YouTube-Occ: Learning Indoor 3D Semantic Occupancy Prediction from YouTube Videos

Haoming Chen[1], Lichen Yuan[1], Tianfang Sun[1], Jingyu Gong[1,2,3†], Zhizhong Zhang[1,3], Xin Tan[1,2], Yanyun Qu[4], and Yuan Xie[1,2†]

[1] School of Computer Science and Technology, East China Normal University, China
[2] Chongqing Key Laboratory of Precision Optics, Chongqing Institute of East China Normal University, China
[3] Shanghai Key Laboratory of Computer Software Evaluating and Testing, China
[4] School of Informatics, Xiamen University, China

Abstract. 3D semantic occupancy prediction is crucial for fine-grained scene understanding, yet its advancement in privacy-sensitive indoor environments is fundamentally hindered by the scarcity of large-scale annotated 3D data. To overcome this limitation, we explore learning indoor 3D semantic occupancy prediction from abundant, uncalibrated in-the-wild internet videos while simultaneously bypassing the extensive manual annotation. Specifically, we introduce *YouTube-Occ*, including an automated data pipeline that leverages 2D and 3D foundation models to process raw web videos, estimating camera geometry, reconstructing scene point clouds, and enriching them with dense semantic pseudo-labels. However, these plausible pseudo-labels fail to yield performance gains under naive supervision. To address this impasse, we further propose a pre-training framework driven by feature distillation with a dual-alignment strategy. Within it, an intra-frame alignment utilizes a voxel-anchored Gaussianization module to align 3D features with corresponding 2D priors, whereas a cross-scene alignment achieves global semantic consistency via class-prototype distillation. Empirically, YouTube-Occ delivers consistent gains across three mainstream architectures on the NYUv2 and Occ-ScanNet benchmarks, especially under limited-data conditions. We will publicly release our code and data, hoping to inspire future research.

Keywords: Indoor 3D Semantic Occupancy · Pre-Training · In-the-Wild Videos

1 Introduction

As one of the most promising abilities of artificial intelligence, 3D scene understanding is becoming increasingly important for applications such as autonomous driving [2] and robotic navigation [37]. Just as humans possess a natural ability to comprehend 3D environments through vision, 3D scene perception plays a fundamental role in understanding surroundings for machines. Among them,

† Corresponding author

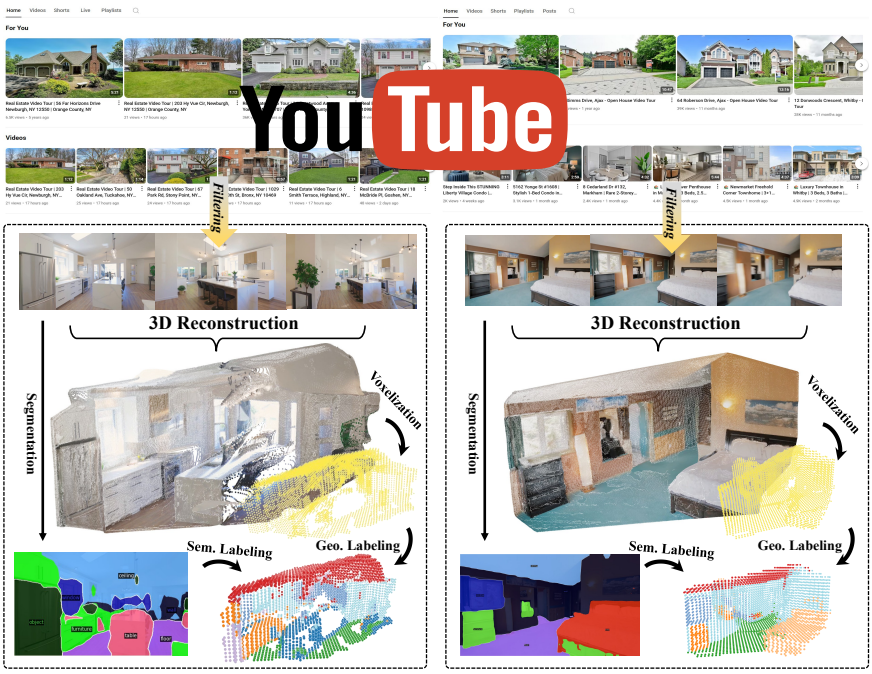

Fig. 1: YouTube-Occ could acquire diverse and rich indoor videos from the YouTube platform and support 3D semantic occupancy training without pre-configured camera.

semantic occupancy prediction [3, 18] discretizes 3D space into voxel grids with semantic labels, which provide a fine-grained and unified geometric representation, leading to a detailed perception of object shapes and scene-level structures.

However, the advancement of this field, especially for privacy-sensitive indoor environments, is fundamentally bottlenecked by the scarcity of large-scale and annotated 3D data. Indoor scenes present formidable challenges due to intricate geometries, severe occlusions, and wide variations in object scales [55]. Existing indoor semantic occupancy benchmarks are limited in scale and diversity (*e.g.*, NYUv2 [38] is only 1,449 images). Moreover, the manual collection and labeling for indoor scene is labor-intensive and unscalable. The high cost of specialized 3D scanning equipment, compounded by significant privacy concerns, also creates a critical barrier. To bypass these barriers, we ask a compelling question: *Can we learn 3D indoor semantic occupancy from abundant, unlabeled, and uncalibrated in-the-wild internet videos?*

To answer this question, we propose **YouTube-Occ**, which incorporates an automatic data pipeline capable of producing geometric and semantic pseudo-labels from in-the-wild internet videos. As illustrated in Fig. 1, our pipeline curates video clips and leverages modern 3D foundation models to jointly estimate camera geometry and reconstruct scene point clouds. These point clouds are then enriched with dense semantic labels derived from 2D foundation models,

culminating in a rich data source for pre-training existing 3D semantic occupancy models.

While utilizing foundation models to generate pseudo-labels seems intuitive, our empirical analysis indicates that direct supervision with pseudo-labels yields unsatisfactory results. Inspired by advancements in outdoor image-to-LiDAR distillation [5, 34, 51, 63], we design a pre-training framework tailored for both in-domain and cross-domain fine-tuning. Crucially, this framework empowers existing semantic occupancy models to effectively harness unconstrained internet videos. Specifically, the core of our framework is a dual alignment strategy driven by two complementary contrastive objectives. First, an intra-frame feature alignment mechanism utilizes a voxel-anchored Gaussianization module to render 3D features, matching 2D priors from the corresponding view. Second, a cross-scene feature alignment module enforces global semantic consistency by aligning representations with dynamically updated class prototypes. Extensive evaluations on NYUv2 and Occ-ScanNet show consistent performance gains across three semantic occupancy architectures. Moreover, leveraging internet videos could enhance model performance in data-scarce scenarios.

- To address data scarcity and annotation bottlenecks in indoor semantic occupancy prediction, we propose an automated data pipeline that, to our knowledge, is the first to use unlabeled, uncalibrated in-the-wild videos for 3D semantic occupancy prediction.
- We introduce a pre-training framework compatible with mainstream semantic occupancy architectures. Using voxel-anchored Gaussian rendering and class-prototype distillation, it learns generalizable representations in both intra- and cross-domain settings.
- Extensive experiments demonstrate that our approach reliably boosts performance on three popular indoor semantic occupancy architectures, with particularly substantial improvements in few-shot scenarios.

2 Related Work

2.1 3D Semantic Occupancy

3D semantic occupancy prediction, which represents scenes as fine-grained semantic grids, has become a pivotal task for holistic 3D understanding. A significant body of prior work [3, 15, 17, 25–27, 30, 45, 52, 57] has focused on advancing performance through architectural innovations, such as sophisticated feature fusion and temporal modeling [6, 23, 53, 59], or by integrating high-level reasoning with large language models [44, 50]. While these data-driven methods have achieved impressive results, they are constrained by large-scale, manually annotated 3D datasets. The prohibitive cost and privacy concerns associated with acquiring such data severely bottleneck the development of these models. Our work deviates from this model-centric paradigm by proposing a new data-centric approach to mitigate this core limitation.

2.2 3D Self-Supervised Learning

3D self-supervised learning (SSL) offers a compelling paradigm to reduce reliance on expensive manual annotations. A common approach involves learning representations by enforcing consistency between a 3D model and its 2D renderings [13,16,56], often employing differentiable renderers like Neural Radiance Fields [31] or 3D Gaussian Splatting [20]. However, these methods are predominantly tailored for multi-view outdoor scenes and presume the availability of pre-calibrated sensor data. In contrast, our approach is specifically designed to tackle the unique challenges of cluttered and varied indoor environments by leveraging uncalibrated in-the-wild videos. Our pre-training framework is introduced to effectively learn from this challenging data, addressing the object complexity and data scarcity inherent to indoor scenes.

2.3 Web-data Learning

Web-data learning offers a scalable and cost-effective alternative to creating datasets that involve substantial financial and operational burdens. Many methods have designed sophisticated data-cleaning pipelines to harness this data for various tasks [8,10–12,21,29,32,35,36]. For instance, YouTube-VLN [28] demonstrated success by collecting large-scale indoor video frames from YouTube for Vision-and-Language Navigation. However, the application of web-data learning to 3D semantic occupancy prediction remains an unexplored area. To our knowledge, our work is the first to bridge this gap. We present a complete pipeline that automatically curates, reconstructs, and pseudo-labels unstructured internet videos to create a promising pre-training dataset for 3D occupancy models.

3 Automated Internet Data Curation

We introduce an automated data pipeline to harvest and process massive, uncalibrated indoor videos (*e.g.*, unknown camera information.) from the internet for 3D semantic occupancy prediction. First, we revisit the manual indoor occupancy annotation process in Sec. 3.1. Then, we detail our data curation pipeline, which leverages 2D/3D foundation models to generate pseudo-labels in Sec. 3.2. Finally, we discuss the challenges involved in using pseudo-labels in Sec. 3.3.

3.1 Revisiting Manual Indoor Occupancy Annotation

In this section, we briefly summarize how indoor semantic occupancy labels are generated in Occ-ScanNet [55]. Occ-ScanNet uses RGB images, depth maps, camera intrinsics, and camera poses from ScanNet [9], together with scene-level semantic voxel annotations from CompleteScanNet [46]. To obtain frame-level semantic occupancy labels, it defines a voxel origin from the current camera pose, specifying a 3D volume in front of the camera. This involves two main components: (1) **Scene Data Collection**: ScanNet [9] provides RGB-D video of indoor scenes captured by handheld depth sensors, with accurate depth calibration and server-side 6-DoF camera trajectories; (2) **Scene-Level Semantic**

Voxels: CompleteScanNet [46] combines 3D CAD models from ShapeNet [48] with CAD-to-scan alignments from Scan2CAD [1] to produce scene-level semantic voxel grids encoding complete object geometry.

To alleviate the labor-intensive and prohibitively expensive burden of manual data collection and annotation, we propose a scalable paradigm that harvests indoor videos from the online video platform (*e.g.,* YouTube). By leveraging off-the-shelf 2D and 3D foundation models, we automatically generate pseudo-labels at scale. These pseudo-labels will be used in our pre-training framework to enable efficient pre-training of different semantic occupancy models.

3.2 Automated Data Curation Pipeline

The data pipeline, as in Fig. 2, consists of video collection and filtering, geometry reconstruction, semantic segmentation, and semantic occupancy generation.

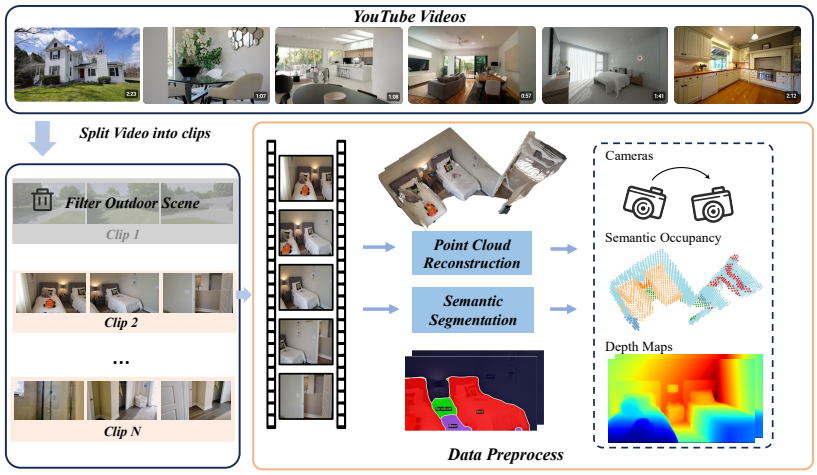

Fig. 2: The data pipeline for generating pseudo-labels from internet videos. Starting from YouTube videos, we first split them into clips and filter out the irrelevant ones. We then use a 3D foundation model to reconstruct the surface point clouds of scene objects, and a 2D foundation model to extract pixel-level semantic information. Finally, each frame is associated with pseudo-labels, including camera intrinsics, camera poses, depth maps, 2D semantic maps, and 3D semantic occupancy grids.

Video Collection. Inspired by [61], we choose the category of real estate footage as our video data source. In general, these real estate videos mainly contain indoor footage, which tend to have a smooth and steady camera motion. These shots mainly provide three major advantages for our data construction: 1) Easier 3D Reconstruction: Stable inter-frame motion helps reconstruction and reduces errors from motion blur and camera jitter. 2) Broader Scene Diversity: Internet videos span many environments, providing richer data that improves

3D perception model generalization. 3) Mostly Static Content: Beyond camera ego-motion, these videos are largely static, making them analogous to indoor benchmarks Occ-ScanNet. Specifically, we begin by collecting video IDs from real estate channels on the YouTube website and then use the video download tool yt-dlp to store these videos locally. Next, we split each video into separate clips following the procedure in [61]. To discard irrelevant segments, we utilize the off-the-shelf image classifier from [60] to filter out outdoor scenes.

Point Cloud Reconstruction. Since in-the-wild internet videos provide only uncalibrated RGB images without camera parameters, we cannot directly apply existing 3D occupancy models. To reconstruct the 3D geometry, we employ Dust3R [41] to jointly estimate camera intrinsics, poses, and point clouds. To address misalignments and scale ambiguity in the raw output, we apply a post-processing pipeline: we use heuristic filters to clean the data, transform the scene to a z-up orientation parallel to the floor, and enforce an approximate metric scale assuming a 2.8-meter wall height. Our design choices are driven by two considerations. (1) We adopt Dust3R [41] as it marks the genesis of the recent boom in 3D foundation models (*e.g.*, VGGT [40] and Pi3 [43]). (2) Since estimating exact metric scale from uncalibrated web data lacking any camera priors is an inherently ill-posed problem, the 2.8-meter architectural prior crucially normalizes diverse scenes into a consistent canonical metric space, ensuring stable occupancy voxelization at a fixed resolution of 0.08 m.

Semantic Segmentation. Following [5], we utilize SAN [49] for open-vocabulary semantic segmentation. The segmentation masks are obtained by directly inputting the textual label of each category from the indoor occupancy benchmarks (both NYUv2 [38] and Occ-ScanNet [55] contain 11 semantic classes) into the model. We deliberately eschew elaborate and transient prompt engineering to mitigate the risk of overfitting to specific formulations. This design choice guarantees a strictly automated pipeline, eliminating the need for task-specific human intervention while ensuring effortless scaling and future extensibility in complex open-world environments.

Occupancy Generation. Given the point cloud and pixel-level semantics, we assign each point both spatial and semantic attributes. We next voxelize the semantic point cloud using a fixed voxel size (*e.g.*, 0.08 m) to form semantic scene voxels, assigning each voxel the most common class label of its contained points. Finally, to derive frame semantic voxels, we specify a 3D bounding box over the front area and label each voxel within it as either occupied or empty.

3.3 Exploiting Pseudo-Labels

A naive way to use our auto-labeled internet videos is to directly train existing 3D semantic occupancy networks (*e.g.,* Symphonies [18]) with these pseudo-labels under standard supervision. However, empirical results (Sec. 5.2) show that direct zero-shot inference or joint training leads to marginal or even negative transfer on benchmarks like Occ-ScanNet. Based on qualitative visualizations and quantitative analyses, we conjecture the reason is two misalignments in pseudo-labels: (1) Geometric scale misalignment: Monocular internet videos

lack absolute metric scale, causing 3D foundation model predictions to be systematically mis-scaled relative to the real-world metric spaces of standard indoor datasets. (2) Semantic misalignment: 2D foundation model predictions are inconsistent with occupancy benchmark categories, lacking a strict one-to-one mapping to the target dataset taxonomy.

The inherent lack of geometric constraints and the high scene diversity in in-the-wild internet videos make direct supervision with unscaled pseudo-labels sub-optimal. To address this, we propose a pre-training framework that leverages such imperfect labels through voxel-anchored Gaussianization and class-prototype-based distillation. As detailed in Sec. 5.3, this approach yields consistent performance gains across three mainstream architectures and two semantic occupancy benchmarks, demonstrating its robustness to the noisy nature of web-sourced data.

4 Pre-training Framework

Our pre-training framework utilizes self-supervised distillation to enhance popular 3D semantic occupancy networks. First, we describe the background of a vision-centric occupancy prediction network in Sec. 4.1. Then, we show the details of our pre-training method in Sec. 4.2. Finally, we discuss the differences between our method and related outdoor self-supervised paradigms in Sec. 4.3.

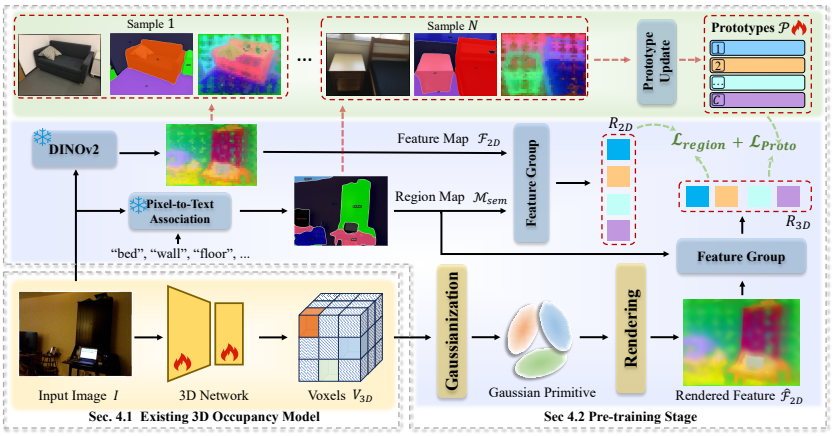

Fig. 3: Overview of our pre-training framework. It supports distilling foundation model knowledge into mainstream 3D occupancy networks. To bridge the 3D-to-2D modality gap, we propose a Voxel-Anchored Gaussianization module to project 3D voxel features onto the 2D image plane. We design two complementary alignment objectives: an intra-frame region loss ($\mathcal{L}_{region}$) for local semantic coherence and a cross-scene prototype loss ($\mathcal{L}_{proto}$) for global semantic consistency.

4.1 Revisiting Occupancy Prediction

Classically, a vision-centric semantic occupancy model takes a single RGB image $I \in \mathbb{R}^{M\times 3}$ with M pixels as input and predicts a 3D semantic occupancy volume $\hat{Y} \in \mathbb{R}^{N}$ of a scene, where N denotes the number of scene voxels. Each voxel y_i in $\hat{Y}$ is either empty or occupied with one semantic class. A 3D occupancy model typically consists of an occupancy network $\mathcal{N}$ and an occupancy head $\mathcal{H}$ as follows:

$$V_{3D} = \mathcal{N}(I), \hat{Y} = \mathcal{H}(V_{3D}), \tag{1}$$

where the network $\mathcal{N}$ generates 3D voxel features with D dimensions $V_{3D} \in \mathbb{R}^{N\times D}$ from 2D image input I, and the decoder $\mathcal{H}$ produces the semantic occupancy prediction $\hat{Y}$.

Existing methods train occupancy models with ground-truth occupancy semantics as direct supervision. We instead propose a self-supervised pre-training paradigm that learns powerful 3D voxel representations via distillation from vision foundation models, avoiding manual annotations. Our approach requires only minor architectural changes to the occupancy network: the sole difference between pre-training and fine-tuning is the dimensionality D of the feature V_{3D}. Thus, the framework is model-agnostic, and we evaluate it on three distinct occupancy networks in Sec. 5.

4.2 Pre-Training Stage

The goal of our pre-training framework is to learn dense 3D representations by distilling knowledge from powerful 2D foundation models. As illustrated in Fig. 3, our framework is built upon two complementary alignment objectives: an intra-frame feature alignment for local semantic coherence, and a cross-scene feature alignment for global semantic consistency.

Overview. Building on the occupancy network $\mathcal{N}$, we construct the 3D voxel representation V_{3D}. For an image I, we first obtain a semantic region map $\mathcal{M}_{sem}$ via our pixel-to-text association module. This map groups features from two heterogeneous streams: a 2D teacher and a 3D student. The 2D teacher stream extracts dense DINOv2 features $\mathcal{F}_{2D}$ and aggregates them with $\mathcal{M}_{sem}$ to form region-level embeddings R_{2D}. The 3D student stream projects 3D voxel features into a 2D feature map $\hat{\mathcal{F}}_{2D}$ (via voxel-anchored Gaussianization), then uses the same $\mathcal{M}_{sem}$ to obtain rendered region-level embeddings R_{3D}. Given R_{2D} and R_{3D}, we perform intra-frame alignment with a region contrastive loss $\mathcal{L}_{region}$, which aligns R_{3D} to the corresponding R_{2D} in each frame. To enforce cross-scene feature alignment, we propose a semantic prototype update module that utilizes $\mathcal{F}_{2D}$ and $\mathcal{M}_{sem}$ to update class prototypes $\mathcal{P}$, thereby achieving a prototype contrastive loss $\mathcal{L}_{proto}$ that pulls R_{3D} toward their corresponding semantic prototypes in $\mathcal{P}$.

Intra-Frame Feature Alignment This module aligns region-level embeddings between 3D and 2D modalities within a single frame, enhancing robustness against artifacts and boundary noise inherent in foundation models.

(a) 2D Teacher Feature Generation. The teacher stream produces distillation targets by integrating dense 2D features $\mathcal{F}_{2D}$ from DINOv2 with semantic masks $\mathcal{M}_{sem}$. As in Sec. 3.2, an open-vocabulary semantic model generates C binary masks based on pre-defined class prompts. The target region feature R^c_{2D} for each class c is then computed via masked average pooling over $\mathcal{F}_{2D}$.

(b) Voxel-Anchored Gaussianization. To bridge the modality gap, we transform the 3D voxel feature V_{3D} into 3D Gaussian primitives for differentiable rendering. Each Gaussian primitive is parameterized as follows: (1) *Position* is derived from the voxel coordinates; (2) *Features* are inherited from V_{3D}; (3) *Opacity* is predicted via an MLP; (4) *Scale and Rotation* are fixed to the voxel size and identity matrix, respectively. We then rasterize these primitives to produce a rendered 2D feature map $\hat{F}_{2D}$. Finally, the rendered region embedding R^c_{3D} is obtained by applying $\mathcal{M}_{sem}$ to $\hat{F}_{2D}$ through masked average pooling.

(c) Region Contrastive Loss. Given the R_{3D} and the R_{2D}, we derive the region-level loss $\mathcal{L}_{region}$:

$$\mathcal{L}_{region} = -\sum_{i=1}^{C} \log \frac{\exp(\langle f^i_{3D}, f^i_{2D} \rangle / \tau_{region})}{\sum_{j=1}^{C} \exp(\langle f^i_{3D}, f^j_{2D} \rangle / \tau_{region})}, \tag{2}$$

where f^i_{3D}/f^i_{2D} is the i^{th} region feature embedding from R_{3D}/R_{2D}, $\langle \cdot, \cdot \rangle$ denotes scalar product, and τ_{region} is a temperature hyper-parameter. This enables the 3D occupancy network to learn representations aligned with the spatial and semantic understanding of the powerful 2D foundation model.

Cross-Scene Feature Alignment. While intra-frame alignment ensures local coherence, a robust 3D representation must also exhibit global semantic consistency across diverse scenes. To achieve this, we introduce a cross-scene feature alignment mechanism based on semantic prototype contrastive learning. This module is depicted in the upper part of Fig. 3.

(a) Prototype Generation. We maintain a dynamic memory bank of class-specific prototypes $\mathcal{P} = \{p_1, p_2, ..., p_C\}$, where C is the number of semantic classes. Each prototype p_c is a feature vector that represents the centroid of a particular semantic class in the embedding space. This memory bank is updated online during training. For each training sample, we collect $\mathcal{F}_{2D}$ and $\mathcal{M}_{sem}$ produced during the 2D teacher feature generation step. These features are then used to update their corresponding class prototypes in the memory bank, typically via an exponential moving average with momentum β. This update rule allows the prototypes to evolve smoothly, capturing the canonical representation of each class from diverse samples observed throughout the entire dataset.

(b) Prototype Contrastive Loss. We regularize the R_{3D} by contrasting them against the global class prototypes $\mathcal{P}$. We define the prototype contrastive loss as $\mathcal{L}_{proto}$:

$$\mathcal{L}_{proto} = -\sum_{i=1}^{C} \log \frac{\exp(\langle f^i_{3D}, p_i \rangle / \tau_{proto})}{\sum_{j=1}^{C} \exp(\langle f^i_{3D}, p_j \rangle / \tau_{proto})}. \tag{3}$$

By minimizing this loss, we enforce a globally consistent and well-structured feature space, compelling the 3D semantic occupancy model to learn representations that are truly discriminative and generalizable across diverse scenes. Ultimately, our pre-training loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{region} + \mathcal{L}_{proto}. \tag{4}$$

4.3 Discussion

Similar to recent methods in self-supervised outdoor semantic occupancy (*e.g.*, GaussTR [19], AGO [24], AutoOcc [62], SelfOcc [16], and S4C [13]), our framework leverages 2D and 3D foundation models. However, our approach markedly diverges from these excellent works in the following two aspects: (1) Robustness to Uncalibrated Video: These outdoor methods rely on the assumption that either exact camera setups or predefined vehicle platforms are available, and their evaluation is conducted almost exclusively within the same data domain. Transferring these approaches to unconstrained, in-the-wild videos devoid of camera priors is nearly infeasible. (2) Model-Agnostic Pre-training: Rather than developing specialized, tightly coupled architectures, we introduce a model-agnostic pre-training framework that could enhance diverse indoor occupancy networks.

5 Experiments

The experimental setups are presented in Sec. 5.1. Sec. 5.2 evaluates the data pipeline using various methods, and Sec. 5.3 evaluates the pre-training paradigm on different architectures and datasets. Sec. 5.4 provides ablation studies to assess each component.

5.1 Experiment Settings

Datasets. We evaluate on two standard 3D semantic occupancy benchmarks, NYUv2 [38] and Occ-ScanNet [55], alongside our proposed YouTube-Occ. All datasets share a $60 \times 60 \times 36$ occupancy grid at 8 cm resolution, annotated with 13 classes (11 semantic, 1 free, 1 unknown). NYUv2 and Occ-ScanNet contain 795/654 and 45,755/19,764 train/test images, respectively, while YouTube-Occ contributes 100,372 frames across 5,241 scenes. Following [3], we report IoU and mIoU for all semantic classes.

Pre-training and Fine-tuning Details. We validate our method across three diverse architectures: MonoScene (Voxel-based CNN) [3], Symphonies (Voxel-based Transformer) [18], and EmbodiedOcc (Gaussian-based Transformer) [47]. Pre-training utilizes DINOv2 [32] for 2D features and SAN [49] for text-to-pixel association. During fine-tuning, the occupancy network retains pre-trained weights while the head is trained from scratch. Both stages employ the AdamW optimizer (initial lr=2e-4) with a 0.1 multi-step decay at predefined epochs. Following [55], we train for 30 epochs on NYUv2 and 10 on Occ-ScanNet and YouTube-Occ. Experiments are conducted on four RTX A6000 GPUs with a batch size of 4.

Table 1: Evaluation of pseudo labels with various methods on Occ-ScanNet. Symphonies serves as the base model, and pseudo-labels are generated by our data pipeline.

Method	Training Data	mIoU↑	IoU↑
Pseudo-Label	None	10.06	27.00
Zero-Shot	YouTube-Occ	8.41	12.74
Joint-Training	NYUv2 + Occ-ScanNet	40.17	51.97
Joint-Training	YouTube-Occ + Occ-ScanNet	37.38	55.06
Fine-Tuning	Occ-ScanNet	**47.71**	**59.89**

5.2 Evaluation of Data Pipeline

Prior to utilizing unconstrained internet data, we first evaluate the quality of the pseudo-labels from our pipeline on Occ-ScanNet, which provides ground-truth 3D semantic occupancy annotations. Specifically, we adopt common supervised learning strategies (see Tab. 1), and present qualitative results in Fig. 4. Directly training Symphonies on YouTube-Occ and validating on Occ-ScanNet (Zero-Shot) yields an unsatisfactory result of 8.41% mIoU, which is even lower than evaluating the raw pseudo-labels (10.06% mIoU). Furthermore, naively jointly training Occ-ScanNet and YouTube-Occ fails to bring consistent improvements, degrading the mIoU compared to the NYUv2 joint-training baseline (37.38% vs. 40.17%). These results indicate that simply applying standard supervision is insufficient to harness the potential of these internet-derived pseudo-labels, necessitating a more sophisticated approach. To this end, we develop a pre-training framework to exploit the potential of our collected in-the-wild videos.

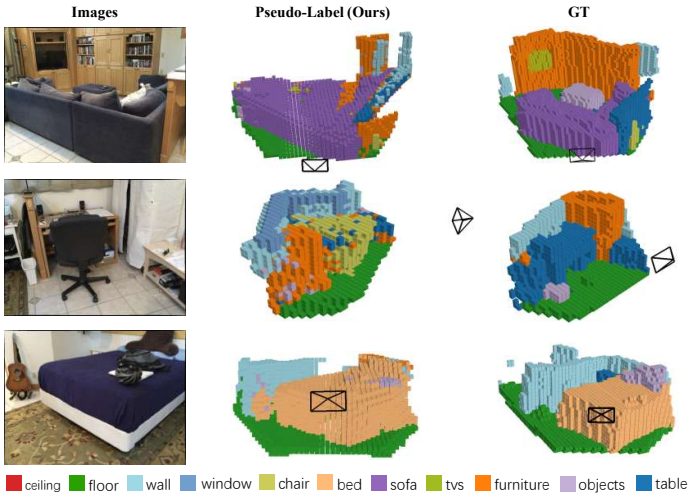

Fig. 4: 3D Semantic Occupancy Pseudo-label visualization in Occ-ScanNet

5.3 Evaluation of Pre-training Framework

Due to the absence of prior work that leverages web data for semantic occupancy, we evaluate the pre-training framework from three perspectives: (1) existing models with/without pre-training, (2) impact of pre-training on web data, and (3) comparison with alternative plug-and-play methods.

Table 2: Performance comparison on NYUv2 and Occ-ScanNet. All methods are evaluated using the standard IoU and mIoU metrics. **Bold** denotes the highest score. †indicates reproduced results using the official public code.

Method	ceiling	floor	wall	window	chair	bed	sofa	table	tvs	furniture	objects	mIoU↑	IoU↑
NYUv2													
LMSCNet [33]	4.49	88.41	4.63	0.25	3.94	32.03	15.44	6.57	0.02	14.51	4.39	15.88	33.93
AICNet [22]	7.58	82.97	9.15	0.05	6.93	35.87	22.92	11.11	0.71	15.90	6.45	18.15	30.03
3DSketch [7]	8.53	90.45	9.94	5.67	10.64	42.29	29.21	13.88	9.38	23.83	8.19	22.91	38.64
MonoScene [3]	8.89	93.50	12.06	12.57	13.72	48.19	36.11	15.13	15.22	27.96	12.94	26.94	42.51
NDC-Scene [54]	12.02	93.51	13.11	13.77	15.83	49.57	39.87	17.17	24.57	31.00	14.96	29.03	44.17
ISO [55]	14.21	93.47	15.89	15.14	**18.55**	**50.01**	40.82	**18.25**	**25.90**	34.08	17.67	31.25	47.11
MonoScene†	8.52	93.50	11.67	10.81	12.87	47.23	35.81	14.82	12.56	26.06	13.17	26.09	41.87
w/ Pre-training	8.73	**93.63**	12.38	11.66	13.77	48.13	37.09	15.67	14.66	27.91	13.62	27.02	42.61
EmbodiedOcc†	3.50	67.20	2.30	6.90	7.10	35.60	33.10	11.50	17.00	18.30	8.10	19.14	30.46
w/ Pre-training	5.20	70.00	6.10	7.30	8.00	34.90	32.90	12.10	17.90	18.20	9.30	20.18	31.85
Symphonies †	14.54	86.59	25.95	15.69	16.78	46.60	38.06	15.37	15.33	32.16	19.58	29.70	49.91
w/ Pre-training	**14.98**	93.13	**26.54**	**18.36**	16.88	48.15	**41.37**	17.62	15.80	**34.72**	**20.64**	**31.65**	**52.15**
Occ-ScanNet													
MonoScene [3]	15.17	44.71	22.41	12.55	26.11	27.03	35.91	28.32	6.57	32.16	19.84	24.62	41.60
ISO [55]	19.88	41.88	22.37	16.98	29.09	42.43	42.00	29.60	10.62	36.36	24.61	28.71	42.16
EmbodiedOcc [47]	40.90	50.80	41.90	33.00	41.20	55.20	61.90	43.80	35.40	53.50	42.90	45.48	53.95
EmbodiedOcc++ [39]	36.40	53.10	41.80	34.40	42.90	57.30	**64.10**	45.20	34.80	54.20	44.10	46.20	54.90
RoboOcc [58]	45.36	53.49	44.35	34.81	43.38	56.93	63.35	46.35	36.12	**55.48**	44.78	47.67	56.48
EmbodiedOcc†	40.30	50.30	41.60	32.10	40.80	52.70	60.00	43.60	36.00	52.30	42.00	44.69	53.23
w/ Pre-training	42.20	50.00	41.70	34.60	41.40	57.20	61.70	44.20	38.50	53.70	42.70	46.17	53.80
Symphonies†	42.14	57.33	47.92	38.50	43.58	56.99	60.16	45.42	37.44	51.54	43.80	47.71	59.89
w/ Pre-training	**46.81**	**57.85**	**50.29**	**40.47**	**44.75**	**57.36**	60.33	**47.00**	**38.88**	53.26	**46.21**	**49.38**	**61.25**

Existing Models Equipped with Pre-Training Tab. 2 demonstrates that our pre-training framework consistently enhances performance across multiple architectures and benchmarks. Integrated with Symphonies [18], our method achieves state-of-the-art (SOTA) results, reaching 31.65% mIoU (+1.95%) on NYUv2 and 49.38% mIoU (+1.67%) on Occ-ScanNet. Crucially, these improvements are model-agnostic. We observe consistent gains across diverse baselines, including EmbodiedOcc (+1.04% and +1.48% on NYUv2 and Occ-ScanNet, respectively) and MonoScene (+0.93% on NYUv2). This indicates that our framework learns robust and generalizable 3D priors. Qualitative results (Fig. 5) further confirm that our approach yields more structurally coherent and accurate semantic completions than the baselines.

Pre-training on YouTube-Occ Beyond fine-tuning on the full dataset, we further conduct a few-shot evaluation to assess the quality of representations learned from YouTube-Occ under limited supervision (Tab. 3). We pre-train the Symphonies model on two scales of our dataset, the full YouTube-Occ and

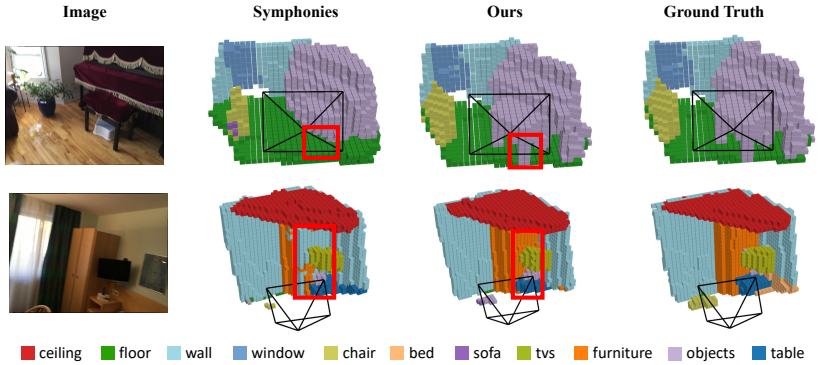

Fig. 5: Qualitative results of the model with pre-training on the Occ-ScanNet dataset.

a 50K-sample subset (YouTube-Occ-50K), and compare them against random initialization and in-domain pre-training baselines.

Evaluations demonstrate that YouTube-Occ enhances data-efficient learning. On NYUv2, our pre-training provides a superior initialization compared to random weights, achieving 18.81% mIoU (+3.29%) with only 5% of the labels, and a +5.09% gain at the 10% split. These benefits extend to the Occ-ScanNet benchmark, where pre-training delivers a +2.36% mIoU boost in the 10% few-shot setting. Furthermore, we observe a clear scaling trend: the full YouTube-Occ dataset consistently outperforms the 50K subset. Overall, these results confirm that the scale and diversity of our dataset equip 3D occupancy models with robust, generalizable semantic representations, improving both downstream performance and data efficiency.

Table 3: Impact of pre-training on few-shot 3D semantic occupancy prediction. We evaluate Symphonies pre-trained on different data sources, fine-tuned with varying labeled data ratios on NYUv2 and Occ-ScanNet. **Bold** numbers mark the best result for each setting, and green numbers show the gain over random initialization.

Pre-training	**5%**		**10%**		**20%**		**50%**		**Full**	
DataSource	**mIoU↑**	**IoU↑**	**mIoU↑**	**IoU↑**	**mIoU↑**	**IoU↑**	**mIoU↑**	**IoU↑**	**mIoU↑**	**IoU↑**
NYUv2										
Random	15.52	42.61	17.43	43.99	23.26	46.49	29.15	49.71	29.70	49.91
YouTube-Occ-50K	18.25	43.19	22.36	44.58	25.20	46.82	29.51	49.77	30.87	51.37
(gain)	+2.73	+0.58	+4.93	+0.59	+1.94	+0.33	+0.36	+0.06	+1.17	+1.46
YouTube-Occ	18.81	44.04	22.52	45.85	26.01	47.43	29.97	50.36	31.45	**52.88**
(gain)	+3.29	+1.43	+5.09	+1.86	+2.75	+0.94	+0.82	+0.65	+1.75	+2.97
NYUv2	**19.87**	**44.23**	**23.58**	**46.37**	**26.64**	**48.37**	**30.30**	**50.83**	**31.65**	52.15
(gain)	+4.35	+1.62	+6.15	+2.38	+3.38	+1.88	+1.15	+1.12	+1.95	+2.24
Occ-ScanNet										
Random	31.77	49.43	34.18	50.16	38.06	53.26	43.74	57.58	47.71	59.89
YouTube-Occ-50K	33.18	49.92	36.26	51.59	38.90	**54.19**	44.34	58.03	48.15	60.33
(gain)	+1.41	+0.49	+2.08	+1.43	+0.84	+0.93	+0.60	+0.45	+0.44	+0.43
YouTube-Occ	33.50	**50.42**	36.54	51.78	39.33	53.57	44.60	57.90	48.47	60.63
(gain)	+1.73	+0.99	+2.36	+1.62	+1.27	+0.31	+0.86	+0.32	+0.76	+0.73
Occ-ScanNet	**33.87**	49.60	**37.80**	**52.56**	**39.63**	53.94	**45.86**	**58.59**	**49.38**	**61.25**
(gain)	+2.10	+0.17	+3.62	+2.40	+1.57	+0.68	+2.12	+1.01	+1.67	+1.35

Comparison to Existing Strategies We compare our pre-training framework against other training strategies in Tab. 4. Our approach outperforms methods that rely on auxiliary losses, specifically HASSC [42] and GaussRender [4]. Moreover, previous 3D pre-training method Pri3D [14] even underperforms the vanilla baseline, likely because it only enforces pixel-voxel consistency with an isolated model, limiting transferability. These results demonstrate the effectiveness of our pre-training approach for improving 3D semantic occupancy estimation.

Table 4: Comparison of different training strategies on the NYUv2 dataset. Results are reported in terms of mIoU/IoU (%).

Method	Training Strategy	Occupancy Networks	
		Symphonies	EmbodiedOcc
Vanilla	Standard	29.70/49.91	19.14/30.46
w/ HASSC [42]	Auxiliary Loss	30.70/51.15	19.54/30.42
w/ GaussRender [4]	Auxiliary Loss	30.37/50.77	19.22/30.96
w/ Pri3D [14]	Pre-training	28.40/49.32	17.94/29.01
w/ Pre-training (Ours)	Pre-training	**31.65/52.15**	**20.18/31.85**

5.4 Ablation Study

To investigate the individual contributions of our proposed modules, Tab. 5 ablates the impact of the two loss components, $\mathcal{L}_{region}$ and $\mathcal{L}_{proto}$. The baseline model yields 29.70% and 47.71% mIoU on NYUv2 and Occ-ScanNet, respectively. Incorporating $\mathcal{L}_{region}$ for intra-frame alignment improves the NYUv2 mIoU to 31.12%, while introducing $\mathcal{L}_{proto}$ for cross-scene consistency achieves 31.24%. Combining both components yields the best performance, reaching 31.65% on NYUv2 and 49.38% on Occ-ScanNet. These correspond to absolute gains of +1.95% and +1.67% over the baseline, demonstrating that local region alignment and global prototype consistency are highly complementary for 3D occupancy representation learning.

Table 5: Ablation study of the two feature-alignment losses in our pre-training framework.

Region Loss	Prototype Loss	NYUv2		Occ-ScanNet	
		mIoU↑	IoU↑	mIoU↑	IoU↑
		29.70	49.91	47.71	59.89
✓		31.12	52.09	48.15	60.31
	✓	31.24	51.76	48.76	60.47
✓	✓	**31.65**	**52.15**	**49.38**	**61.25**

Tab. 6 ablates the proposed projection and prototype mechanisms on the NYUv2 dataset. The default configuration, utilizing Gaussianization and a momentum of $\beta = 0.999$, yields the best performance at 31.65% mIoU and 52.15%

IoU. Replacing Gaussianization with direct 3D projection or introducing learnable scale and rotation leads to notable performance drops of -1.07% and -1.76% mIoU, respectively, justifying the effectiveness of the introduced geometric prior. Similarly, altering the prototype momentum to 0.9 or applying a semantic threshold of 0.3 degrades performance. This could indicate that our prototype-based module can alleviate the inevitable noise in pseudo labels.

Table 6: Ablation study of the projection and prototype mechanisms on the NYUv2 dataset.

Configuration	**mIoU↑**	**IoU↑**
Default (Gaussianization and $\beta = 0.999$)	**31.65**	**52.15**
Projection Variants		
Direct 3D Projection	30.58	50.83
Learnable Scale & Rotation	29.89	50.53
Prototype Variants		
Momentum $\beta = 0.9$	30.37	50.34
Semantic Threshold = 0.3	30.42	50.30

6 Conclusion and Limitation

Our work introduces YouTube-Occ, an automated pipeline for generating pseudo-labeled 3D data from uncalibrated internet videos, alongside a self-supervised distillation framework that boosts 3D occupancy networks by leveraging foundation models. Experimental results on NYUv2 and Occ-ScanNet validate that our data pipeline and pre-training framework deliver consistent improvements, particularly in data-scarce settings. This study demonstrates that rich 3D priors can be extracted from in-the-wild videos, underscoring web-scale data as a promising resource for 3D perception. **Limitations** of our method include handling invisible or occluded voxels, extending to multi-view or temporal networks, and generalizing beyond the wall height assumption.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62502159, W2521174, 62476090, U23A20343, 62302167, 62222602), Fundamental Research Funds for the Central Universities, Natural Science Foundation of Shanghai (Grant No. 25ZR1402135), the Shanghai Committee of Science and Technology (Grant No. 25511103300, 25511104302, 25511102700), "Chen Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (Grant No. 25CGA22), Natural Science Foundation of Chongqing (Grant No. CSTB2025NSCQ-GPX0445), Young Elite Scientists Sponsorship Program by CAST (Grant No. YESS20240780), Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2501), Zhejiang University, Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU.

References

1. Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Nießner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 2614–2623 (2019) 5
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 1
3. Cao, A.Q., De Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022) 2, 3, 10, 12
4. Chambon, L., Zablocki, E., Boulch, A., Chen, M., Cord, M.: Gaussrender: Learning 3d occupancy with gaussian rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 27010–27020 (October 2025) 14
5. Chen, H., Zhang, Z., Qu, Y., Zhang, R., Tan, X., Xie, Y.: Building a strong pre-training baseline for universal 3d large-scale perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19925–19935 (June 2024) 3, 6
6. Chen, J., Xu, H., Wang, Y., Chau, L.P.: Occprophet: Pushing efficiency frontier of camera-only 4d occupancy forecasting with observer-forecaster-refiner framework. arXiv preprint arXiv:2502.15180 (2025) 3
7. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4193–4202 (2020) 12
8. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023) 4
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 4
10. Feng, R., Chang, H.J., Tse, T.H.E., Kim, B., Chang, Y., Gao, Y.: High-resolution spatiotemporal modeling with global-local state space models for video-based human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8929–8938 (2025) 4
11. Gong, J., Xu, J., Tan, X., Song, H., Qu, Y., Xie, Y., Ma, L.: Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11673–11682 (2021) 4
12. Gong, J., Xu, J., Tan, X., Zhou, J., Qu, Y., Xie, Y., Ma, L.: Boundary-aware geometric encoding for semantic segmentation of point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1424–1432 (2021) 4
13. Hayler, A., Wimbauer, F., Muhle, D., Rupprecht, C., Cremers, D.: S4c: Self-supervised semantic scene completion with neural fields. arXiv preprint arXiv:2310.07522 (2023) 4, 10

14. Hou, J., Xie, S., Graham, B., Dai, A., Nießner, M.: Pri3d: Can 3d priors help 2d representation learning? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5693–5702 (October 2021) 14
15. Huang, J., Huang, G., Zhu, Z., Yun, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021) 3
16. Huang, Y., Zheng, W., Zhang, B., Zhou, J., Lu, J.: Selfocc: Self-supervised vision-based 3d occupancy prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19946–19956 (2024) 4, 10
17. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2302.07817 (2023) 3
18. Jiang, H., Cheng, T., Gao, N., Zhang, H., Lin, T., Liu, W., Wang, X.: Symphonize 3d semantic scene completion with contextual instance queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20258–20267 (2024) 2, 6, 10, 12
19. Jiang, H., Liu, L., Cheng, T., Wang, X., Lin, T., Su, Z., Liu, W., Wang, X.: Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11960–11970 (2025) 10
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023) 4
21. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 4
22. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3351–3359 (2020) 12
23. Li, J., He, X., Zhou, C., Cheng, X., Wen, Y., Zhang, D.: Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers. arXiv preprint arXiv:2405.04299 (2024) 3
24. Li, P., Ding, S., Zhou, Y., Zhang, Q., Inak, O., Triess, L., Hanselmann, N., Cordts, M., Zell, A.: Ago: Adaptive grounding for open world 3d occupancy prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8645–8655 (2025) 10
25. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
26. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arXiv preprint arXiv:2203.17270 (2022) 3
27. Li, Z., Yu, Z., Austin, D., Fang, M., Lan, S., Kautz, J., Alvarez, J.M.: FB-OCC: 3D occupancy prediction based on forward-backward view transformation. arXiv:2307.01492 (2023) 3
28. Lin, K., Chen, P., Huang, D., Li, T.H., Tan, M., Gan, C.: Learning vision-and-language navigation from youtube videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8317–8326 (October 2023) 4

29. Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., Wang, X.: Deep dual consecutive network for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 525–534 (2021) 4
30. Ma, Q., Tan, X., Qu, Y., Ma, L., Zhang, Z., Xie, Y.: Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19936–19945 (2024) 3
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 4
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 4, 10
33. Roldao, L., De Charette, R., Verroust-Blondet, A.: Lmscnet: Lightweight multiscale 3d semantic completion. In: 2020 International Conference on 3D Vision (3DV). pp. 111–119. IEEE (2020) 12
34. Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-lidar self-supervised distillation for autonomous driving data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9891–9901 (2022) 3
35. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022) 4
36. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) 4
37. Shah, D., Osiński, B., Levine, S., et al.: Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on robot learning. pp. 492–504. PMLR (2023) 1
38. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. pp. 746–760. Springer (2012) 2, 6, 10
39. Wang, H., Wei, X., Zhang, X., Li, J., Bai, C., Li, Y., Lu, M., Zheng, W., Zhang, S.: Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler. arXiv preprint arXiv:2504.09540 (2025) 12
40. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025) 6
41. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024) 6
42. Wang, S., Yu, J., Li, W., Liu, W., Liu, X., Chen, J., Zhu, J.: Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14792–14801 (June 2024) 14

43. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025) 6
44. Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: Occllama: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv:2409.03272 (2024) 3
45. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. arXiv preprint arXiv:2303.09551 (2023) 3
46. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Scfusion: Real-time incremental scene reconstruction with semantic completion. In: 2020 International Conference on 3D Vision (3DV). pp. 801–810. IEEE (2020) 4, 5
47. Wu, Y., Zheng, W., Zuo, S., Huang, Y., Zhou, J., Lu, J.: Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. arXiv preprint arXiv:2412.04380 (2024) 10, 12
48. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015) 5
49. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2945–2954 (June 2023) 6, 10
50. Xu, T., Lu, H., Yan, X., Cai, Y., Liu, B., Chen, Y.: Occ-llm: Enhancing autonomous driving with occupancy-based large language models. arXiv preprint arXiv:2502.06419 (2025) 3
51. Xu, X., Kong, L., Shuai, H., Zhang, W., Pan, L., Chen, K., Liu, Z., Liu, Q.: 4d contrastive superflows are dense 3d representation learners. In: European Conference on Computer Vision. pp. 58–80. Springer (2024) 3
52. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., Zhou, J., Dai, J.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. ArXiv (2022) 3
53. Yang, Y., Mei, J., Ma, Y., Du, S., Chen, W., Qian, Y., Feng, Y., Liu, Y.: Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. arXiv preprint arXiv:2408.14197 (2024) 3
54. Yao, J., Li, C., Sun, K., Cai, Y., Li, H., Ouyang, W., Li, H.: Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9421–9431. IEEE Computer Society (2023) 12
55. Yu, H., Wang, Y., Chen, Y., Zhang, Z.: Monocular occupancy prediction for scalable indoor scenes. In: European Conference on Computer Vision. pp. 38–54. Springer (2024) 2, 4, 6, 10, 12
56. Zhang, C., Yan, J., Wei, Y., Li, J., Liu, L., Tang, Y., Duan, Y., Lu, J.: Occnerf: Advancing 3d occupancy prediction in lidar-free environments. arXiv preprint arXiv:2312.09243 (2023) 4
57. Zhang, Y., Zhu, Z., Du, D.: Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. arXiv preprint arXiv:2304.05316 (2023) 3
58. Zhang, Z., Zhang, Q., Cui, W., Shi, S., Guo, Y., Han, G., Zhao, W., Ren, H., Xu, R., Tang, J.: Roboocc: Enhancing the geometric and semantic scene understanding for robots. arXiv preprint arXiv:2504.14604 (2025) 12

59. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. arXiv preprint arXiv: 2311.16038 (2023) 3
60. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 6
61. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018) 5, 6
62. Zhou, X., Wang, J., Wang, Y., Wei, Y., Dong, N., Yang, M.H.: Autoocc: Automatic open-ended semantic occupancy annotation via vision-language guided gaussian splatting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3367–3377 (2025) 10
63. Zou, J., Huang, T., Yang, G., Guo, Z., Luo, T., Feng, C.M., Zuo, W.: Unim 2 ae: Multi-modal masked autoencoders with unified 3d representation for 3d perception in autonomous driving. In: European Conference on Computer Vision. pp. 296–313. Springer (2024) 3