PANOPTIC3D: LEVERAGING 3D PSEUDO SUPERVISION FOR PANOPTIC OCCUPANCY PREDICTION

Dian Jia[⋆]*, Pei Yu*[†]*, Xiaoqian Ruan*[⋆]*, Hyeonjeong Park*[⋆]*, Wei Tang*[⋆]

[⋆]University of Illinois Chicago, Chicago, IL, USA [†]Microsoft, Redmond, WA, USA

{djia7, xruan9, hpark233, tangw}@uic.edu, pei.yu@microsoft.com

ABSTRACT

Panoptic occupancy prediction enables holistic 3D scene understanding by unifying semantic and instance-level reconstruction. However, current methods rely on expensive, often unobtainable 3D panoptic annotations. While weakly supervised approaches exist for semantic occupancy, extending them to panoptic settings remains challenging due to the difficulty of inferring 3D instance boundaries. To address this, we propose Panoptic3D, a novel framework for panoptic occupancy prediction without 3D ground truth. Our core contribution is a novel pseudo-label generation method that lifts 2D panoptic segmentations into 3D space, providing simultaneous semantic and instance-level supervision. Specifically, we localize instance centers by fitting oriented 3D bounding boxes to lifted object contours via principal direction identification. To mitigate noise from depth inaccuracies and occlusions, we incorporate semantic and geometric filtering, temporal regularization, and iterative refinement. Extensive experiments on the Occ3D-nuScenes benchmark demonstrate the effectiveness of Panoptic3D.

***Index Terms*—** Panoptic occupancy prediction, 3D scene understanding, weakly supervised learning

1. INTRODUCTION

Understanding the geometry and semantics of 3D scenes is a fundamental problem in computer vision [1], playing a crucial role in autonomous driving and robotics. Recent advances in monocular semantic occupancy prediction [2] offer cost-effective per-voxel reconstruction. However, these works focus on category-level semantics and cannot separate individual object instances, which is essential for motion forecasting and obstacle avoidance. Panoptic occupancy prediction [3] addresses this by unifying "stuff" regions and "thing" instances. Despite its potential, current approaches rely on full supervision, requiring costly 3D panoptic annotations that are difficult to scale. Other research using Gaussian Splatting [4] shows promise for semantic occupancy, but struggles in panoptic settings: it prioritizes surface fidelity but lacks the explicit volumetric structure required for consistent 3D instance grouping and holistic panoptic reasoning.

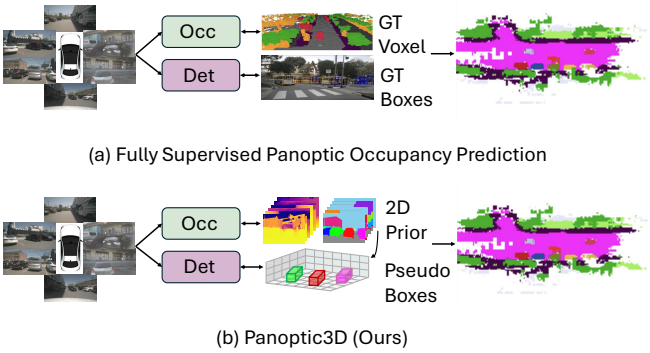

Fig. 1. Comparison of (a) fully supervised panoptic occupancy and (b) our Panoptic3D, which constructs 3D pseudo-supervision from 2D priors without 3D ground truth.

To close this gap, we propose Panoptic3D, which, to the best of our knowledge, is the first method to learn panoptic occupancy prediction without any ground truth 3D panoptic annotations. Our approach introduces two major technical innovations:

First, we propose a panoptic 3D pseudo-label generation method. We leverage vision foundation models [5] to lift 2D panoptic segmentation into 3D volumetric space. The core challenge lies in localizing 3D instance centers from monocular views, as direct projection of 2D centers yields surface points that deviate from true object centroids. We address this by fitting oriented 3D bounding boxes to lifted object contours through identifying their principal directions. By analyzing boundary pixels corresponding to silhouette tangencies, our method captures the object's lateral extent to produce accurate 3D center estimates.

Second, we develop a robust learning framework to handle inherent noise in pseudo labels, such as segmentation uncertainty and depth inaccuracies. We design semantic and geometric filters to exclude unreliable voxels and incorporate a temporal consistency loss to regularize predictions across adjacent frames. Finally, we leverage learned knowledge to refine pseudo labels and perform re-training to progressively sharpen instance boundaries and increase spatial coverage.

Our contributions are summarized as follows: (1) We present Panoptic3D, which to our knowledge is the first work to learn panoptic occupancy prediction without 3D ground truth, providing a data-efficient approach for 3D scene understanding (Fig. 1). (2) We propose a pseudo-label generation method that accurately locates 3D instance centers by fitting oriented boxes to lifted contours via principal direction identification. (3) We develop a robust framework incorporating semantic and geometric filtering, temporal regularization, and iterative refinement to mitigate pseudo-supervision noise. (4) Extensive experiments on the Occ3D-nuScenes benchmark demonstrate the effectiveness of Panoptic3D.

2. RELATED WORK

2.1. Panoptic Occupancy Prediction

Panoptic occupancy prediction unifies 3D reasoning by jointly predicting per-voxel semantic categories and consistent instance identities, extending the capabilities of purely semantic occupancy [6]. This task requires not only accurate category assignment but also spatially coherent instance grouping across dense volumetric representations. Compared to semantic-only methods, it faces challenges in separating adjacent objects, resolving complex occlusions, and maintaining viewpoint consistency. A series of recent camera-based baselines have established strong results for this task. PanoOcc [7] introduces voxel queries that aggregate multi-frame features in a coarse-to-fine manner for long-range context. SparseOcc [3] leverages sparse 3D convolutions and mask transformers to improve instance delineation. Panoptic-FlashOcc [8] adopts a bottom-up design where a lightweight 2D head predicts centerness maps for efficient voxel grouping. While effective, these methods rely on exhaustive 3D panoptic annotations that are prohibitively expensive to collect. In contrast, our work constructs 3D panoptic pseudo-supervision from 2D priors, enabling panoptic occupancy learning without any human-provided 3D labels.

2.2. Weakly-supervised Occupancy Prediction

Weakly supervised occupancy reduces label reliance using 2D masks, depth, or photometric consistency. Many methods leverage NeRF-based differentiable rendering to optimize semantic-density fields [9], often incorporating LiDAR projections [10] or foundation models [5] to handle limited viewpoints. To reduce computational overhead, recent works explore Gaussian Splatting [4]; for instance, GaussianOcc [4] uses 3D primitives to accelerate training. However, Gaussian Splatting prioritizes surface fidelity but lacks the explicit volumetric structure required for consistent instance grouping, limiting its capacity for holistic 3D panoptic reasoning. While prior works focus on semantic or surface occupancy, we address volumetric panoptic occupancy without 3D labels using 2D-derived pseudo-supervision.

3. METHOD

Panoptic occupancy prediction unifies semantic and instance-level occupancy for holistic 3D scene understanding, which is essential for tasks that require reasoning about individual entities. To bypass the requirement for expensive 3D panoptic annotations, we propose a framework to learn this task without 3D ground truth. Specifically, we describe our network architecture in Sec. 3.1, panoptic 3D pseudo-label generation in Sec. 3.2, and robust learning from noisy 3D supervision in Sec. 3.3.

3.1. Network Architecture

Our network builds upon Panoptic-FlashOcc [8], which predicts 3D instance centers with per-voxel semantic occupancy and reconstructs panoptic labels via voxel-space clustering. We adopt it for its effectiveness and efficiency. As shown in Fig. 2, the network consists of three components. (1) *BEV Feature Generation:* Multi-view images are transformed into BEV features $\boldsymbol{F}^{\text{bev}} \in \mathbb{R}^{C\times H\times W}$ (C and $H \times W$ represent the feature dimension and spatial resolution, respectively) using a ResNet-50 backbone, an LSS view transformer, and a ResNet-18 FPN encoder [11]. (2) *Semantic and Instance Heads:* The semantic head predicts voxel-wise probabilities $\boldsymbol{O}^{\text{sem}} \in [0,1]^{Z\times H\times W}$ across Z vertical voxels. The instance head outputs class-aware heatmaps $\boldsymbol{O}^{\text{heat}} \in [0,1]^{K\times H\times W}$ for K "thing" categories, and a 3D offset field $\boldsymbol{O}^{\text{off}} \in \mathbb{R}^{3\times H\times W}$ for instance localization. (3) *Panoptic Occupancy Assembly:* Semantic and instance predictions are fused by assigning each "thing" voxel to the nearest class-consistent center. These centers are extracted from $\boldsymbol{O}^{\text{heat}}$ via local-maximum detection and regressed offsets from $\boldsymbol{O}^{\text{off}}$, while "stuff" classes share a single instance ID.

3.2. Panoptic 3D Pseudo Label Generation

2D Priors from Foundation Models. We adopt *Grounded-SAM* [5] to generate the panoptic segmentation maps. To resolve redundant overlaps and unclear boundaries in direct mask fusion, we apply: (1) *Class-aware NMS* to remove duplicate bounding boxes; (2) *Score-guided pixel assignment* to let high-confidence masks dominate. This yields stable, order-independent per-view 2D panoptic maps. Finally, we estimate metric depth using *Metric3D V2* [12] for geometrically consistent mapping.

Semantic Occupancy Construction. We construct 3D semantic voxels through three steps:

(1) Semantic Point Cloud Generation. For each view i, a pixel (u, v) is lifted to the ego coordinate system as:

$$\boldsymbol{p}_{i,u,v} = \boldsymbol{T}_i \begin{bmatrix} \boldsymbol{D}_i(u,v)\, \boldsymbol{K}_i^{-1}[u,v,1]^\top \\ 1 \end{bmatrix}, \qquad (1)$$

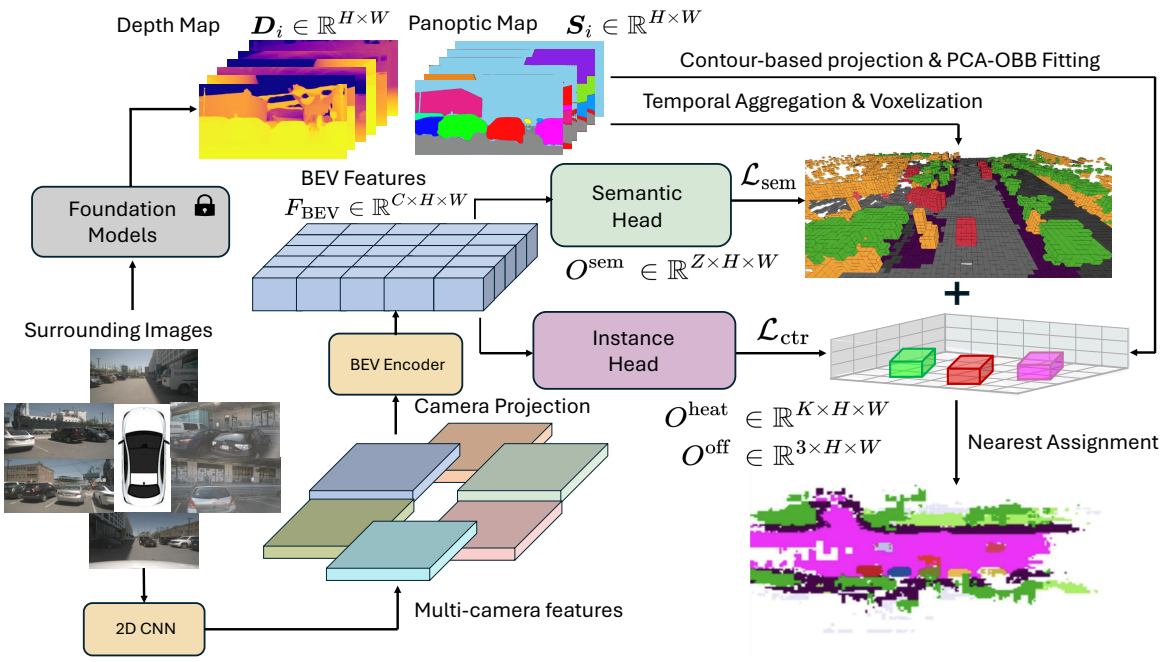

Fig. 2. Overall framework of Panoptic3D. We learn panoptic occupancy without 3D ground truth. The pipeline includes: (1) Architecture: BEV generation, semantic and instance heads, and voxel assembly (Sec. 3.1); (2) Supervision: 2D-to-3D pseudo-label training (Sec. 3.2); (3) Learning: Noise-robust optimization via temporal and refinement losses (Sec. 3.3).

where $\boldsymbol{D}_i$ and $\boldsymbol{K}_i$ are the estimated depth and camera intrinsics, and $\boldsymbol{T}_i$ transforms the i-th camera to the ego view. Each $\boldsymbol{p}_{i,u,v}$ is associated with its semantic label $\boldsymbol{S}_i(u,v)$, forming per-view clouds $\boldsymbol{P}_i$. Aggregating all M views yields the global semantic point cloud: $\boldsymbol{P} = \bigcup_{i=1}^{M} \boldsymbol{P}_i$.

(2) Temporal Aggregation. To densify the representation, we aggregate point clouds across frames $\mathcal{S}_t$:

$$\hat{\boldsymbol{P}}_t = \bigcup_{t\in\mathcal{S}_t} \boldsymbol{T}_t\, \boldsymbol{P}_t. \tag{2}$$

We adopt different strategies: for "thing" classes, we fuse the current and preceding frame to balance density and motion robustness; for static "stuff" classes, points from all frames are aggregated to maximize spatial completeness.

(3) Voxelization. We discretize $\hat{\boldsymbol{P}}_t$ into voxels. A voxel is activated if it receives sufficient multi-view or temporal support by exceeding a certain threshold, with its class decided by majority voting among contributing points.

3D Instance Center Localization. To provide 3D supervision for the instance head, we generate pseudo instance labels from 2D priors. Directly projecting 2D centers into 3D yields surface-biased points. We overcome this by fitting an oriented bounding box (OBB) to the lifted instance using Principal Component Analysis (PCA).

For a set of N lifted points $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$, we compute the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then solve the eigendecomposition: $\boldsymbol{\Sigma E} = \boldsymbol{E\Lambda}$, where $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3]$ are orthonormal eigenvectors. These define the principal axes of the point distribution. Each point is transformed into the PCA frame as:

$$\boldsymbol{\xi}_n = \boldsymbol{E}^T(\boldsymbol{x}_n - \boldsymbol{\mu}). \tag{3}$$

We calculate the 3D box bounds in the PCA frame: $\boldsymbol{\xi}^{\min} = \min_n \boldsymbol{\xi}_n$ and $\boldsymbol{\xi}^{\max} = \max_n \boldsymbol{\xi}_n$. Finally, the 3D box center $\boldsymbol{c}$ is transformed back to the ego system:

$$\boldsymbol{c} = \boldsymbol{\mu} + \boldsymbol{E}\frac{1}{2}(\boldsymbol{\xi}^{\min} + \boldsymbol{\xi}^{\max}). \tag{4}$$

Mask versus Contour. Using full instance masks for $\mathcal{X}$ causes biased estimation as points collapse on the front facet. In contrast, we use the mask contour (boundary pixels). Contour pixels correspond to silhouette tangencies along viewing rays. When lifted, they distribute around the object's lateral hull, yielding a centroid that accurately approximates the true 3D volumetric center (Fig. 3).

3.3. Robust Learning from Noisy 3D Supervision

Noisy Label Filtering. To mitigate noise from depth inaccuracies and occlusions, we apply a filtering mechanism to assign an "ignore" flag to unreliable voxels. It consists of: (1) *Semantic filter*: We calculate the class probability distribution $\boldsymbol{y}_v$ and retain voxels with confidence $\max_c y_{v,c} > 0.8$; (2) *Geometric filter*: We remove voxels without at least two occupied neighbors in a $3\times3\times3$ neighborhood. This eliminates

Table 1. Comparison of 3D panoptic occupancy on Occ3D-nuScenes [13]. † denotes upper-bound variants; ‡ SparseOcc uses 8-frame input.

Method	Backbone	Input Size	Epochs	RayPQ	$\text{RayPQ}_{1\text{m}}$	$\text{RayPQ}_{2\text{m}}$	$\text{RayPQ}_{4\text{m}}$
SparseOcc (Full GT) ‡ [3]	R50	704×256	24	14.1	10.2	14.5	17.6
Panoptic-FlashOcc (Full GT) [8]	R50	704×256	24	13.2	9.2	13.5	16.8
Panoptic-FlashOcc† (Voxel GT only)	R50	704×256	24	11.0	9.0	11.4	12.7
Panoptic-FlashOcc† (Box GT only)	R50	704×256	24	9.2	7.1	9.6	10.9
BaseModel A	R50	704×256	24	4.8	3.5	4.8	6.2
BaseModel B	R50	704×256	24	6.2	4.5	6.3	7.9
Panoptic3D (ours)	R50	704×256	24	8.6	6.7	8.9	10.2

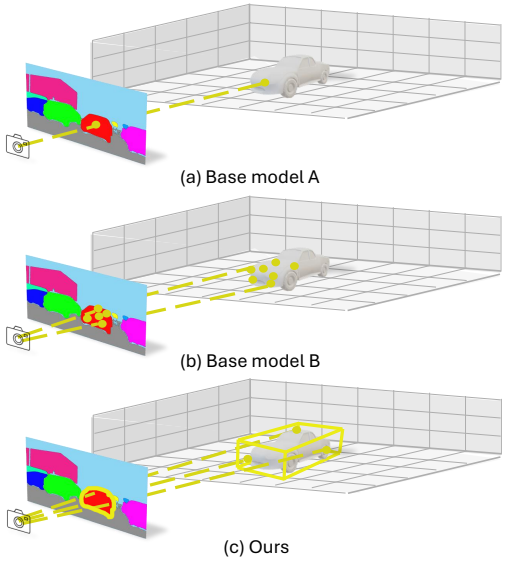

Fig. 3. Comparison of pseudo-center generation strategies. (a) 2D center back-projection and (b) mask-based centroid are surface-biased. (c) Our approach adopts contour-based projection followed by PCA-aligned OBB fitting to obtain a consistent 3D volumetric center approximation.

floating fragments and yields a coherent semantic volume for supervision.

Learning Objective. Following [6], we optimize a semantic occupancy loss $\mathcal{L}^{\text{sem}}$ (including focal loss, semantic/geometry affinity [2], and Lovász-Softmax [14]) and a center loss $\mathcal{L}^{\text{ctr}}$ that supervises center heatmap and 3D offset regression with a smooth L_1 objective for instance grouping on valid voxels $\mathcal{V}^{\text{valid}}$. To enhance stability under noisy labels, we introduce a temporal consistency loss:

$$\mathcal{L}^{\text{temp}} = \frac{1}{|\mathcal{V}^{\text{valid}}|} \sum_{v \in \mathcal{V}^{\text{valid}}} \left\| \widehat{\boldsymbol{y}}_{t,v} - \widehat{\boldsymbol{y}}_{t-1\to t,v} \right\|_1, \quad (5)$$

where $\widehat{\boldsymbol{y}}_{t-1\to t,v}$ is the previous frame's prediction warped to the current ego-coordinates.

Refinement. We leverage *propagated knowledge* [15] by using high-confidence model predictions to refine pseudo labels and re-training the network. This process suppresses noise, sharpens boundaries, and fills missing regions for final prediction.

4. EXPERIMENTS

4.1. Dataset and Metrics

We evaluate our method on the Occ3D-nuScenes dataset [13], a large-scale benchmark for autonomous driving. It contains 700 training and 150 validation sequences, annotated at 2 Hz. The 3D occupancy volume covers $[-40\text{m}, 40\text{m}]$ for x, y axes and $[-1\text{m}, 5.4\text{m}]$ for the z axis, with a voxel resolution of 0.4m. We adopt Ray Panoptic Quality (RayPQ) [3] as the primary metric to evaluate panoptic occupancy, which accounts for depth inconsistency by assessing coherence along viewing rays. We also report mIoU for semantic occupancy evaluation.

4.2. Implementation Details

Our model is trained on 4 NVIDIA A40 GPUs with a batch size of 32 for 24 epochs. We employ a ResNet-50 backbone and resize input images to 704×256. We use the AdamW with a weight decay of 1×10^{-2} and a base learning rate of 1×10^{-4} following a linear warm-up. All loss weights are set to 1.0. Inference speed is measured on an NVIDIA A100 GPU using the PyTorch `fp32` backend to ensure fair comparison.

4.3. Panoptic Occupancy Evaluation

To the best of our knowledge, Panoptic3D is the first framework to perform panoptic occupancy prediction without any 3D annotations. We compare our approach with the fully supervised Panoptic-FlashOcc [8] (the Oracle upper bound) and two specifically designed baselines to validate our pseudo-center generation.

Comparison with Baselines. As shown in Tab. 1, we analyze three levels of pseudo-supervision: (1) *Baseline A:* Directly back-projects 2D instance centers into 3D. This model performs poorly (4.8 RayPQ) due to severe depth ambiguity and surface bias. (2) *Baseline B:* Computes the geometric centroid of all projected pixels. While improving

Table 2. Semantic occupancy results on Occ3D-nuScenes [13].

Method	Training Labels	mIoU	FPS
SelfOcc (CVPR'24) [17]	OpenSeeD	10.54	1.2
OccNeRF (TIP'25) [16]	GroundedSAM	10.81	1.3
GaussianOcc (ICCV'25) [4]	GroundedSAM	11.26	5.6
GaussTR (CVPR'25) [18]	Foundation Models	13.26	0.2
Panoptic3D (Ours)	Foundation Models	**14.64**	**35.1**

to 6.2 RayPQ, it still lacks volumetric structural awareness. (3) *Ours:* By incorporating contour-based projection and PCA-OBB fitting, our method achieves 8.6 RayPQ. These results demonstrate that our pseudo-supervision narrows the performance gap to 3D-supervised methods. Notably, our weakly-supervised approach achieves 8.6 RayPQ, reaching near-parity with models trained on 3D bounding box annotations (e.g., 9.2 RayPQ).

4.4. Semantic Occupancy Evaluation

We also evaluate the semantic occupancy performance as it provides the foundation for instance grouping. As shown in Tab. 2, our method achieves 14.64 mIoU, outperforming existing weak-supervision methods such as OccNeRF [16] (10.81) and SelfOcc [17] (10.54). Notably, our framework maintains a high inference speed of 35.1 FPS. It is important to emphasize that the heavy foundation models are only utilized offline for pseudo label generation; during inference, the network operates independently without these priors. This ensures our approach is not only effective but also significantly more efficient than recent Gaussian-based or Transformer-based alternatives.

Table 3. Ablation on pseudo-center construction (Chamfer Distance ↓).

Method	Chamfer Dist.
Base-A (2D center → 3D)	1.861
Base-B (Full mask → centroid)	1.688
w/o OBB (Contour → centroid)	1.624
Ours (Contour → PCA-OBB)	**1.412**

4.5. Ablation Study

Effectiveness of PCA-OBB Fitting. In Tab. 3, we use Chamfer Distance to measure the localization error of pseudo-centers. Simple back-projection (Base-A) and pixel averaging (Base-B) lead to high errors. Moving from full masks to contours reduces the error to 1.624, while our full PCA-OBB scheme achieves the lowest error of 1.412. This proves that PCA-based alignment effectively captures the volumetric center of objects even from partial 2D observations.

Learning Framework Components. We ablate the contribution of temporal consistency and pseudo-label refinement in Tab. 4. Temporal consistency improves stability under noisy labels (5.53 → 6.04 RayPQ). The refinement process brings the largest gain (+2.61 RayPQ). The joint model achieves 8.60 RayPQ and 14.64 mIoU, which is the best overall performance among all settings. This shows that temporal regularization and pseudo-label refinement together form a more robust learning framework.

Table 4. Ablation study on learning framework components.

Refinement	Temp-Consist	RayPQ	mIoU
		5.53	10.91
	✓	6.04	12.82
✓		8.14	14.25
✓	✓	**8.60**	**14.64**

Table 5. Effect of instance score threshold τ on panoptic occupancy.

τ	0.1	0.2	0.3	0.4	0.5
RayPQ	8.13	8.36	**8.60**	8.37	8.35

Table 6. Ablation on pseudo label refinement threshold.

Setting	RayPQ	mIoU
w/o refinement	6.04	12.82
Refinement ($\tau = 0.7$)	8.21	14.48
Refinement ($\tau = 0.8$)	**8.60**	**14.64**
Refinement ($\tau = 0.9$)	7.94	13.71

Hyper-parameter Analysis. We investigate the instance score threshold τ (Tab. 5) and the refinement confidence threshold (Tab. 6). For instance assembly, $\tau = 0.3$ serves as the optimal point to balance center recall and precision. For refinement, a threshold of 0.8 ensures that only the most reliable model predictions are used to update the pseudo-labels, preventing error propagation.

5. CONCLUSION

We presented Panoptic3D, a weakly-supervised framework for panoptic occupancy prediction without 3D annotations. By lifting 2D priors through contour-based PCA-OBB fitting and robust refinement, our method generates reliable 3D pseudo supervision and achieves competitive performance on Occ3D-nuScenes. Our results show that effective panoptic occupancy learning is possible using only 2D-derived supervision, reducing the reliance on costly 3D annotations.

Acknowledgements. This work was supported in part by National Science Foundation (NSF) grants ECCS-2400900 and IIS-2442540, the National Artificial Intelligence Research Resource (NAIRR) Pilot, Amazon Web Services (AWS) provided through CloudBank, and NCSA Delta GPU resources provided through ACCESS.

6. REFERENCES

[1] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1746–1754.

[2] Anh-Quan Cao and Raoul De Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3991–4001.

[3] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma, "Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15035–15044.

[4] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya, "Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 28980–28990.

[5] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[6] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.

[7] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang, "Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 17158–17168.

[8] Zichen Yu, Changyong Shu, Qianpu Sun, Yifan Bian, Xiaobao Wei, Jiangyong Yu, Zongdai Liu, Dawei Yang, Hui Li, and Yan Chen, "Panoptic-flashocc: An efficient baseline to marry semantic occupancy with panoptic via instance center," *arXiv preprint arXiv:2406.10527*, 2024.

[9] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[10] Peizheng Li, Shuxiao Ding, You Zhou, Qingwen Zhang, Onat Inak, Larissa Triess, Niklas Hanselmann, Marius Cordts, and Andreas Zell, "Ago: Adaptive grounding for open world 3d occupancy prediction," *arXiv preprint arXiv:2504.10117*, 2025.

[11] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[12] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[13] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64318–64330, 2023.

[14] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17850–17859.

[15] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10687–10698.

[16] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu, "Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields," *CoRR*, 2023.

[17] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu, "Selfocc: Self-supervised vision-based 3d occupancy prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19946–19956.

[18] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang, "Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11960–11970.