CSCI 5561

Feed-forward 3D Reconstruction for Spatial Semantic Understanding

Minseong Kweon, Haotian Zhai, Yang Liu

"Can We Make Feature-3DGS Faster, Better, and Smaller?"

Introduction:

3D Gaussian Splatting can build 3D scenes fast, and feature fields make these scenes actually understand what they see. But current pipelines are slow to set up and create way too many Gaussians. We look at how to make Feature 3DGS faster, better, and much smaller.

Preliminaries: 3DGS

3DGS basically turns a scene into a big cloud of Gaussians, each one storing its position, shape, color, opacity, and even a feature vector. We train all of these Gaussians just for this one scene using its multi-view photos. Once trained, we can drop a new camera anywhere, project the Gaussians, blend them, and instantly get a new RGB view and a feature map.

Motivation/Contribution

1.Why use VGGT: While the SfM process suffers from weak textures, occlusions, and inefficient preprocessing, VGGT[1] initializes point clouds by predicting depth at every pixel, making it more robust and enabling a fully end-to-end pipeline.

2.Why add semantic consistency loss: Semantic consistency loss encourages neighboring Gaussians to share similar feature vectors, preventing semantic drift that arises from independent Gaussian optimization. This is especially useful in underwater scenes where sparse and noisy visual cues can cause fragmented semantics.
By enforcing local coherence through kNN-based regularization, the model achieves more stable reconstruction and cleaner region boundaries.

3.Why semantic-aware Gaussian pruning:
LightGaussian[3] decides which Gaussians to remove purely from visibility, so it tends to keep large flat surfaces like walls and floors while dropping small structures and object boundaries. These small regions are visually minor but semantically important because they define what the scene actually contains. Semantic-aware pruning measures how much each Gaussian affects the semantic loss, allowing the model to preserve meaningful details and remove only background that contributes little to scene understanding.

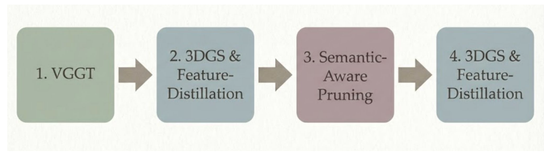
Evaluation:

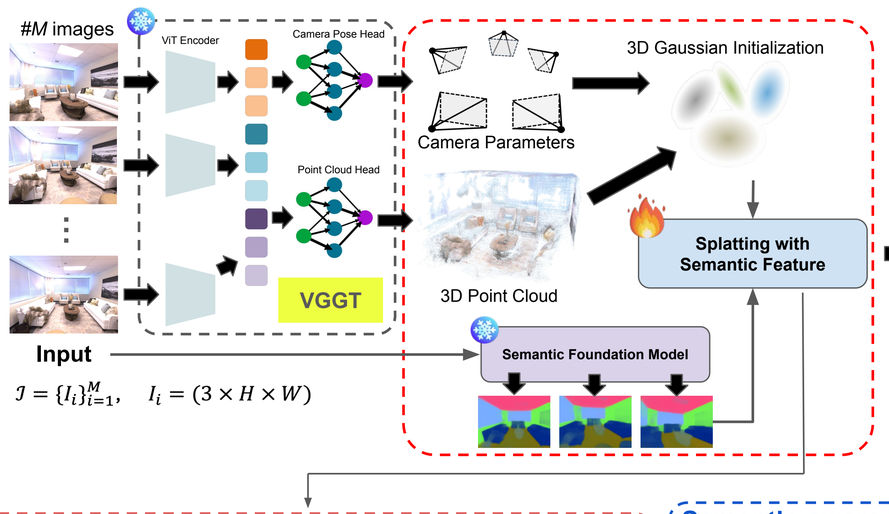
Baseline: Feature 3DGS[4]
Ours: VGGT[1] init + semantic-aware pruning & consistency loss
We test mIoU(↑) and FPS(↑) / # GS(↓)
on six scenes(replica dataset and Gopher and LindHall) .

References:

[1] Wang et al., "Vggt: Visual geometry grounded transformer", *CVPR 2025*.
[2] Li et al., "Language-driven Semantic Segmentation", *ICLR 2022*.
[3] Fan et al., "LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS", *NIPS 2024*
[4] Zhou et al., "Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields", *CVPR 2024*

Overall Pipeline





Semantic Consistency Loss

$\mathcal{L}_1 = \| S - \hat{S} \|_1$ **Feature-3DGS**

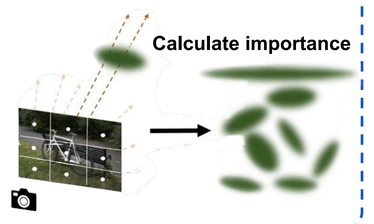
$\mathcal{L}_{\text{similarity}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} w_{ij} \left(1 - \hat{\mathbf{f}}_i^{\top} \hat{\mathbf{f}}_j\right)$

$\mathcal{L}_{\text{semantic}} = \lambda_{\text{sim}}\, \mathcal{L}_{\text{similarity}} + \lambda_1\, \mathcal{L}_1$ **(Ours)**

$\mathbf{x}_i$: 3D Gaussian center
$\mathbf{f}_i$: semantic embedding
$\hat{\mathbf{f}}_i$: normalized embedding
$\mathcal{P}$: kNN neighbor pairs
$w_{ij} = e^{-\alpha \|\mathbf{x}_i - \mathbf{x}_j\|^2}$: distance weight
$S, \hat{S}$: GT / predicted semantic map

Semantic-aware Pruning

1. **Compute each Gaussian's contribution to the scene using its semantic features[2].**
$S_i = w_{\text{vis}}\, s_i^{\text{vis}} + w_{\text{sem}}\, s_i^{\text{sem}}$
2. **prune low importance.**

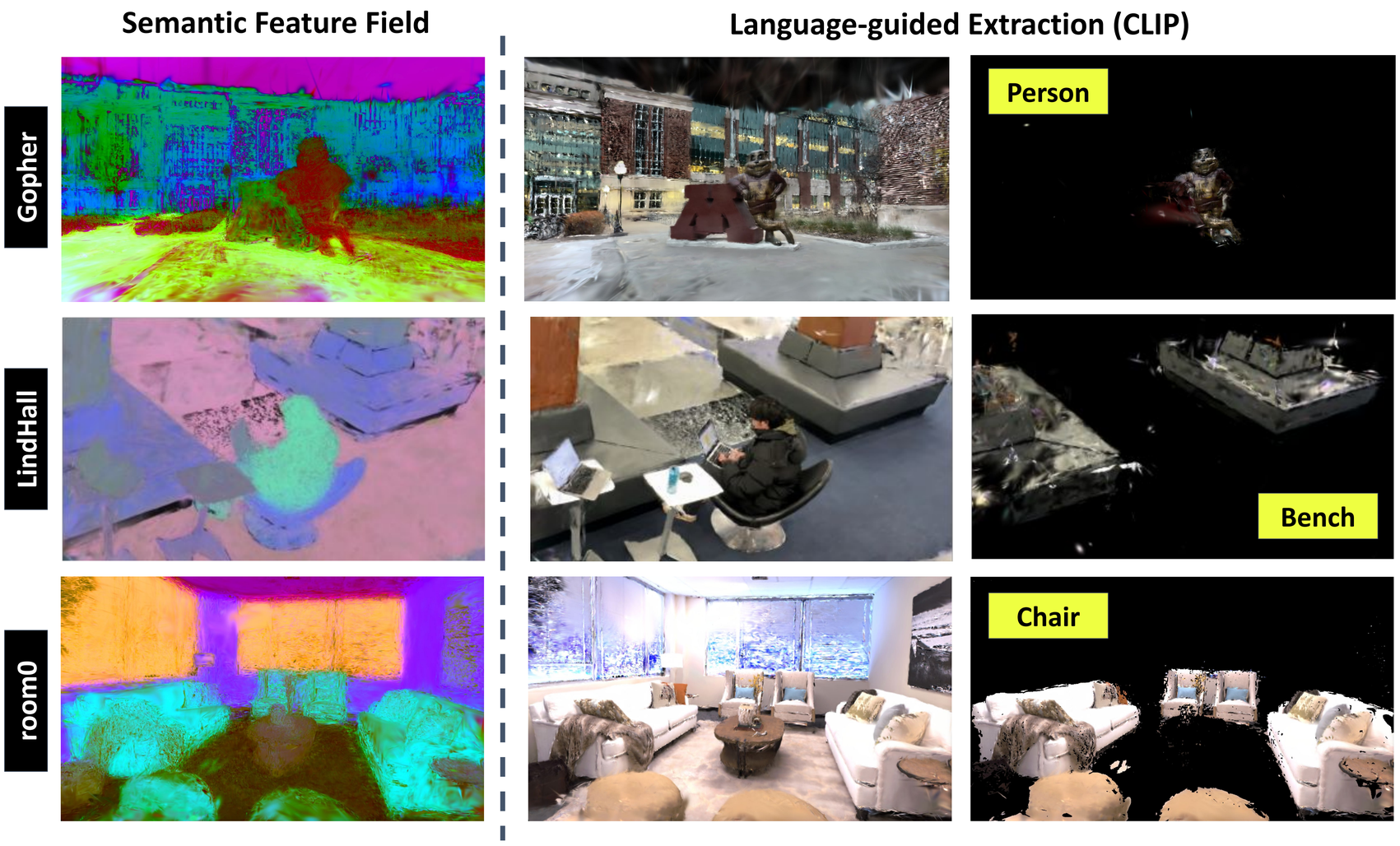


Quantitative Results

mIoU (↑)	office3	office4	room0	room1	Gopher	LindHall
Baseline [4]	**0.838**	0.702	0.845	**0.799**	0.724	**0.531**
Ours	0.831	**0.752**	**0.846**	0.771	**0.748**	0.500

FPS(↑) / # GS(↓)	office3	↓office4	room0	room1	Gopher	LindHall
Baseline [4]	67.63 / 241K	65.97 / 258K	27.09 / 596K	47.10 / 642K	26.13 / 2.4M	33.00 / 889K
Ours	**84.12 / 81K**	**78.79 / 85K**	**32.08 / 203K**	**61.74 / 210K**	**57.54 / 807K**	**56.93 / 303K**

Qualitative Results



computer science & engineering
MINNESOTA ROBOTICS INSTITUTE
UNIVERSITY OF MINNESOTA