

# EmbodiedOcc: Embodied 3D Occupancy Prediction for Vision-based Online Scene Understanding

## Supplementary Material

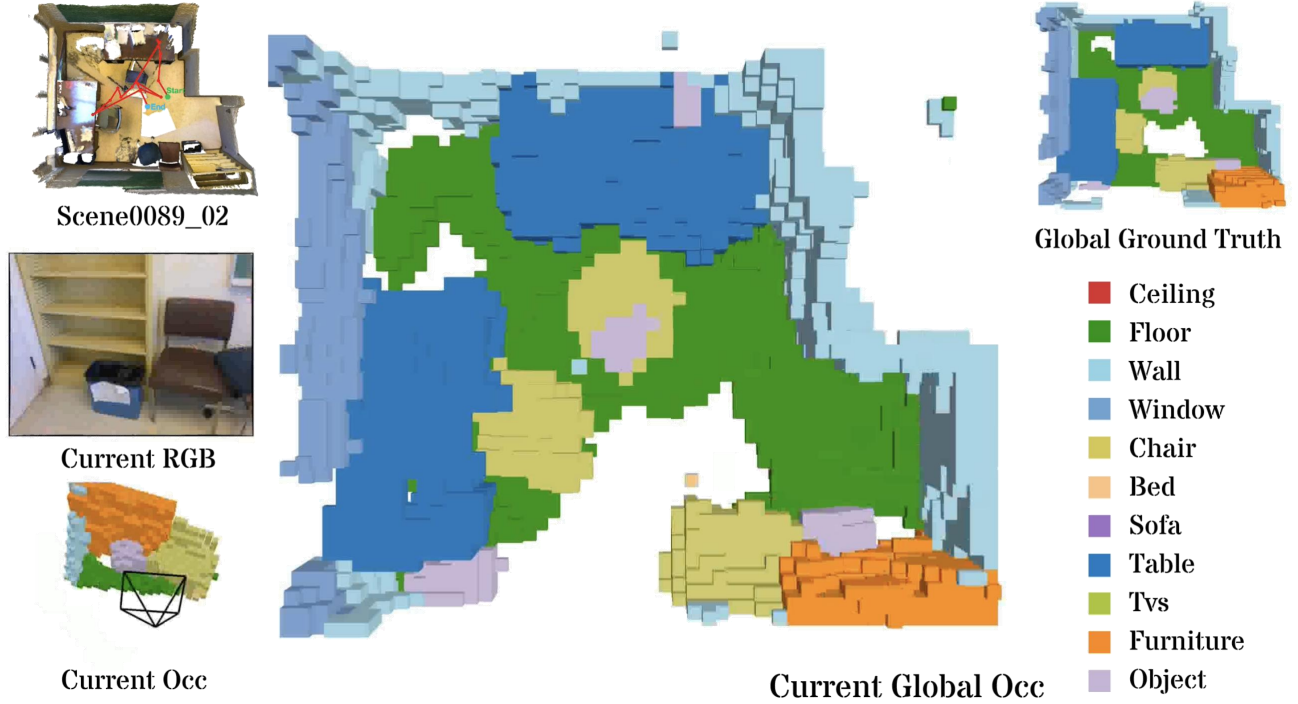


Figure 1. **Visualizations of the proposed EmbodiedOcc for Embodied 3D Occupancy Prediction on the EmbodiedOcc-ScanNet.** We visualize the current monocular RGB input and local occupancy prediction given by our EmbodiedOcc in the bottom left corner, and the global occupancy ground truth of the current scene in the top right corner. The global occupancy for the current scene given by our EmbodiedOcc is right in the center.

### A. EmbodiedOcc-ScanNet Dataset Details

We reorganize our EmbodiedOcc-ScanNet dataset following the data formulation used in NYUv2 [2] and Occ-ScanNet [4]. We noted that the Occ-ScanNet dataset consists of frames sampled from the original ScanNet [1] dataset randomly, which means that different frames may come from the same indoor scene. However, during the training and evaluation of our EmbodiedOcc framework, we have to ensure that scenes in the training set are different from those in the evaluating set. So we split the scenes again and obtained our EmbodiedOcc-ScanNet dataset, which comprises 537/137 scenes in the train/val splits.

For each scene, we first obtain a global occupancy of it from the voxel labels in the CompleteScanNet [3] dataset using the K-Nearest Neighbors algorithm. Next, we count and resample the frames of this scene in the Occ-ScanNet dataset using a certain interval to obtain 30 posed images. For each frame, we select a specific area in front of the cam-

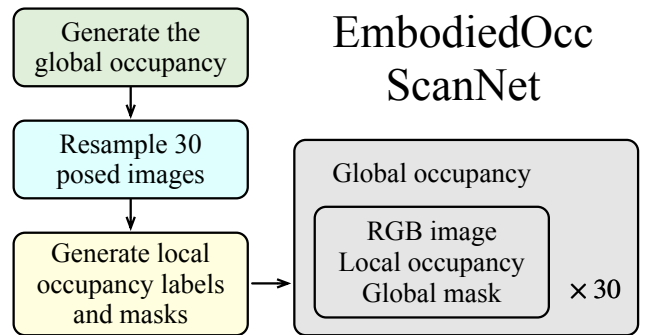


Figure 2. **Pipeline of our EmbodiedOcc-ScanNet.**

era as the range of local occupancy. The selection of the local voxel origin is consistent with the Occ-ScanNet [4]. Then, we obtain the current local occupancy from the global occupancy using the K-Nearest Neighbors algorithm. In addition to this, we maintain a mask in global resolutions for each frame, which marks the intersection of the current local voxel and frustum. This allows us to obtain the occu-

pancy ground truth of the explored area by splicing together the masks of processed frames, enhancing the flexibility of our EmbodiedOcc-ScanNet. The pipeline to generate one scene in our EmbodiedOcc-ScanNet is shown in Figure 2.

## B. Additional Analysis

**Analysis of Loss Functions.** We train our model using a composite loss consisting of the focal loss, the lovasz-softmax loss and the scene-class affinity loss. We use Table 1 to analyze the effect of different loss functions.

Table 1. Ablation of different losses on the mini set.

Method	Local Prediction		Embodied Prediction	
	IoU	mIoU	IoU	mIoU
EmbodiedOcc w/o L (focal)	50.00	37.83	49.15	35.23
EmbodiedOcc w/o L (lov)	46.61	37.30	49.57	40.28
EmbodiedOcc w/o L (geo)	<b>54.62</b>	45.88	47.62	40.37
EmbodiedOcc w/o L (sem)	50.83	44.58	48.00	35.26
EmbodiedOcc	53.93	<b>46.20</b>	<b>50.78</b>	<b>41.45</b>

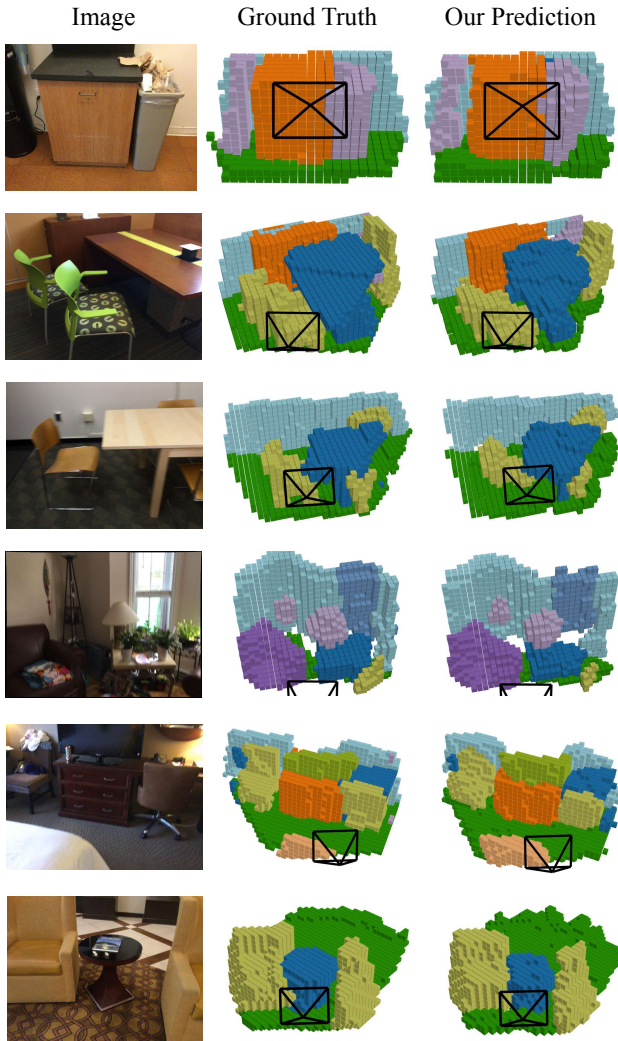


Figure 3. Additional visualizations of our local refinement module.

**Analysis of Robustness to Pose Errors.** For each frame, we use the pose from the original dataset (the poses which result from a global optimization using all frames of the current scene). However, during the real-world scenarios, the embodied agent can only obtain camera poses estimated from previous frames and the current frame, which inevitably introduces some errors. To demonstrate the robustness of our model to camera pose errors, we estimate poses based on previous frame (this brings a translation error of 0.10 m and a rotation error of  $2^\circ$ ), and use this pose to evaluate our model. This results in 39.03 (mIoU) and 47.57 (IoU), which is still better than the baselines.

## C. Additional Visualizations

Due to space limitations, we only selected a few frames in the main text to demonstrate the performance of our local refinement module. In Figure 3, we use a more diverse set of monocular samples to further showcase the visual effects of the local occupancy obtained by our local module.

To fully demonstrate the working process of our EmbodiedOcc, we use a video demo to showcase the performance of EmbodiedOcc when exploring indoor scenes. Figure 1 shows a sampled image from the video demo for embodied 3D occupancy prediction on the EmbodiedOcc-ScanNet. The video demo and our implementation code are both included in the zip folder.

## References

- [1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1
- [2] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, pages 746–760, 2012. 1
- [3] Shun-Cheng Wu, Kesuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 801–810, 2020. 1
- [4] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. *arXiv preprint arXiv:2407.11730*, 2024. 1