Visual Computing & Artificial Intelligence
Prof. Matthias Nießner

Semantic Novel View Synthesis with 3D Gaussians

Gábor Markó, Daniel Stjepanovic, Supervisor: Dávid Rozenberszki

Advanced Deep Learning for Computer Vision

End-to-end hybrid Gaussian- and voxel-based method with 2D to 3D segmentation lifting

Dense Gaussian Representation

Advantages

- ✅ Real-time rasterization on modern GPUs
- ✅ Compact form for geometric scene representation
  → Enables reconstruction with low hardware cost even for large scale scenes

0. Starting point

- Multi-view RGB images
- Camera intrinsics and extrinsics
- ScanNet++ dataset

Sparse Voxel Representation

Advantages

- ✅ Compact grid structure for efficient data storage
- ✅ Significantly fewer voxels than Gaussians
  → allows storing high-dimensional feature vectors
  → enables efficient calculations

1. Geometric Scene Reconstruction

Method: 3D Gaussian Splatting
Output: scene reconstruction with ~1-2 million Gaussians used as geometry prior

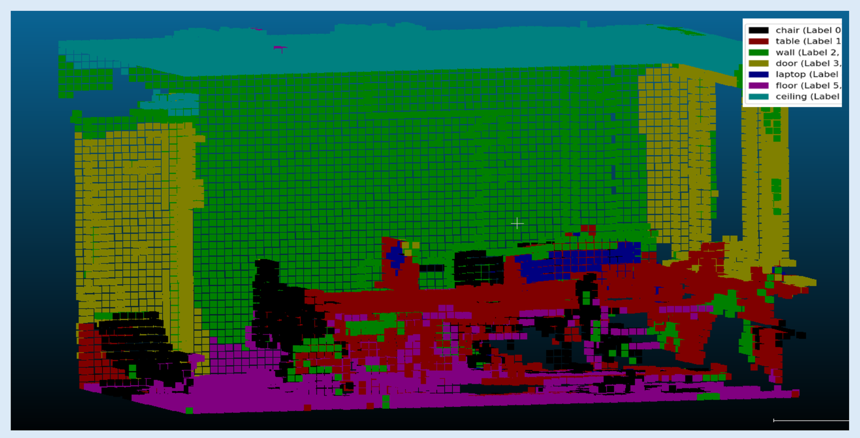
2. Sparse Voxel Grid Initialization

Method: opacity, local density and scale-based filtering of Gaussian centers
Output: ~10-100k initialized voxels around the geometry of the scene, saved as a sparse voxel grid

3. 2D Per-Pixel Semantic Map

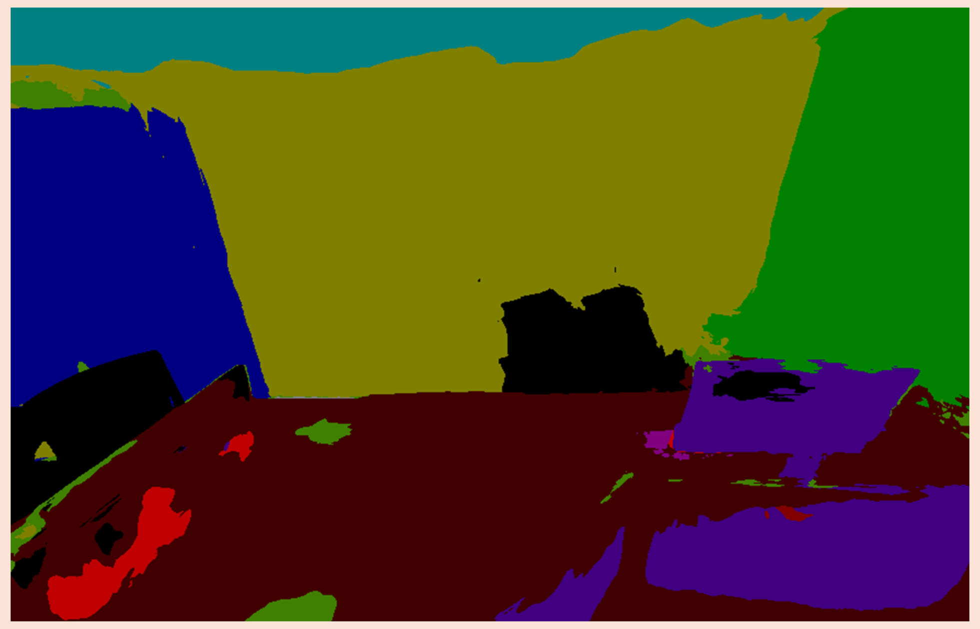
Method: 2D-semantic segmentation with LSeg
Output: 512-D raw per pixel feature embeddings

4. Semantic Feature Map projection



Method: Lifting 2D segmentation results by projecting per-pixel feature embeddings from 2D to 3D sparse voxels using efficient ray tracing
Output: per-voxel semantic vectors

5. Semantic Rasterization



Method:

1. Open vocabulary query of the text space embeddings stored in the voxels
2. Associating Gaussians with voxels and replacing RGB+SH with the label logits
3. Alpha blending over the logits, argmax to get per-pixel semantic label in 2D

Output: Gaussians with semantics logits, enabling semantic novel view synthesis

7. Evaluation: 2D semantic metrics

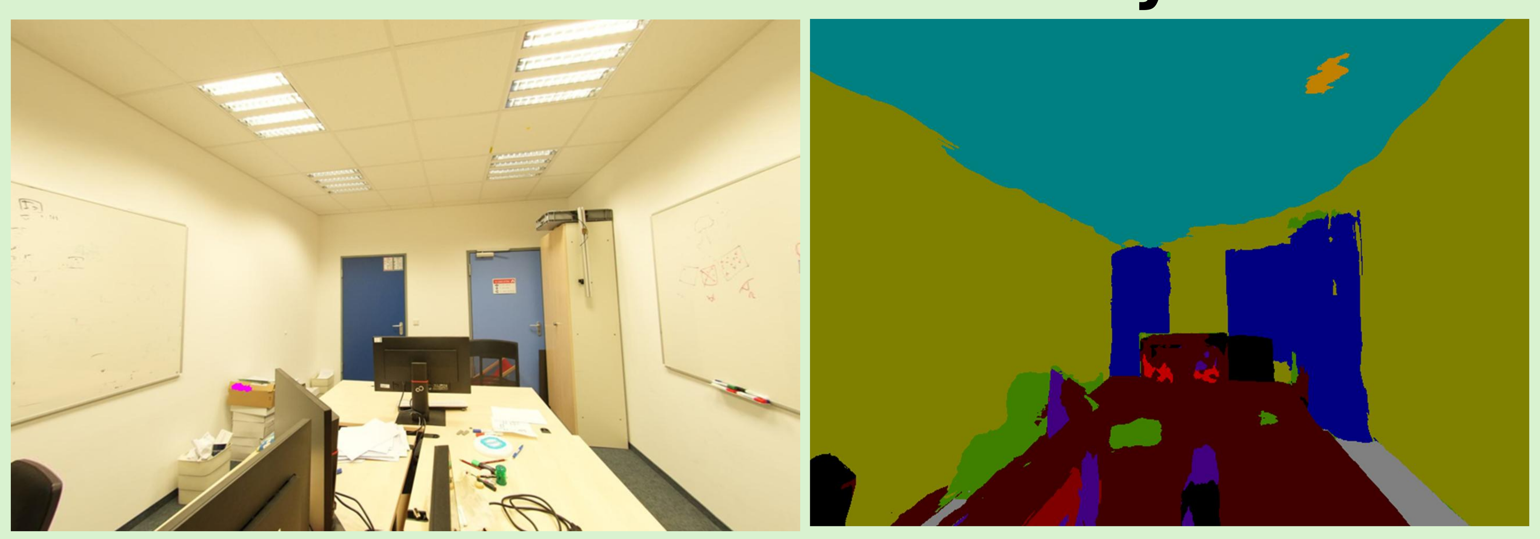
mIoU	0.391
fwIoU	0.494

mIoU, fwIoU metrics, compared to the rendered GT label values of ScanNet++, over the given query inputs

Large structures (e.g. walls, ceilings) → decent segmentation

Small/fine objects → less robust performance, as expected from the steps including resolution decrease (e.g. sparse voxel grid, downsampling to avoid OOM)

6. Result: Semantic Novel View Synthesis



8. Future work:

- Improving sparse voxel grid initialization around the geometry, smaller cell size
- Incorporating depth values of Gaussians at the rasterization step
- Optimization of the semantic logits stored in the Gaussians by supervision with ground truth 2D semantic maps
- More comprehensive evaluation