benchmarks, including novel view appearance and semantic synthesis, as well as 3D semantic reconstruction.

Overview

- Given posed RGB images as input, our method lifts noisy 2D & 3D predictions to the 3D space via decomposed 3D Gaussians, and enables holistic scene understanding in 2D and 3D space.

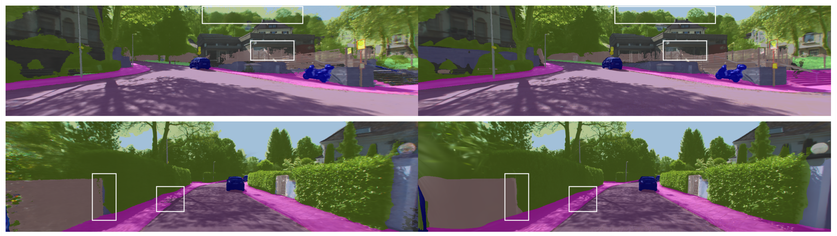


HUGS Lifts Noisy Predictions to the 3D Space

Results on KITTI-360

MARS Ours MARS Ours MARS Ours

Qualitative Comparison on KITTI-360 RGB



PNF Ours

Qualitative Comparison on KITTI-360 Semantic

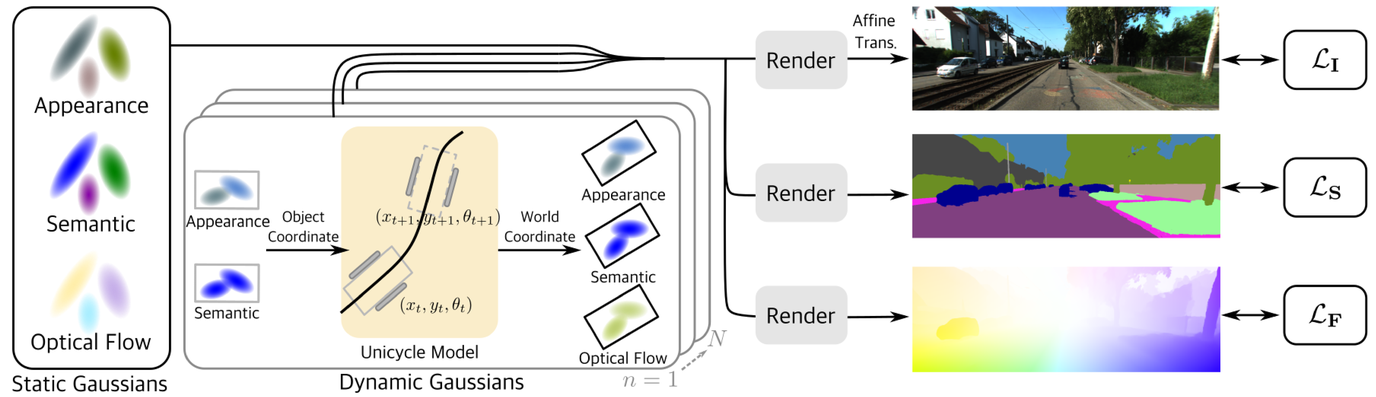
Contact

- Project Page: `https://xdimlab.github.io/hugs_website`
- Paper: `https://arxiv.org/abs/2403.12722`
- E-mail: `hongyu_zhou@zju.edu.cn`

Project Page **Paper**

Pipeline

- Scenes are decomposed into static regions and rigidly moving objects represented by 3D Gaussians.
- Using N unicycle models to individually represent the motion of N dynamic objects.
- Each 3D Gaussian encompasses multi-modal information, enabling holistic understanding within a unified model.



Results on KITTI & vKITTI

NSG MARS Ours

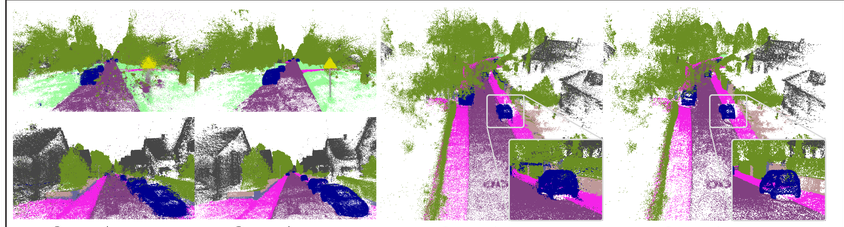
Qualitative Comparison with Predicted / Noisy BBox

Editing

Original Background Edited

Decomposed Foreground & Background Scene Editing

Ablation

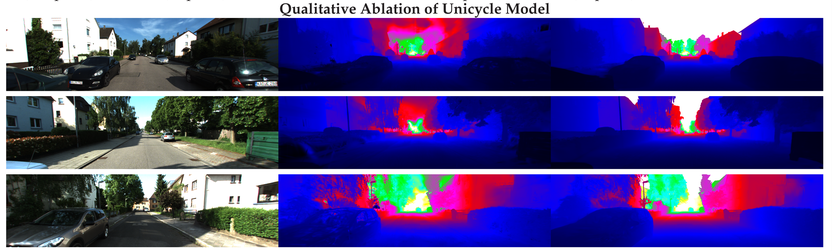


Ours w/ $_{\text{2D\_norm}}$ Ours w/ $_{\text{3D\_norm}}$ Ours w/ $_{\text{2D\_norm}}$ Ours w/ $_{\text{3D\_norm}}$

Qualitative Ablation of Semantic Reconstruction in 3D Space

w/o opt. w/o uni. w/ opt. w/o uni. Ours w/o opt. w/o uni. w/ opt. w/o uni. Ours

Qualitative Ablation of Unicycle Model



RGB w/o L_S, w/o L_F Ours

Qualitative Ablation on Depth. The depth of sky region is set to infinite.

RGB w/o L_S, w/o L_F Ours

Qualitative Ablation on Optical Flow.