TGSFormer: Scalable Temporal Gaussian Splatting for Embodied Semantic Scene Completion

Supplementary Material

1. Experimental Setup

1.1. Dataset

Occ-ScanNet. Occ-ScanNet [16] is a large-scale indoor monocular semantic occupancy dataset containing 45,755 training frames and 19,764 validation frames. Each sample is annotated with 12 semantic categories, including free space and eleven occupied classes: ceiling, floor, wall, window, chair, bed, sofa, table, television, furniture, and generic objects. The ground-truth occupancy is provided as a voxel grid covering a 4.8m × 4.8m × 2.88m region in front of the camera, discretized into a 60 × 60 × 36 resolution. This dataset is used as the benchmark for our local monocula occupancy prediction experiments. We also report results on Occ-ScanNet-mini, a reduced subset containing 5,504 training frames and 2,376 validation frames.

EmbodiedOcc-ScanNet. EmbodiedOcc-ScanNet [15] provides an embodied variant of ScanNet, consisting of 537 training scenes and 137 validation scenes. Each scene contains a short exploration sequence of 30 posed RGB images together with corresponding volumetric occupancy ground truth. The global occupancy for each scene is generated by voxelizing the entire traversed region in world coordinates, using the same voxel size and semantic label set as the local task. In the embodied setting, the model receives sequential observations and updates a global scene estimate conditioned on the known camera poses.

1.2. Temporal-Occ-ScanNet Variant.

As Occ-ScanNet lacks sequential exploration, we additionally construct a temporalized variant of the dataset, termed Temporal-Occ-ScanNet for clarity. Beyond enabling controlled evaluation of our two-stage training strategy, this set also offers a lightweight benchmark that still contains sufficient temporal complexity for analyzing temporal fusion behaviours. Concretely, we reorganize all frames by scene, divide each scene into fixed-length **batches** that preserve intra-scene continuity, and enqueue the batches in sequential order. Frames within the same batch therefore form short yet meaningful temporal windows.

To ensure that each batch contains sufficient temporal context, we apply a scene-consistent padding strategy: if a batch contains fewer frames than the target temporal length, we automatically pad it using neighboring frames from the same scene, prioritizing earlier frames, followed by later ones, and finally repeating existing frames when necessary. This guarantees smooth temporal transitions even for scenes with limited frame counts.

During training, batches are shuffled at the group level rather than at the frame level, preserving temporal coherence within each batch while still exposing the model to diverse scene orders. Each data sample thus provides a compact sequence of RGB images, depth maps, intrinsic/extrinsic parameters, and voxel occupancies aligned in time, enabling fair and stable temporal supervision on Occ-ScanNet without altering its original annotations.

1.3. Evaluation metrics

We evaluate semantic scene completion performance using Intersection-over-Union (IoU) and mean IoU (mIoU) over the 12 semantic categories.

For local occupancy prediction, we follow the evaluation protocol of ISO [16], computing IoU strictly within the current frame's camera frustum.

For embodied occupancy prediction, we adhere to the EmbodiedOcc [15] protocol. The evaluation is performed over the global voxel grid of each scene, considering only regions that are observed at least once throughout the 30-frame exploration sequence.

1.4. Implementation Details

In our framework, the image encoder employs a pretrained EfficientNet-B7 [9] as backbone, while the depth branch utilizes a frozen fine-tuned *Depth-Anything-V2* [15] model.

Stage 1: Monocular Pretraining. In the first stage, we train TGSFormer on monocular SSC to establish a strong and frame-agnostic perceptual prior. The Gaussian Lifter operates on a uniformly downsampled grid of 30×40 points, following SplatSSC [8]. For Confidence-aware Voxel Fusion (CAVF), we set the sharpness parameter p and the maximum entropy threshold $H_{\max}$ to 3.0. The loss weights λ_1 and λ_2 in the final objective $\mathcal{L}_{\text{total}}$ are set to 100 and 2, respectively. We use the AdamW optimizer [7] with a weight decay of 0.01, and apply a learning-rate multiplier of 0.1 to the image backbone. The learning rate follows a cosine schedule with a 1000-iteration warmup, reaching a peak value of 8×10^{-4}. The model is trained for 10 epochs on Occ-ScanNet and for 20 epochs on Occ-ScanNet-mini, using 4 NVIDIA RTX 3090 GPUs with a batch size of 2 per GPU (global batch size 8).

Stage 2: Embodied Fine-tuning. In the second stage, we adapt the pretrained model to the embodied setting. To preserve the frame-agnostic perceptual prior established in Stage 1, all components of TGSFormer are frozen, except

Table 1. **Experiment settings for different ablation studies and efficient analysis.** Experiments above the dashed line are included in our main manuscript, while those below the dashed line are newly introduced in this appendix.

Experiments	Experiment Settings			
	Training Dataset	Training Device	Max Learning Rate	Total Batch Size
Comparison of Training Strategy	Occ-ScanNet-mini	2 NVIDIA RTX 3090	6×10^{-4}	6
Ablation on CAVF	Temporal-Occ-ScanNet-mini	4 NVIDIA RTX 3090	4×10^{-4}	4
Ablation on Training Objective	Temporal-Occ-ScanNet-mini	4 NVIDIA RTX 3090	4×10^{-4}	4
Ablation on Gaussian Initialization	Occ-ScanNet-mini	2 NVIDIA RTX 3090	6×10^{-4}	6
Ablation on Temporal Encoder	Temporal-Occ-ScanNet-mini	4 NVIDIA RTX 3090	4×10^{-4}	4
Ablation on CCA modulation strategies	Occ-ScanNet-mini	2 NVIDIA RTX 3090	6×10^{-4}	6
Ablation on Uncertainty Estimation	Occ-ScanNet-mini	2 NVIDIA RTX 3090	6×10^{-4}	6
Ablation on Confidence-aware Loss	Occ-ScanNet-mini	2 NVIDIA RTX 3090	6×10^{-4}	6
Efficiency Analysis	EmbodiedOcc-ScanNet-mini	4 NVIDIA RTX 3090	4×10^{-4}	4

Table 2. **Ablation on Uncertainty Estimation.** The temperature parameter used in the normalization step is denoted as T. Among all variants, the power transform with $T = 0.2$ achieves the best performance. The results of our proposed setting are highlighted in light gray.

Uncertainty Transform	Normalize	Temperature	IoU↑	mIoU↑
sharp sigmoid	softmax	1.0	64.19	51.94
power transform	softmax	1.0	64.20	52.04
power transform	softmax	0.2	**64.32**	**52.10**
power transform	softmax	0.5	64.23	51.99

for the Dual Temporal Encoder (DTE), which is exclusively responsible for temporal fusion. During finetuning, we apply a learning-rate multiplier of 0.1 to the DTE parameters, while the remainder of the network remains fixed. This selective optimization enables the model to learn stable cross-frame interactions and temporal consistency without perturbing the underlying single-frame representation. Stage 2 is trained for 5 epochs on EmbodiedOcc-ScanNet using 4 NVIDIA RTX 3090 GPUs, with a batch size of 1 per GPU (global batch size 4).

Further experimental settings. The configurations used in our ablation studies and efficiency analyses are summarized in Tab. 1. Each experiment follows the same training and inference protocol as its corresponding main result, with changes applied only to the component being examined. All experiments are conducted on a single NVIDIA RTX 3090 GPU, and the inference dataset is identical to the training dataset unless otherwise specified.

2. Further Experiment Results

2.1. Uncertainty Estimation

We further study the effect of different uncertainty-to-confidence mappings used in CAVF, as summarized in Tab. 2. Among all variants, the power transform with a temperature of 0.2 achieves the highest IoU and mIoU, indicating that a stronger contrast in confidence weighting leads to more reliable voxel-level fusion. In comparison, the sharp sigmoid baseline is:

$$c = \sigma\big(-\beta\,(H-\gamma)\big), \tag{1}$$

where H denotes the entropy and (β, γ) are set to $(10.0, 1.5)$ in our experiments. Utilizing the sharp sigmoid leads to inferior performance on both IoU and mIoU compared to our power transform, since its steep nonlinearity pushes confidence toward near-binary values, making the fusion less stable. In contrast, our power transform offers smoother scaling, with a lower temperature gives marginally more reliable weights, yielding the best result.

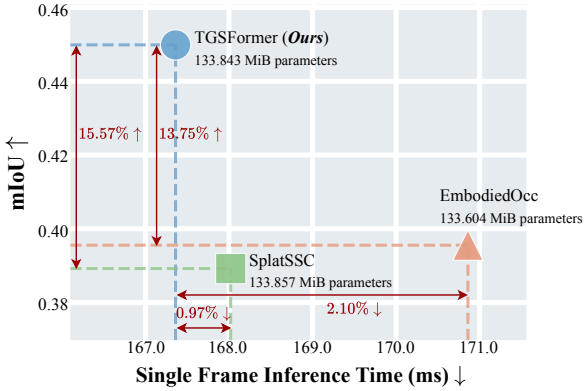

Figure 1. **Single-frame inference latency versus mIoU in Embodied Sequence**. All models share comparable parameter counts. EmbodiedOcc incurs the highest latency due to accumulating Gaussian features and memory entries. TGSFormer achieves both the lowest latency and the highest mIoU, benefiting from its lightweight Gaussian Lifter, DTE, and CAVF modules.

2.2. Efficiency Analysis

Runtime Efficiency. We evaluate the single-frame inference latency of TGSFormer, SplatSSC, and EmbodiedOcc in Fig. 1. All three models share comparable parameter counts, ensuring a fair comparison. EmbodiedOcc exhibits the highest latency due to the continual growth of its Gaussian features and memory entries, which increases splatting and aggregation cost. SplatSSC achieves faster infer-

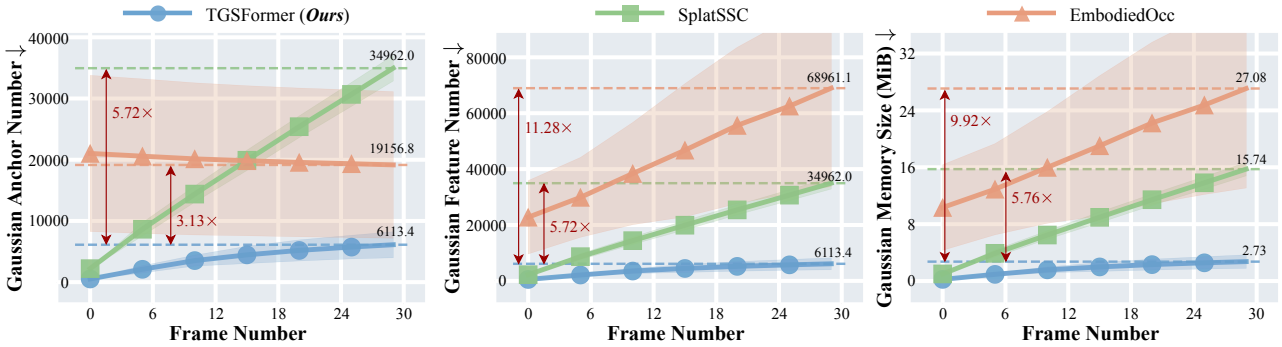

Figure 2. **Efficiency comparison in Embodied Sequence.** EmbodiedOcc shows unbounded growth in Gaussian features and memory, while SplatSSC exhibits steady accumulation due to the lack of temporal regulation. TGSFormer maintains a compact and bounded Gaussian representation through CAVF, resulting in up to 11.28× and 9.92× reductions in feature count and memory size, respectively.

ence but remains slightly slower than TGSFormer, consistent with its more complex multi-branch feature fusion design and the gradual accumulation of Gaussian features over long sequences.

Despite incorporating both temporal fusion and memory regulation, TGSFormer attains the *lowest* single-frame latency while simultaneously achieving the *highest* mIoU. This efficiency stems from two design choices: (1) a simple depth-guided Gaussian Lifter without heavy geometric modules, and (2) lightweight Dual Temporal Encoder (DTE) and CAVF modules that introduce negligible computational overhead. Moreover, CAVF actively reduces the number of active Gaussians during the update process, further lowering per-frame splatting and rendering cost.

Gaussian Complexity and Memory Growth. We further analyze the evolution of Gaussian anchors, features, and memory consumption throughout the embodied sequence in Fig. 2. EmbodiedOcc initializes a moderate number of Gaussian anchors but lacks a mechanism to constrain the associated feature representations. Consequently, while the anchor count remains relatively stable across frames, its Gaussian features and memory usage grow rapidly due to unbounded accumulation. SplatSSC exhibits similar growth in anchors and features, as it performs frame-wise depth-guided lifting without temporal regulation.

In contrast, TGSFormer maintains substantially fewer Gaussian anchors, features, and memory entries throughout the sequence. The CAVF module merges overlapping Gaussians in a confidence-aware manner, producing a compact and bounded representation that converges within only a few frames. Compared to EmbodiedOCC and SplatSSC, TGSFormer achieves up to **5.72× fewer anchors**, **11.28× fewer features**, and **9.92× less memory**.

3. Discussions

3.1. Depth Sensitivity and Failure Modes

TGSFormer relies on depth-based lifting to initialize 3D Gaussians. This design assumes that the predicted depth

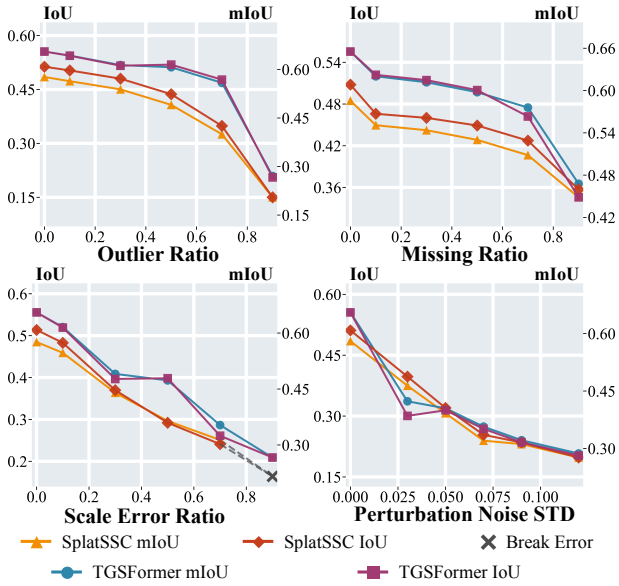

Figure 3. **Sensitivity to depth quality.** We evaluate the robustness of TGSFormer to depth priors under four noise types.

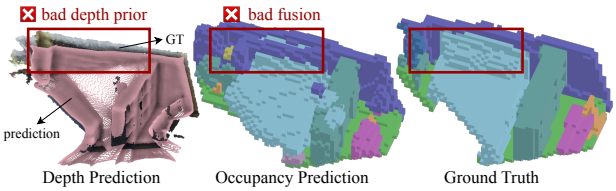

Figure 4. **Failure case on Temporal-Occ-ScanNet-mini.** Large depth errors under severe occlusion (GT in color vs. prediction in pink) lead to misaligned lifted Gaussians, causing inconsistent occupancy between the current frame and the historical estimate.

is metrically reasonable and locally consistent. When this assumption is violated, geometry initialization errors propagate through the system. We analyze robustness under four corruption types: outlier ratio, missing ratio, scale error ratio, and perturbation noise. As shown in Fig. 3, the model is tolerant to missing points and sparse outliers, since local fusion and temporal aggregation can compensate small irregularities. However, performance degrades under large

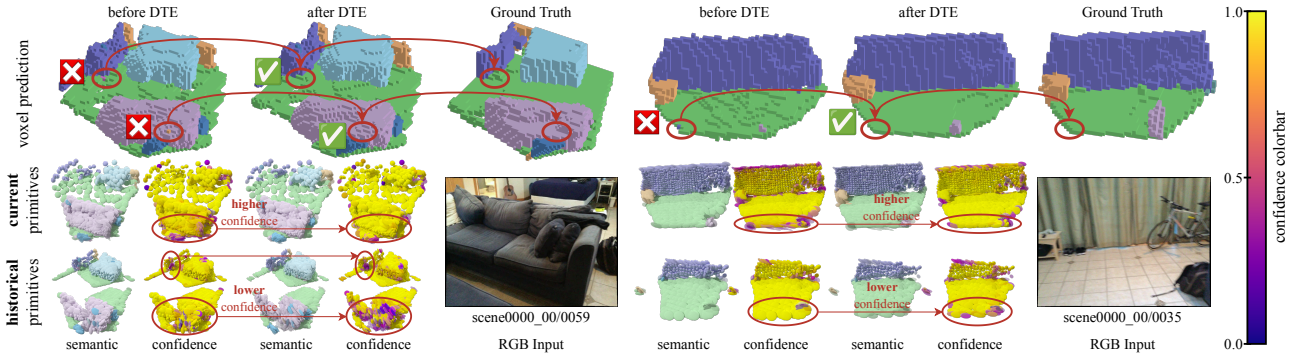

Figure 5. **Intuitive analysis of DTE.** DTE performs confidence-aware temporal fusion between current and historical primitives, selectively enhancing reliable semantic cues while suppressing ambiguous ones. As a result, voxel predictions become more structurally complete and semantically consistent after DTE.

Table 3. **Preliminary study on Confidence-aware Loss Strategies.** Each method adds an extra confidence dimension to Gaussians and renders it by confidence-weighted splatting. During training, the confidence C_i' is used to reweight the target loss with a weighting term parameterized by λ.

Reweighted Terms	Reweighting Formula	C_i' Mapping	IoU↑	mIoU↑
/	/	/	**63.13**	**52.89**
$\mathcal{L}_{\text{focal}}$	$C_i'\mathcal{L}_{\text{target}} + \lambda e^{-C_i'}$	$\exp(c_i')$	<u>62.19</u>	<u>52.12</u>
$\mathcal{L}_{\text{geo}}$	$C_i'\mathcal{L}_{\text{target}} + \lambda e^{-C_i'}$	$\exp(c_i')$	61.81	51.86
$\mathcal{L}_{\text{focal}}, \mathcal{L}_{\text{geo}}$	$C_i'\mathcal{L}_{\text{target}} + \lambda e^{-C_i'}$	$\exp(c_i')$	60.47	51.91
$\mathcal{L}_{\text{focal}}, \mathcal{L}_{\text{geo}}$	$C_i'\mathcal{L}_{\text{target}} - \lambda \log(C_i')$	$1 + \exp(c_i')$	58.95	51.22

perturbation noise or global scale errors. This degradation reveals a structural limitation: the Dual Temporal Encoder (DTE) performs feature refinement under the assumption of geometric consistency, and the Confidence-aware Voxel Fusion (CAVF) merges Gaussians only within local voxel neighborhoods. Neither module is designed to correct large metric-scale misplacements. A representative failure case is shown in Fig. 4. When severe occlusion induces depth scale bias, the lifted Gaussians (pink) deviate significantly from the true geometry (RGB), and the misalignment persists through temporal fusion, leading to incorrect global occupancy estimation.

3.2. Intuitive Analysis of the DTE

To provide a more intuitive understanding of the Dual Temporal Encoder (DTE), we visualize the evolution of Gaussian primitives in Fig. 5. The top row illustrates how DTE resolves semantic ambiguity during temporal fusion. Instead of directly aggregating features from current and historical frames, DTE performs cross-attention between the current Gaussian set and the historical Gaussian set. In this process, the current observations act as queries that selectively retrieve compatible information from historical primitives. This attention mechanism implicitly serves as a confidence-aware filtering process. Historical primitives that are geometrically or semantically consistent with the current observation receive higher attention weights, while outdated or inconsistent primitives are automatically suppressed. Consequently, valid foreground structures in the current frame are reinforced, while ambiguous or conflicting historical features are down-weighted. As shown in Fig. 5, this selective interaction produces cleaner semantic predictions and more complete geometric structures compared to direct feature aggregation.

3.3. Future Works

Uncertainty Quantification. Our current confidence estimation is primarily based on semantic entropy and does not explicitly account for geometric uncertainty arising from depth errors or lifting drift. A more principled treatment of model and data uncertainty may benefit both temporal alignment and voxel fusion. Classical approaches such as MC dropout [3], deep ensembles [6], and MIMO [4] offer generic mechanisms for estimating epistemic and aleatoric uncertainty, and represent natural directions for extending our framework. Recent SSC works [1, 5] incorporate uncertainty into 3D occupancy reasoning but are typically built upon SurroundOcc-style pipelines [14], which do not directly align with our Gaussian-memory representation. EmbodiedOcc++ [10] adopts MC dropout for semantic uncertainty, yet still faces challenges in balancing efficiency and reconstruction performance.

Another paradigm [11, 13] utilize confidence as an auxiliary prediction to weight the loss, rather than as a directly supervised input to fusion. Inspired by this, we added an additional confidence attribute $c_i' \in [0, 1]$ to each Gaussian primitive and rendered it by confidence-weighted splatting:

$$\hat{C}(\mathbf{x}) = \frac{\sum_{i \in \mathcal{N}(\mathbf{x})} p(\mathbf{x} \mid G_i)\, c_i'}{\sum_{j \in \mathcal{N}(\mathbf{x})} p(\mathbf{x} \mid G_j)}. \tag{2}$$

Nevertheless, directly using this confidence to reweight the training loss (Tab. 3) led to a slight drop in performance. In particular, the formulation $C_i'\mathcal{L}_{\text{target}} - \lambda \log(C_i')$ with $C_i' = 1 + \exp(c_i')$ follows the confidence-aware loss design of Dust3R [13], while the alternative $C_i'\mathcal{L}_{\text{target}} + \lambda e^{-C_i'}$ with

$C_i' = \exp(c_i')$ is adopted to avoid negative loss values. Both variants underperform in our setting, indicating that loss reweighting schemes to Gaussian–based occupancy prediction are non-trivial and require further adaptation.

Designing unified semantic and geometric uncertainty estimators that integrate effectively with Gaussian memory remains an important direction for future work.

RNN-style Scene Reconstruction. As shown in Fig. 2, the number of Gaussian primitives in TGSFormer still grows over time in embodied settings, albeit much more slowly and non-linearly compared to existing baselines. This suggests that maintaining long-term memory consistency remains a challenge under extended exploration. Recent works [2, 12] demonstrate the feasibility of recurrent or state-space formulations for sustaining global 3D consistency over long sequences. Exploring RNN-style mechanisms for managing Gaussian memory and mitigating long-horizon drift represents a natural extension of this work.

References

[1] Anh-Quan Cao, Angela Dai, and Raoul de Charette. Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14554–14564. CVPR, 2024. 4

[2] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 5

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 4

[4] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *Proceedings of the Ninth International Conference on Learning Representations*. ICLR, 2021. 4

[5] Severin Heidrich, Till Beemelmanns, Alexey Nekrasov, Bastian Leibe, and Lutz Eckstein. OCCUQ: exploring efficient uncertainty quantification for 3d occupancy prediction. In *IEEE International Conference on Robotics and Automation*, pages 1–8. ICRA, 2025. 4

[6] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*. NeurIPS, 2017. 4

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations*. ICLR, 2019. 1

[8] Rui Qian, Haozhi Cao, Tianchen Deng, Shenghai Yuan, and Lihua Xie. Splatssc: Decoupled depth-guided gaussian splatting for semantic scene completion. *arXiv preprint arXiv:2508.02261*, 2025. 1

[9] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[10] Hao Wang, Xiaobao Wei, Xiaoan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler. In *Proceedings of the 33rd ACM International Conference on Multimedia*. MM, 2025. 4

[11] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306. CVPR, 2025. 4

[12] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10510–10522. CVPR, 2025. 5

[13] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709. CVPR, 2024. 4

[14] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21672–21683. ICCV, 2023. 4

[15] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. ICCV, 2025. 1

[16] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *Proceedings of the European Conference on Computer Vision*, pages 38–54. ECCV, 2024. 1