Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Open-Vocabulary 3D Indoor Scene Segmentation: A Systematic Review

Hussam Azzuni

h.azzuni@ecu.edu.au

Edith Cowan University

Syed Mohammed Shamsul Islam

Edith Cowan University

Viet Huynh

Edith Cowan University

Mohammed Bennamoun

The University of Western Australia

Farid Boussaid

The University of Western Australia

Syed Afaq Ali Shah

Edith Cowan University

Research Article

Keywords: Open-vocabulary learning, 3D Scene Understanding, Vision-Language models, 3D Segmentation

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11475551/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Open-Vocabulary 3D Indoor Scene Segmentation: A Systematic Review

Hussam Azzuni[1*], Syed Mohammed Shamsul Islam[1,2], Viet Huynh[1], Mohammed Bennamoun[2], Farid Boussaid[3], Syed Afaq Ali Shah[1]

[1*]Computer Science, Edith Cowan University, Joondalup, Western Australia, Australia.
[2]Department of Computer Science and Software Engineering, The University of Western Australia, Crawley, Western Australia, Australia.
[3]Department of Electrical, Electronic and Computer Engineering, The University of Western Australia, Crawley, Western Australia, Australia.

*Corresponding author(s). E-mail(s): h.azzuni@ecu.edu.au;
Contributing authors: syed.islam@ecu.edu.au; v.huynh@ecu.edu.au; mohammed.bennamoun@uwa.edu.au; farid.boussaid@uwa.edu.au; afaq.shah@ecu.edu.au;

Abstract

Open-vocabulary 3D indoor scene segmentation seeks to overcome the fixed-category limitation of traditional 3D segmentation tasks by transferring the rich semantic knowledge of large-scale vision-language models to the 3D space. This topic has gained traction with the emergence of vision-language models such as Contrastive Language-Image Pre-training (CLIP) model and continues to develop due to its fundamental importance across various domains in computer vision, robotics, and autonomous driving. Given the growing significance of this field and the rapid expansion of research on open-vocabulary 3D scene understanding, this systematic review examines 93 peer-reviewed papers to provide the first systematic analysis focused on open-vocabulary 3D indoor scene segmentation. These papers are identified and extracted from three databases (Scopus, IEEE Xplore, and ACM Digital Library) following multiple inclusion and exclusion criteria. We critically examine the studies while thematically grouping them based on their main representation: point cloud-based, NeRF-based, and 3DGS-based methods. Then, we synthesise the evaluation protocols adopted across the

existing literature. Finally, we discuss the challenges and promising future directions to guide researchers toward the main research gaps in open-vocabulary 3D segmentation.

Keywords: Open-vocabulary learning, 3D Scene Understanding, Vision-Language models, 3D Segmentation

1 Introduction

Visual scene understanding is a crucial computer vision objective that involves understanding objects, their surrounding environment, and the relationships between them. This encompasses a range of tasks including scene classification, object detection, image segmentation, and scene graph generation. Although 2D scene understanding methods have exhibited strong performance, they inherently lack explicit depth and geometric information. Three-dimensional representations address these limitations by providing accurate spatial reasoning capabilities, which are critical for robotics, autonomous driving, and augmented/virtual reality applications.

Despite the advantages of 3D scene representations, developing deep neural networks for 3D scene understanding continues to face significant challenges. In particular, scalability remains a major obstacle due to the labour-intensive process of semantic 3D annotation and the open-ended nature of object categories found in real-world scenes.

Zero-shot learning (ZSL) was introduced as an initial attempt to address this challenge by enabling models to recognise unseen categories without requiring explicit training samples. Consequently, unseen objects are often treated as background during training, leading to suboptimal generalisation and recognition performance [1].

This limitation has motivated an alternative approach that projects image pixels and text features onto the same feature space [2], which is the leading principle of **open-vocabulary learning (OVL)**. OVL reframes tasks such as object detection and segmentation to use text-based supervision, instead of the time-consuming dense annotations. This becomes apparent with the rise of language models (e.g., BERT [3]) followed by vision-language models (e.g., ALIGN [4], CLIP [5]) as they are fundamental components in enabling OVL across many visual tasks, such as **open-vocabulary 2D object detection (OV-2DOD)**, **2D segmentation (OV-2DS)**, **3D object detection (OV-3DOD)**, and **3D segmentation (OV-3DS)**.

Despite the rapid development of the field, OV-3DS techniques are fragmented across different pipelines and representation types, making it difficult to have a broad understanding of the topic. The existing survey papers primarily focus on 2D open-vocabulary learning whilst some provide a partial exploration of the 3D space. This demonstrates the importance of a dedicated systematic review paper that mainly focuses on open-vocabulary learning for 3D indoor scene segmentation.

Table 1 Comparison with closely related surveys, with symbols denoting discussion depth. ● **Extensive**: (Moderate to extensive discussion including methodologies, taxonomy, and datasets.) ◐ **Moderate:** (Slight to moderate discussion.) ◴ **Slight:** (Short methodology description without an independent section.) ○ **Not Mentioned:** (The topic is not mentioned). The scope column denotes open-vocabulary tasks: 2DOD (2D object detection), 2DS (2D segmentation), 3DOD (3D object detection), 3DS (3D segmentation), NeRF-3DS (NeRF-based 3D segmentation), and 3DGS-3DS (3DGS-based 3D segmentation).

Title	Year	Venue	Scope (Open-vocabulary Tasks)					
			2DOD	2DS	3DOD	3DS	NeRF-3DS	3DGS-3DS
[6]	2023	Arxiv	○	○	○	◐	○	○
[1]	2024	IEEE TPAMI	●	●	◐	◐	○	○
[7]	2024	IEEE TPAMI	●	●	◐	◐	○	○
[8]	2024	IEEE TPAMI	◐	◐	○	○	○	○
[9]	2025	Arxiv	◴	○	●	◴	○	○
[10]	2025	Procedia Computer Science	●	○	○	○	○	○
[11]	2026	IJCV	○	○	○	○	◐	○
Ours	-	-	○	○	○	●	●	●

1.1 Overview of Previous Review Papers

Previous survey papers on OVL and related topics are summarised in Table 1. Understanding their strengths and limitations is essential for assessing their coverage depth and ensuring that this systematic review provides a thorough analysis of OV-3DS.

Lei et al. [6] explore multi-modal 3D scene understanding including the integration of 3D representation with 2D and text modalities. This paper covers various topics in 3D vision-language scene understanding such as visual grounding, dense captioning, question answering, and open-vocabulary 3D recognition. They briefly discuss some OV-3DS approaches and categorise them into distillation-based approaches and methods that predict instance masks using a pre-trained segmentation network followed by an open-vocabulary semantic label assignment step. However, they do not discuss open-vocabulary approaches that use NeRF and 3DGS-based representations. Wu et al. [1] provide a holistic exploration of OVL focusing on tasks such as OV-2DOD, OV-2DS, video and 3D scene understanding. They discuss a few open-vocabulary approaches related to 3D recognition, detection, and segmentation. However, they do not delve deeper into the overarching taxonomy of OV-3DS networks and do not discuss NeRF and 3DGS-based approaches. Furthermore, they do not discuss the available relevant datasets for 3D tasks. Similarly, Zhu et al. [7] examine the existing literature of open-vocabulary detection and segmentation, distinguishing them from ZSL by noting that the former leverages weak supervision signals. However, they only provide a brief overview of 3D scene understanding approaches. This includes briefly discussing multiple open-vocabulary approaches across various tasks such as object detection, semantic segmentation, and instance segmentation. However, it does not provide an extensive analysis and taxonomy for the domain, and it also does not discuss NeRF and 3DGS-based segmentation methodologies. Unlike the previous approaches, Zhang et al. [8] provide a systematic review on vision-language modelling for various tasks, where they briefly discuss **knowledge distillation (KD)** and its capacity to enable OV-2DOD and OV-2DS. However, they do not discuss applying open-vocabulary learning to the 3D domain. Sapkota et al. [9] direct the focus toward 3D object detection and how the field has witnessed a shift toward VLM-driven 3D approaches, enabling open-vocabulary detection. This review briefly

mentions some OV-3DS segmentation approaches but does not discuss any approaches related to NeRF or 3DGS representations. Yu et al. [10] primarily focus on OV-2DOD and categorise its approaches into four main categories: distillation-based, enhanced candidate proposals, region-word alignment-based, and semantic augmentation training. They do not discuss any approach that operates in the 3D space. Finally, Nguyen et al. [11] provide a comprehensive review of semantic-aware **neural radiance fields (NeRF)** in the context of 3D scene understanding. They discuss queryable NeRF-based methodologies. However, they do not discuss point-based or 3DGS-based segmentation approaches.

Unlike existing review papers, this work fills the current research gap and focuses on exploring OV-3DS methods in-depth across point-based and rendering-based representations such as NeRF and **3D Gaussian splatting (3DGS)**.

1.2 Objective of this Systematic Review

The main goal of this systematic review is to provide the reader with a self-contained analytical exploration of OV-3DS across multiple 3D representation types. We start with a preliminary background section, followed by identifying and selecting eligible studies using **Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)** protocol [12]. Finally, we provide a unified methodology taxonomy, dataset descriptions, and performance assessment settings of OV-3DS approaches. To sum up, the **key contributions** of this review are the following:

1. To our knowledge, we conduct **the first** systematic review focused on open-vocabulary 3D indoor scene segmentation, which includes an analytical synthesis of point-based, NeRF-based, and 3DGS-based methods, overall trends, key takeaways, and evaluation settings.
2. We develop a unified taxonomy for point-based and rendering-based approaches based on their main methodological pipeline and 3D representation.
3. We identify the key challenges and propose promising research avenues, guiding future research on OV-3DS.

2 Background, Terminology, and Scope

This section covers relevant background information for a thorough understanding of OV-3DS. This includes 2D **foundation models (FMs)**, knowledge distillation, 3D representations, zero-shot learning, and an introduction to OV-3DS.

2.1 Foundation Models

The development of **vision-language models (VLMs)** has played a major role in advancing 2D scene understanding tasks, such as object detection and image segmentation. Consequently, the emergence of FMs, defined in [13], as "the base models trained on large-scale data in a self-supervised or semi-supervised manner that can be adapted for several other downstream tasks", has extended OVL to 3D scene understanding. These foundation models include:

- Textually-prompted FMs that use an input image with a text query to perform various tasks including image-text matching (e.g., CLIP [5]), object detection (e.g., Grounding DINO [14]), image segmentation (e.g., Grounded-SAM [15]), a general-purpose VLM (e.g., LLaVA [16]), and image generation (e.g., Stable Diffusion [17]).
- Visually-prompted FMs that use an input image and a visual prompt (e.g., point, bounding box, mask) to perform various tasks (e.g., SAM [18]).

Open-vocabulary 3D segmentation methods use these FMs to adapt their knowledge into the 3D space. For example, textually or visually-prompted detection or segmentation networks play a huge role in region extraction which commonly occurs before extracting open-vocabulary features. The feature extraction step may use various techniques to extract image-level or region-level features. For instance, a method can utilise CLIP or one of its variants directly, or it can use a general-purpose VLM to generate captions or descriptions before extracting their features. Alternatively, the features can be extracted from an image generation network such as Stable Diffusion. The image generation network can also be used as a class-conditioned image generator to establish the capabilities of multiple foundation models before fusing their features.

2.2 Knowledge Distillation

Knowledge Distillation (KD) is the process of training a small student network through the supervision of a larger teacher model [19]. It consists of three main components: knowledge type, training strategy, and the teacher-student network architecture [19]. The knowledge type can be categorised into response-based, feature-based, and relation-based knowledge; depending on whether logits, features, or the relationships between different layers or data samples are used as a target supervision signal during distillation [19]. Distillation strategies are categorised into offline, online, or self-distillation. Offline KD uses a pre-trained teacher network to provide a target supervision signal. Online KD jointly trains the teacher and student branches, while self-distillation transfers the knowledge within the same network family across different layers and stages [19]. Finally, a teacher-student framework is incorporated to enable knowledge distillation.

Knowledge distillation plays a fundamental role across many open-vocabulary 3D segmentation methods; it transfers the knowledge from pre-trained foundation models into a 3D backbone or representation to enable and refine the segmentation performance using text queries. The transfer of knowledge occurs by aligning point-caption features using the direct application of cosine or distribution-based alignment losses, or through a contrastive learning training strategy. Other approaches use soft-labels extracted from pre-trained vision-language models such as probabilistic spatial-semantic labels, or hard-labels such as CLIP-based pseudo-labels or mask predictions as the supervision signal. Furthermore, some methods use self-distillation to refine the learned 3D representation by employing an exponential moving average (EMA) and geometric priors, or by using a temporal ensembling self-distillation approach that utilises pseudo-labels from the previous epochs as target supervision signals. In addition, some rendering-based approaches leverage novel view synthesis to enable self-distillation using rendered novel viewpoints.

2.3 3D Representations

A prerequisite for understanding OV-3DS methods is familiarity with different representation types. They are categorised into explicit (e.g., point cloud, voxels, mesh, 3DGS) and implicit representations (e.g., NeRF). We will only explain NeRF and 3DGS representations in detail, as the remaining types are more common concepts.

Neural Radiance Field (NeRF) is an implicit representation that represents 3D scenes as a continuous function using multi-view observations and their corresponding camera poses. NeRFs leverage multilayer perceptrons (MLPs) to map a 5D input vector that consists of 3D point coordinates (x, y, z) and 2D viewing directions (θ, ϕ) to predict a volume density and a view-dependent RGB colour for each point along the camera ray. The main idea is to leverage the concept of volume rendering and predict the colour and volume density for sampled points along the ray to obtain a rendered pixel, which is used to train the network by comparing it with the true pixel colours [20]. Recent radiance-field approaches explore replacing or augmenting the MLP with explicit grids, feature planes, hash encodings or tensor decomposition.

3D Gaussian Splatting (3DGS) is an unstructured and explicit 3D representation that combines the high visual quality of radiance fields with real-time rendering for novel view synthesis. In this approach, Gaussians are represented as ellipsoids initialised from a sparse point cloud and parameterised by their mean 3D position, covariance matrix, opacity, and colour encoded using spherical harmonic coefficients. During training, the 3D Gaussians are projected onto a 2D plane using tile-based rasterisation, which performs visibility-based Gaussian ordering and α-blending to generate the final rendered image [21].

2.4 Zero-Shot Learning

Zero-shot learning (ZSL) aims to train a network using labelled instances of seen classes while enabling it to classify unseen classes with the help of the shared semantic information between seen and unseen categories. Since evaluating performance exclusively on unseen classes can be impractical, the **generalised zero-shot learning (GZSL)** protocol evaluates performance on both seen and unseen classes [22]. These two evaluation protocols have played a major role in assessing the performance of OV-3DS methods on base and novel classes.

2.5 Open-Vocabulary 3D Segmentation: Scope and Definitions

The development of **open-vocabulary 3D learning (OV-3DL)** targets the scalability challenge of ZSL. We define OV-3DS approaches as follows: "Any approach that uses the language-aligned label space of VLMs to enable performing a segmentation task using a free-form text query input". More specifically, it leverages 2D VLMs to use or distil their knowledge to enable a 3D network to segment objects. This includes various tasks such as **open-vocabulary 3D semantic segmentation (OV-3DSS)** that focuses on segmenting the semantic region/object within the scene given a text query. **Open-vocabulary 3D instance segmentation (OV-3DIS)** takes it one step further by identifying and spatially separating individual instances, then assigning each instance a category. **Referring expression segmentation (RES)** segments

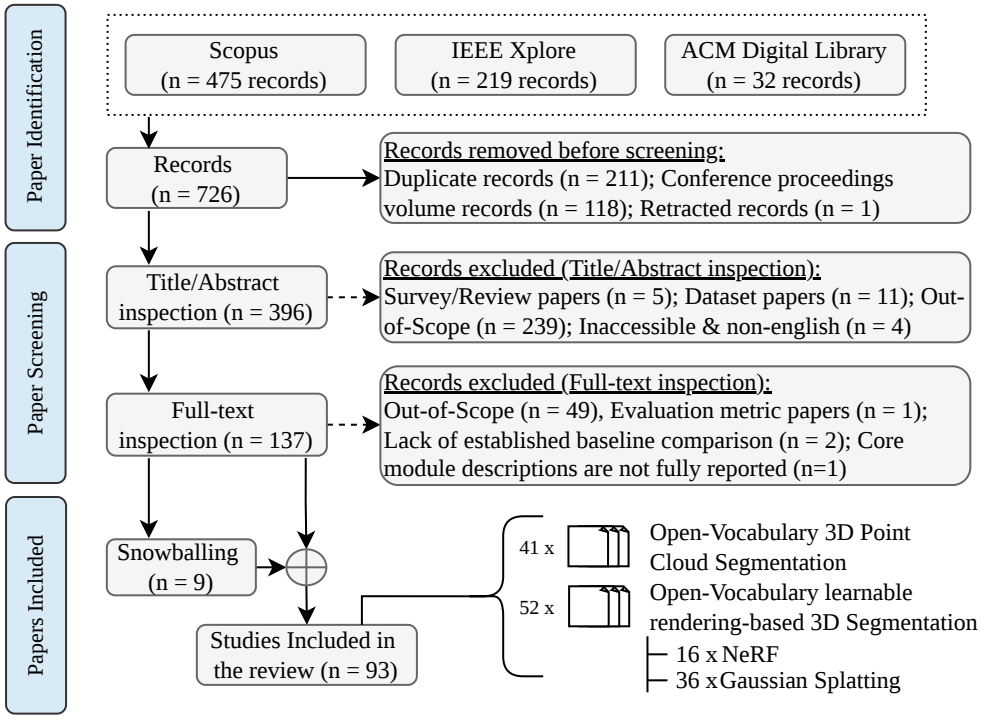


Fig. 1 PRISMA diagram showing the article selection process.

an object instance by providing a query expression (e.g. attribute, inter-object spatial relationship) of the target object. Finally, **open-vocabulary 3D panoptic segmentation (OV-3DPS)** is a combination of semantic and instance segmentation tasks; it assigns semantic labels to all the regions within the scene while also distinguishing individual countable object instances.

3 Systematic Review Methodology

This systematic literature review explores OV-3DS while following the PRISMA protocol [12] to ensure the consistency and reproducibility of our study as shown in Fig. 1. The first step is to define research questions that tackle various aspects of the topic, followed by a paper extraction methodology to select the relevant works and finally, synthesise findings across the papers throughout the survey.

1. RQ1: Which OV-3DS approaches primarily use a point cloud representation?
2. RQ2: What methods address OV-3DS using a neural radiance field representation?
3. RQ3: What methods tackle OV-3DS using a 3D Gaussian splatting representation?
4. RQ4: What datasets and evaluation settings are used for OV-3DS assessment?

3.1 Paper Identification

We collect articles about OV-3DL from three major databases (IEEE Xplore, ACM Digital Library, and Elsevier Scopus) based on the research questions. We explore different relevant methodologies to choose the most suitable keywords for our search query. Our initial search targets papers published between (01/01/2015 - 31/12/2025) and contains these keywords in the **abstract** section as shown in the box below. Then, we use a filtering stage to remove duplicate records, retracted papers, and records that cover the entire conference proceedings volume before going into the screening stage.

("open vocabulary" OR "open world" OR "open set") AND ("3D") AND ("scene" OR "scenes" OR "environment" OR "environments" OR "multi-view" OR "spaces" OR "scenarios" OR "observations" OR "indoor" OR "outdoor")

3.2 Paper Screening

After identifying the unique articles from the previous stage, we proceed with a multi-stage screening process to keep the most relevant articles and ensure reproducibility.

1. **Title/Abstract Inspection:** This stage focuses on removing irrelevant works by inspecting the title and abstract of each paper. As this systematic review targets a comprehensive understanding of open-vocabulary 3D indoor segmentation, we omit documents that are out-of-scope of the study following these **exclusion criteria**:
   - Inaccessible & non-English works.
   - Survey, review & dataset papers.
   - Irrelevant works: face biometric tasks, object-level 3D understanding and retrieval, etc.
   - Papers that primarily focus on other tasks such as scene graph generation, semantic mapping and robotics tasks, even if they perform segmentation.
2. **Full-Text Inspection:** The final filtering step ensures that the selected papers contribute to OV-3DS if they fulfil these **inclusion criteria**:
   - The article focuses on methods that can perform indoor scene understanding.
   - The article primarily contributes to OV-3DS.
   - The article provides a quantitative comparison with baseline methods.

3.3 Handling Data-Inclusion Bias

After extracting the relevant papers, we address missing documents arising from the keyword-dependent nature of the initial data collection process by applying a "backward snowballing" approach. This includes going through the reference list of each of the included papers to identify and include those relevant studies. Also, we conducted multiple rounds of internal revision, refining the inclusion criteria, leading to the final list of included studies.

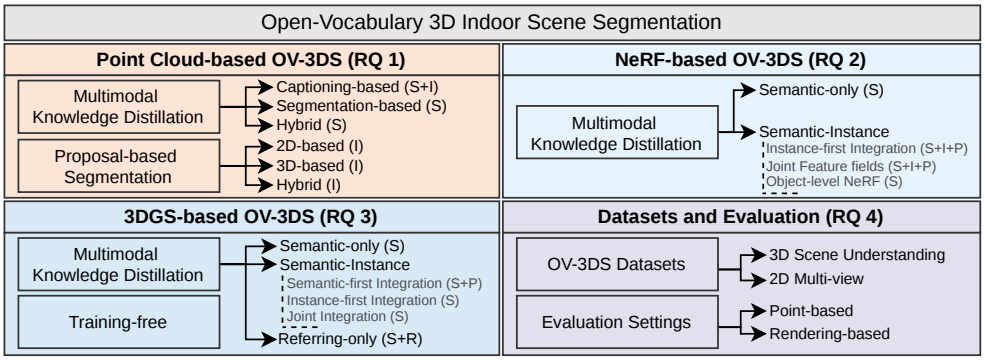


Fig. 2 Overview of the taxonomy and research directions covered in this systematic review, organised according to the four main research questions. We provide a task label after the categories to indicate the main segmentation tasks such as semantic segmentation (S), instance segmentation (I), panoptic segmentation (P), and referring-expression segmentation (R).

3.4 Data Extraction and Synthesis

After identifying the papers and performing multi-step screening, we extract **93** peer-reviewed papers and report the findings by categorising them thematically based on their fundamental method, and summarising the datasets and evaluation protocols to provide a comprehensive analysis of the topic following the outline in Fig. 2.

4 Point Cloud-based OV-3DS (RQ 1)

OV-3DS networks are capable of segmenting objects given arbitrary text queries. This section discusses various methods that use point clouds as their main 3D representation, and are categorised into two major groups: multimodal knowledge distillation and proposal-based segmentation.

Multimodal knowledge distillation uses 2D foundation models with open-vocabulary capabilities and transfers their knowledge into a 3D encoder using pixel-point correspondence to enable OV-3DS. The approaches belonging to this type are categorised based on the supervision source including captioning-based, segmentation-based, and hybrid methods. Alternatively, **proposal-based segmentation** commonly uses a class-agnostic mask proposal module, and then associates the open-vocabulary knowledge using a feature/label computation module. These approaches are categorised based on the modality used for mask proposal prediction: 2D-based, 3D-based, and hybrid proposal prediction.

Each group has its strengths and weaknesses. For example, even though KD approaches suffer from initial training overhead, they can be computationally efficient during inference. On the other hand, proposal-based segmentation approaches do not usually require additional training as they leverage pre-trained networks; however, they suffer from slow inference due to the use of one or multiple FMs. Furthermore, proposal-based approaches focus on OV-3DIS, while distillation-based approaches

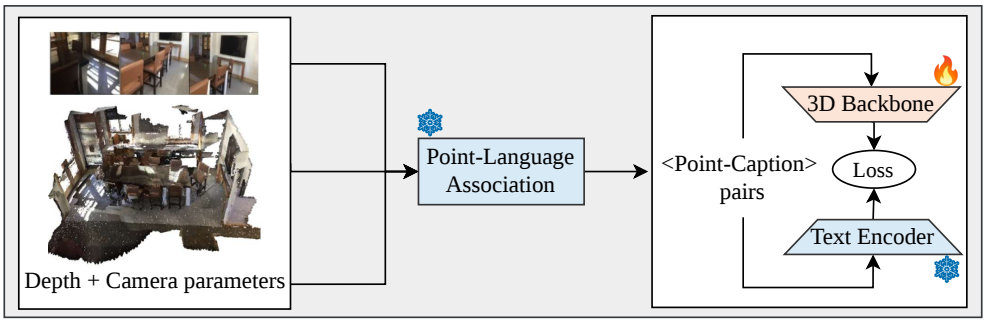


Fig. 3 This figure summarises a captioning-based KD training strategy. It consists of a point-language association stage that leverages pre-trained networks to generate <point-caption> pairs, which are used to train a 3D backbone with the help of a pre-trained text encoder.

target semantic or instance segmentation depending on their pipeline. This section discusses the core methodologies, the challenges, and research directions.

4.1 Multimodal Knowledge Distillation

4.1.1 Captioning-based Methods (Fig. 3 & Table 2)

This pipeline leverages 2D foundation models and image captioning networks to build a large-scale set of <point-language> pairs and uses them to train OV-3DS networks. These approaches are categorised based on caption granularity: image-level [23–27], or entity-level captioning [28–31].

The development of captioning-based algorithms began as a solution to the lack of large-scale 3D-language datasets. Thus, it uses 2D foundation models as a bridge toward using their knowledge for 3D scene understanding. Early approaches [23] extract multi-granular captions from multi-view images before training the 3D backbone. These approaches struggle with various limitations as they rely on pre-trained 2D networks followed by a back-projection step.

The first challenge is imprecise caption prediction and projection. This is addressed as the proposed methodologies move toward region-level captioning followed by entity-level captioning, where they leverage a general-purpose VLM for caption generation to ensure cleaner data. A notable contribution comes from [29] as the authors propose a data generation pipeline and a two-stage trainable OV-3DS network called Mosaic3D which is a step forward toward more scalable 3D-language approaches, whereas [26] improves 3D-language alignment through refining the language branch in the pipeline.

The second challenge is conflicting supervision from multi-view captions, resulting in noisy point-language pairs. Existing approaches address this limitation using different strategies, either by proposing a two-stage training scheme to improve the semantic information and multi-view feature consistency [31], or through utilising contrastive learning to train the 3D backbone to achieve view-invariant entity representations [30].

Key Takeaways: Captioning-based methods use pre-trained 2D foundation models and multi-view RGB images to train a 3D network for OV-3DS. These approaches

Table 2 Comparison of captioning-based distillation approaches for open-vocabulary 3D segmentation. The table summarises how different methods generate <point-language> pairs, and the text encoders employed to train a sparse 3D UNet backbone for all approaches. The asterisk (*) denotes the use of trainable injected tokens, while the (X) indicates that the module or training strategy is not being used. The methods are grouped by type: image-level networks are unshaded, and entity-level methods are grey. Abbreviations: Ext: Extraction; SW: Sliding Window.

Methods	**<Point-Language> Pair Generation**				**Training**
	Region Ext.	Captioning	Dense Captioning	Caption Granularity	Text Encoder
PLA [23], Lowis3D [24]	×	GPT-ViT 2	×	Scene, View, Entity	CLIP
RegionPLC [25]	Detic [32]/SW	CLIP/OFA [33]	GRiT [34]/Kosmos-2 [35]	Proposal	CLIP
AiDe [26]	×	GPT-ViT 2	×	Scene, View, Entity	CLIP*
WS3D++ [27]	2D+3D RPN	GPT-4	×	Scene, Proposal	CLIP
OV3D [28]	SAM/SEEM [36]	LLaVA 1.5 [37]	×	Entity	CLIP
MPEC [30]	PQ3D	[38]	×	Entity	CLIP
PGOV3D [31]	G-SAM	LLaVA-NeXT	×	Entity	CLIP
Mosaic3D [29]	RAM++ & G-SAM+SEEM	Osprey [39]	×	Entity	Recap-CLIP [40]

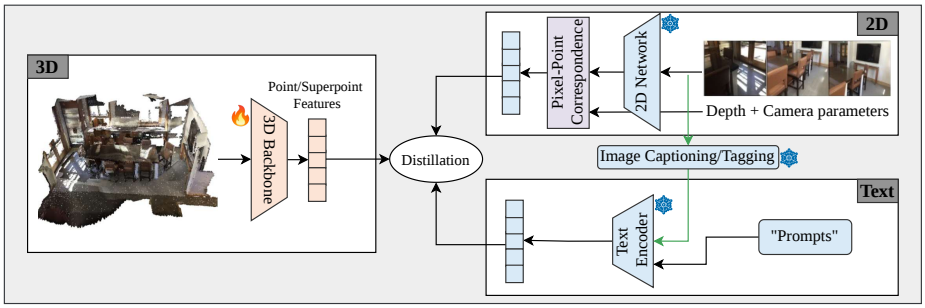


Fig. 4 This figure summarises a segmentation-based KD training strategy. It uses pre-trained 2D and text networks to distil their knowledge into a 3D backbone with the help of pixel-point correspondence.

retain the same pipeline, but progressively shift toward the use of more fine-grained information, evolving from image-level to region-level and, more recently, entity-level representations for better 3D backbone training. This progression shows promising results as it led to enhancing the segmentation performance. Alternatively, adapting the text encoder with 3D semantic information [26] offers a promising avenue as it rivals the performance of entity-level methods. Beyond these findings, the ablation study in [23] demonstrates that using entity-only captions yields better performance than using full-scene captions.

4.1.2 Segmentation-based Methods (Fig. 4 & Table 3)

This type leverages 2D foundation models with OV-2DS capabilities to distil their information directly into a 3D network. These methods are commonly called annotation-free approaches as they do not utilise labels during distillation, and can be categorised based on whether they use a single-visual network [41, 43, 44, 46, 48, 50, 52, 53, 56, 60], or multiple-visual networks [57, 59] to provide the supervision signal. These approaches follow a common pipeline as they predict dense open-vocabulary features, back-project them into 3D, and then train a 3D backbone by aligning the predicted features with the projected features.

Unlike the concurrent captioning-based paradigm that creates <point-language> pairs before training, segmentation-based methods directly distil the dense predictions of 2D foundation models into a 3D backbone. These methods typically leverage one foundation model to provide the supervisory signal, resulting in a multitude of limitations inherited from directly projecting 2D predictions. These challenges may impose a hard performance ceiling on 3D segmentation depending on the 2D network used.

Existing approaches demonstrate multiple pathways to handle this limitation. One notable direction is the use of 3D geometric priors to enhance the 3D representation during feature distillation [46, 53], while others leverage a diffusion-based distillation strategy as it provides stronger localisation compared to CLIP-based approaches [50]. Moreover, the adoption of multiple visual networks and leveraging their feature diversity provide an enhanced training signal, resulting in a more robust student network. This could be accomplished by distilling the information of each foundation model

Table 3 Comparison of segmentation-based distillation methods for open-vocabulary 3D segmentation. The table summarises the teacher network, student backbone, and KD targets. The (X) indicates that the module or training strategy is not being used. The methods are grouped by type: single-visual network methods are unshaded, and multiple-visual network approaches are grey. Abbreviations: Feat: Feature; Ext: Extraction; SD: Stable Diffusion; PL: Pseudo-label; SW: Sliding Window; PC: PointClustering.

Methods	Teacher Network					Student Network	KD Target	
	Text	2D Vision			3D Vision	3D Backbone	Feat.	PL
	Feat. Ext.	**Region Ext.**	**Feat. Ext.**	**Unified Ext.**	**Network**			
OpenScene [41]	×	×	×	OpenSeg [42]	×	MinkowskiNet	√	×
DCL-OW3D [43]	CLIP	×	×	OpenSeg	×	MinkowskiNet	√	×
CLIP2Scene [44]	CLIP	×	×	MaskCLIP [45]	×	MinkowskiNet	√	×
GGSD [46]	CLIP	×	×	LSeg [47]	Self-KD	MinkowskiNet	√	×
Shen et al. [48]	CLIP	×	×	LSeg/OVSeg [49]	×	MinkowskiNet	×	√
Diff2Scene [50]	×	×	×	ODISE [51]	×	MinkowskiNet	×	√
CLIP-FO3D [52]	×	SW+superpixels	CLIP	×	×	MinkowskiNet	√	×
CUS3D [53]	CLIP	MaskFormer [54]	CLIP	×	PC [55]	3D Res-UNet	√	√
LSWKD [56]	CLIP	×	×	×	PLA	Sparse UNet	√	×
CUA-O3D [57]	×	×	SD+DINOv2 [58]	LSeg	×	MinkowskiNet	√	×
SAS [59]	CLIP	SEEM	CLIP	LSeg	Self-KD	MinkowskiNet	√	√

independently with an uncertainty-aware scalar to mitigate inconsistencies across multiple views and models [57], or by aligning the foundation models before leveraging a diffusion and SAM-based approach to establish the capability of each model across different classes [59]. In contrast to the typical approach of using feature-based distillation, one notable method [48] explores the use of a semantic probability-based distillation strategy, resulting in a much more efficient pipeline.

Key Takeaways: Segmentation-based methods have evolved beyond the direct distillation of 2D knowledge into 3D space through incorporating additional strategies to improve their robustness and overall performance. One prominent research direction, demonstrated in [46, 53], involves leveraging 3D geometric priors to mitigate noise introduced during 2D-3D back-projection. Notably, [46] shows that, with self-distillation, the 3D student network can even outperform its teacher model, whereas [53] highlights the benefits of jointly applying feature and pseudo-label distillation to obtain more refined representations. A separate research avenue investigates mask-based distillation as an alternative strategy with better localisation compared to conventional point-based schemes [50]. Furthermore, an emerging line of research [57, 59] combines multiple visual networks while addressing feature heterogeneity to provide a more robust supervision signal, resulting in better trained student networks.

4.1.3 Hybrid Methods (Table 4)

This approach combines captioning and segmentation-based techniques to distil open-vocabulary knowledge into a 3D backbone. This is motivated by their individual weaknesses. For instance, captioning-based methods rely on coarse language supervision, while segmentation-based approaches provide dense supervision signals but suffer from noisy view-dependent semantic predictions. Thus, developing two-branch frameworks can provide complementary information, resulting in better segmentation.

Although hybrid methods share a common pipeline, they perform the alignment with the help of language-aligned 2D networks [60–62] while UniM-OV3D [63] also explores the use of 3D pointLLM-based hierarchical semantic caption generation [64], hierarchical feature extraction, and depth supervision to enhance feature alignment.

Key Takeaways: The performance of hybrid methods is reliant on how captioning-based and segmentation-based information complement each other. One key takeaway is that multi-granular 3D-language alignment provides strong complementary features to enhance the overall performance of OV-3DS. One notable approach is [63] as it targets improving both branches, whereas [61] shows that using scene-level captions for 3D-language alignment is sensitive to optimisation design such as loss weights and the training schedule.

4.2 Proposal-based Segmentation

4.2.1 2D-based Proposal Network (Fig. 5 & Table 5)

This type leverages pre-trained 2D networks to obtain segmentation masks, back-project them into 3D before going through a merging/fusion scheme to build the 3D instances for OV-3DIS.

Table 4 Comparison of hybrid knowledge distillation approaches for open-vocabulary 3D segmentation, including captioning-and-segmentation-based teacher networks, along with the student backbone employed. The (X) indicates that the module is not being used. Abbreviations: Feat: Feature; Ext: Extraction; Cap: Captioning.

Methods	**Teacher Network**						**Student Network**
	Captioning-based			Segmentation-based			3D Backbone
	Cap.	**Feat. Ext.**	**Cap. Granularity**	**Feat. Ext.**	**OV-2DS**	**Geo. Branch**	
TAMC [61]	ViT-GPT 2	CLIP	Scene	×	LSeg	×	MinkowskiNet
XMask3D [60]	ViT-GPT 2	CLIP	View	×	ODISE	×	MinkowskiNet
DMA [62]	LLaVA+RAM	CLIP	Scene	×	FC-CLIP [65]	×	MinkowskiNet
UniM-OV3D [63]	PointLLM [64]	Point-BIND [66]	Global, eye, sector	Point-BIND	×	CLIP2Point	Transformer-based

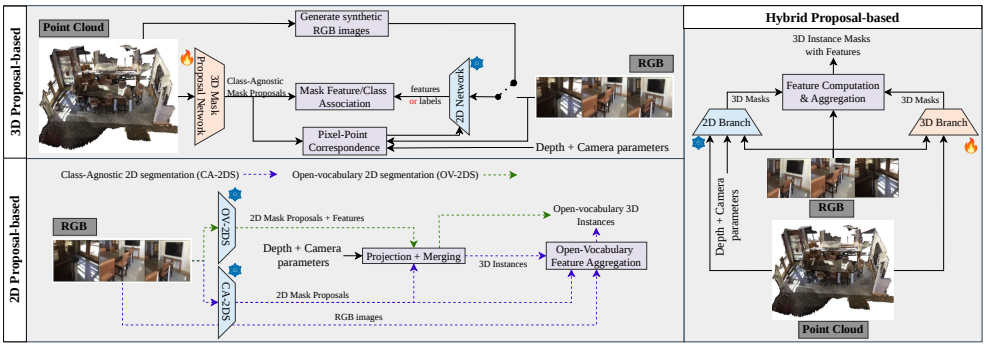


Fig. 5 Overview of the three proposal-based OV-3DS paradigms: 3D proposal-based, 2D proposal-based, and hybrid proposal-based approaches.

These approaches mostly focus on how to reliably project and fuse the masks and their features to produce 3D instances. They achieve this by focusing on various challenges arising from direct fusion such as multi-view instance inconsistency that occurs when using local metrics when merging consecutive frames, the imprecise segmentation boundaries due to noisy segmentation and projection steps, or overlooking semantic information during merging.

One approach [67] tackles the first challenge by proposing a global view consensus rate that merges 2D masks if the remaining masks indicate that they belong to the same 3D instance, whereas [68, 69] handle imprecise segmentation boundaries by leveraging 3D geometric priors or boundary information during the merging process. Subsequent works started to use semantic information such as enabling the merging process if the instances have the same semantic label [70] or through integrating both feature similarity and spatial overlap in the merging process [71].

Key Takeaways: 2D-based proposal segmentation methods generally maintain a two-stage pipeline that separates localisation from semantic classification, where the major changes revolve around enhancing the merging process of the 2D prediction masks. One network stands in particular [68] as it demonstrates the highest AP performance compared to the other 2D-based proposal networks, which could be attributed to using boundary affinity information during the merging process. Another notable direction appears in [71] as it utilises 3D geometric features during the merging process resulting in an improved segmentation performance. This shows an interesting research avenue to leverage the semantic coherence within point clouds to enhance the predicted 3D instances whilst making use of the knowledge within 2D FMs.

4.2.2 3D-based Proposal Network (Fig. 5 & Table 5)

This approach typically uses a closed set 3D instance proposal network followed by a feature computation module. These methods are either modular [75–79] or an end-to-end segmentation pipeline [80, 81]. Modular approaches follow a common pipeline that consists of a pre-trained 3D instance segmentation network [86] to predict class-agnostic 3D instance proposals. These proposals pass through a feature computation

Table 5 Comparison of proposal-based segmentation methods for open-vocabulary 3D segmentation. The table highlights object proposal generation, feature computation strategies, and architectural designs. The (X) indicates that the module is not being used. The methods are grouped by type: 2D-based proposal networks are unshaded, 3D-based proposal networks are grey, and hybrid methods are blue. Abbreviations: Ext: Extraction; View SEL: View selection; Arch: Architecture; SP: superpoints; G-SAM: Grounded-SAM. VAcc: Accelerated Visibility Computation; MVPDist: Multi-view Prompt Distribution.

Methods	Class-Agnostic Object Proposals		Feature Computation			Arch. Type
	3D Branch	2D Branch	View SEL	Crop Scale	Feature Ext.	
MaskClustering [67]	×	CropFormer [72]	Top-K	Multi	CLIP	Modular
OV-BIS [68]	×	SAM	Top-K	Multi	CLIP	Modular
OV-MAP [69]	×	CropFormer+SP	Top-1	Single	CLIP	Modular
FreeMask3D [70]	×	OneFormer	×	Single	CLIP	Modular
OpenSU3D [73]	×	G-SAM+GPT-4V	Top-K	Multi	CLIP	Modular
VoxelFlow [71]	×	G-SAM	Top-λ	Multi	SigLIP [74]	Modular
OpenMask3D [75]	Mask3D	×	Top-K	Multi	CLIP	Modular
Open-YOLO 3D [76]	Mask3D	×	VAcc Top-K	Single	MVPDist	Modular
Search3D [77]	Mask3D	×	Top-K	Multi	SigLIP	Modular
OpenIns3D [78]	Mask3D	×	Rendered K	Single	ODISE	Modular
UniSeg3D [79]	UniSeg3D	×	Top-λ	Multi	CLIP	Modular
SOLE [80]	Mask3D	×	×	×	Mask3D+OpenSeg	End-to-End
MarIns3D [81]	Mask3D	×	×	×	Mask3D+OpenSeg	End-to-End
Dual-Path [82]	Mask3D	G-SAM	Top-K	Both	CLIP	Modular
Open3DIS [83]	ISBNet	G-SAM+SP	Top-λ	Multi	CLIP	Modular
OV3D-CG [84]	Mask3D	SAM+SP	Top-1	×	LLM	Modular
Jung et al. [85]	Mask3D/ISBNet	G-SAM+SP	Top-K	Multi	Alpha-CLIP	Modular

module which includes back-projection, top-k view selection, and ending with computing and assigning open-vocabulary features to the corresponding 3D instance. This provides the flexibility of changing components as better networks are released.

The development of modular 3D-based proposal network methods started with OpenMask3D [75] as the first zero-shot open-vocabulary 3D instance segmentation approach. Subsequent works target specific limitations such as the high computational demands of using SAM and CLIP during feature computation. This led to approaches that utilise a more efficient detection-based feature computation module [76]. Other works focus on specific use cases such as [78] as it focuses on performing open-vocabulary 3D instance segmentation using only a point cloud input if the RGB-D images are not available. It utilises a rendering-based pipeline as a replacement to those images for feature extraction. Unlike the previous approaches, some methods [77] revolve around establishing a more fine-grained understanding of the scene such as part-level or material-level understanding, while others [79] construct a unified 3D scene understanding approach that is capable of performing six segmentation tasks including OV-3DS without relying on task-specific modules.

Alternatively, end-to-end approaches [80, 81] target a limitation apparent in modular methods, where they predict mask proposals without training the network using open-vocabulary features. Thus, end-to-end networks explore this integration to improve mask generation resulting in high-quality mask proposals. More specifically, they integrate open-vocabulary features with Mask3D features before feeding it into a cross-modality decoder resulting in enhanced performance.

Key Takeaways: 3D proposal-based segmentation networks follow the same two-stage approach as 2D-based methods with the main difference being the use of a 3D backbone for mask proposal prediction. These approaches focus on enhancing the quality of the proposed masks, the feature computation module, or building an end-to-end framework. One prominent design choice is the use of a pre-trained 3D instance segmentation network to provide the class-agnostic instance proposals followed by a multi-view image-driven feature computation module for semantic recognition. The former presents a major bottleneck as the performance is constrained by the closed set nature of the mask proposals. A promising direction is exploring the fusion of 2D-based and 3D-based proposal networks to handle predicting familiar and novel object categories. Another potential avenue is leveraging a multi-view detection-based feature computation approach, explored in [76], as it provides a more efficient alternative compared to SAM-based methods whilst providing strong segmentation performance.

4.2.3 Hybrid Proposal Network (Table 5 & Fig. 5)

This type leverages a closed set 3D network and an OV-2DS approach for mask proposal predictions, followed by a feature computation module to associate language-aligned features with the 3D instances. The motivation behind using both modalities is to build a more complete set of class-agnostic mask proposals covering familiar, novel, and geometrically ambiguous objects. These methods can be categorised as early-fusion [83–85], or late-fusion approaches [82] depending on when the proposals and features are integrated.

The earlier works explore both early-fusion [83] and late-fusion approaches [82]. The former starts by predicting and concatenating class-agnostic 3D proposals with 2D-guided 3D instance proposals before the feature computation module, while the latter treats the 2D and 3D branches independently as they obtain mask proposals and their corresponding features before integrating them together. Subsequent works mostly focus on early-fusion methods by either enhancing the object recognition step using contextual information [84], or combining and refining concepts from multiple methodologies to enhance the overall segmentation performance [85].

Key Takeaways: Hybrid approaches utilise both 2D and 3D modalities to increase the recall of mask proposal predictions, where they utilise filtering techniques to remove erroneous and redundant information to maintain the proposal precision. Throughout these works, several studies offer insights for improving the overall performance. The first takeaway, discussed in [83], concerns the 2D branch merging process, where they empirically show that hierarchical agglomerative clustering outperforms sequential Hungarian merging. The second, shown in [82], emphasises the importance of a carefully designed approach to integrate the 2D and 3D branches to ensure the removal of low-quality and redundant proposals. The third, proposed in [85], shows that the use of Alpha-CLIP during feature computation provides better object-centric representation, enhancing the classification performance.

5 Neural Radiance Field-based OV-3DS (RQ 2)

This section examines OV-3DS methods that use NeRF as their rendering-optimised 3D representation. These approaches typically follow a distillation-based pipeline.

Multimodal knowledge distillation uses 2D foundation models and distils the knowledge into a 3D representation using rendering-based training, enabling novel view synthesis with semantic understanding. Initial research [87, 88] encodes closed set semantics into NeRFs, which is the basis of its open-vocabulary capabilities. Distillation-based methods are categorised based on the type of information integrated into NeRF such as semantic-only or semantic-instance integration.

NeRF is a rendering-optimised representation that uses dense image-space supervision for scene understanding, which may benefit reasoning about partially occluded objects when sufficient multi-view observations are available. NeRFs benefit from their continuous representation that enforces multi-view consistency. However, they struggle due to the computational training and rendering costs of the ray-marching volumetric rendering approach. This section discusses the core principles behind the methodologies and their overall trend.

5.1 Multimodal Knowledge Distillation

5.1.1 Semantic-only Integration (Fig. 6 & Table 6)

These approaches focus on embedding language information into NeRF with the help of 2D open-vocabulary networks to enable open-vocabulary querying [89–98]. They follow a common pipeline which extracts dense open-vocabulary features and use them to supervise a NeRF with an additional feature branch. Feature extraction follows two

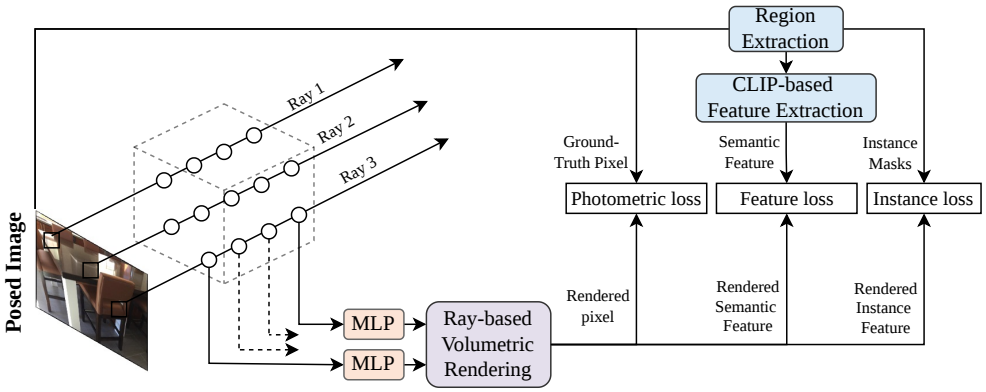


Fig. 6 This figure summarises NeRF-based KD training strategy. It distils knowledge from 2D foundation models into NeRF using ray-based rendering, using either modular or unified networks for region and feature extraction.

primary paradigms. The first leverages pre-trained open-vocabulary 2D segmentation (OV-2DS) networks to provide the supervision signal, or a two-stage approach that extracts masks followed by feature assignment.

Early works [89–91] explore distilling 2D-OVS knowledge into a NeRF by incorporating a feature branch, enabling text-driven semantic segmentation. Subsequent works address one or more challenges to enhance their segmentation performance.

One fundamental challenge is its computational inefficiency due to its ray-based volumetric rendering as it samples many points along the ray to predict values for each pixel, which becomes even more challenging as NeRF rely on per-scene optimisation. There are various approaches to handle this challenge such as modifying the positional encoding of the original NeRF with more efficient alternatives such as multi-resolution hash encoding [90] or a hybrid positional encoding approach [91]. The other alternative is modifying the original representation of NeRF with tensoRF [92, 95] or by using an occupancy representation [96]. Alternatively, there are approaches that move away from the per-scene optimisation approach altogether and construct a 3D volume for a more generalisable scene-agnostic framework [94].

There are other limitations that are directly related to the distillation of open-vocabulary 2D features into NeRF. For example, the imprecise segmentation boundaries which are inherited from the coarse nature of CLIP heatmap predictions. Several methods target this limitation by introducing modules to enhance spatial boundaries such as SAM and DINO, or by providing a 3D geometric prior [98]. Another inherited limitation is multi-view feature inconsistency, which is addressed using various contributions. For instance, some methods [97] associate same-object masks using a tracking module before feature aggregation, whereas [95] uses the inherent 3D consistency of NeRFs, while [94, 96] propose a more selective multi-view feature aggregation scheme.

One approach [99] takes a different direction by reframing the training mechanism through proposing a semantic loss that is directly implemented on the sampled 3D points along the ray opposed to the traditional 2D-based rendering loss.

Table 6 Comparison of NeRF-based knowledge distillation approaches for open-vocabulary 3D scene understanding. The table summarises the teacher network and training configuration. The (×) indicates that the module or training strategy is not being used. In the case of the "Joint" column, the symbol indicates whether the fields are trained concurrently (√) or sequentially (×). The methods are grouped by type: semantic-only methods are unshaded, and semantic-instance methods are grey. Abbreviations: NV: Novel Views; Rad: Radiance; Sem: Semantic; Ins: Instance; Ext: Extraction.

Methods	**Teacher Network**			**Training Configuration**				
	2D Vision			NV	Fields			Joint
	Region Ext.	**Feature Ext.**	**Unified Ext.**		**Rad**	**Sem**	**Ins**	
DFF [89]	×	×	LSeg	×	√	√	×	√
VL-Fields [90]	×	×	LSeg	×	√	√	×	√
NIVLFF [91]	×	×	LSeg	×	√	√	×	√
3D-OVS [92]	Sliding window	CLIP+DINO	×	×	√	√	×	×
Engelmann et al. [93]	×	×	OpenSeg	√	√	√	×	√
GOV-NeSF [94]	×	×	LSeg	√	√	√	×	√
OV-NeRF [95]	SAM	OpenCLIP+DINO	×	√	√	√	×	√
OpenOcc [96]	×	×	OpenSeg	×	√	√	×	√
Zhang et al. [97]	G-DINO+SAM	OpenCLIP	LSeg	×	√	√	×	√
SPNeRF [98]	Ray-based cropping	OpenCLIP	×	×	√	√	×	×
Open3DRF [99]	SAM	OpenCLIP+DINO	×	×	√	√	×	√
Cues3D [100]	×	×	FC-CLIP	×	√	×	√	√
PVLFF [101]	SAM	×	LSeg	×	√	√	√	√
PanopticRecon [102]	Grounded-SAM	DINOv2	×	×	√	√	√	√
OpenObj [103]	CropFormer	CLIP	×	×	√	√	×	√
LE-Object [104]	YOLO-World [105]+SAM	CLIP	×	×	√	√	×	√

Key Takeaways: The progression of semantic-only NeRF-based OV-3DS primarily focuses on three challenges: The computational efficiency of NeRF training and explicitly leveraging modules to enhance boundary segmentation precision and multi-view feature consistency. One recent approach [99] challenges the current 2D-based training paradigm by shifting the focus towards 3D point-based training, resulting in less floaters. Additionally, leveraging novel view synthesis is an under-explored approach and remains a promising avenue for enhancing segmentation performance.

5.1.2 Semantic-Instance Integration (Fig. 6 & Table 6)

These methods integrate semantic and instance information into a NeRF by either integrating instance-level information first followed by instance-level semantic association [100], or jointly training both [101, 102], or by training object-centric NeRFs that train one representation per object [103, 104]. The first two focus on enhancing multi-view instance consistency, while the latter focuses on improving semantic granularity.

Constructing NeRF-based OV-3DS methods using both semantic and instance information can take various directions. One approach [101] follows a contrastive learning training scheme, whereas [102] follows a two-stage reconstruction framework, while [100] leverages the inherent 3D consistency within NeRF using a three-stage training pipeline to ensure multi-view instance consistency.

On the other hand, some methods [103, 104] follow an object-centric approach that starts by extracting object-level open-vocabulary features, followed by multi-view instance association, and ending with training one NeRF per object, inspired by [106], enabling fine-grained understanding.

Key Takeaways: Semantic-instance integration methods vary in how they incorporate semantic and instance information into the neural radiance field. One direction focuses on integrating instance-level information into the NeRF before associating language-aligned semantic information. The second direction jointly trains semantic and instance feature fields directly [101] or through using a multi-step approach [102]. The third direction follows a multi-step approach to train one NeRF per object. Even though these approaches integrate instance-level information into the representation, they focus on different objectives. For example, instance-first and joint feature field methods enable panoptic scene understanding, while object-level NeRF approaches emphasise fine-grained understanding for tasks such as robotic manipulation.

6 3D Gaussian Splatting-based OV-3DS (RQ 3)

This section discusses OV-3DS approaches that utilise Gaussian Splatting as their rendering-optimised 3D representation. These methods are categorised into: multi-modal KD and training-free approaches [107–110].

Multimodal knowledge distillation uses 2D foundation models and distils their knowledge into 3D Gaussians by optimising their internal parameters for geometric reconstruction and semantic understanding. Achieving that means that the 3DGS has been adapted to include semantic and/or instance parameters. On the other hand,

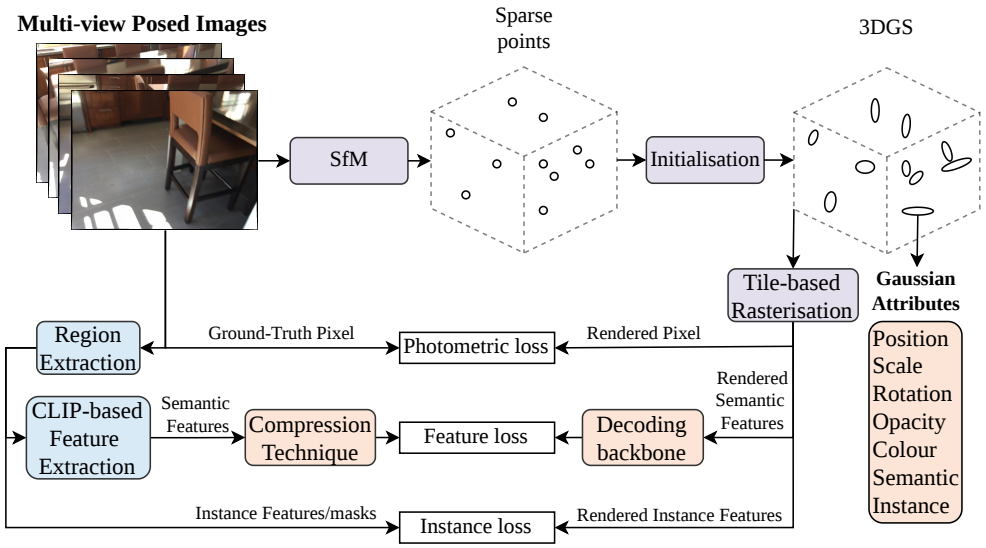


Fig. 7 This figure summarises 3DGS-based KD training strategy. It distils knowledge from 2D foundation models into 3DGS through tile-based rasterisation, using modular or unified networks for region and feature extraction.

training-free approaches associate or back-project semantic features into the 3DGS representation without iterative scene-level optimisation.

Unlike NeRFs, 3DGS is a rendering-based explicit representation that tackles NeRFs' rendering inefficiency by proposing a faster rendering method, but it comes with multiple trade-offs such as high memory consumption, and an imbalance of how each Gaussian represents appearance and semantic information. Similar to NeRFs, 3DGS has limited capabilities when dealing with object boundaries and scenes with sparse input views. In this section, we discuss the fundamental concepts and trends of distillation-based and training-free methods.

6.1 Multimodal Knowledge Distillation

6.1.1 Semantic-only Integration (Fig. 7 & Table 7)

These approaches focus on 3DGS-based methods that directly embed semantic features into the 3DGS representation, or utilise an external parametric feature field [111, 112] to enable semantic-level understanding. They follow a common pipeline that extracts language-aligned semantic features from multi-view RGB images and use those features to optimise a semantic parameter in the 3DGS representation. This feature extraction step either uses an OV-2DS network or a two-stage approach that extracts segmentation masks followed by a CLIP feature assignment. Due to the high-dimensional CLIP features, semantic-only approaches employ feature-modelling techniques to handle the features efficiently.

The development of 3DGS-based OV-3DS methods witnessed a dramatic increase due to their fast rendering and explicit representation. Early approaches [113–115] transfer 2D features to 3DGS through feature supervision alongside RGB images, enabling open-vocabulary feature rendering. Subsequent methods address limitations arising from integrating semantic information into the 3DGS representation.

One major limitation is the memory cost of assigning CLIP features to each Gaussian, prompting various open-vocabulary feature compression techniques. This includes using a scene-specific autoencoder, dictionary lookup, semantic labelling, feature grid mapping and retrieval, student decoding, or modelling the features using an external parametric feature-field. These techniques are essential for preserving semantic information while reducing the memory footprint to enable its integration into 3DGS. A common approach is training a scene-specific autoencoder to map the open-vocabulary features into a low-dimensional continuous space for training, followed by using a decoder to reconstruct them during inference. Alternatively, other approaches utilise a dictionary lookup-based approach that either embeds low-dimensional Gaussian features that points to the corresponding learned and quantised semantic entry in a codebook [113, 116], or an object ID that retrieves the uncompressed CLIP feature [117, 118]. Furthermore, some approaches [114, 119, 120] replace compressing CLIP features by constructing a trainable student decoder to upsample the low-dimensional feature maps embedded into the 3DGS to match the uncompressed open-vocabulary features. On another note, there are uncommon approaches to handle CLIP features such as creating semantic label masks from those features to act as supervision signals which improves rasterisation efficiency, constructing a semantic feature grid map that maps a low-dimensional feature to the uncompressed multi-view CLIP features of an object, or utilising an external parametric semantic feature-field, specifically a triplane representation [121] with an MLP, which is more memory-efficient than storing uncompressed CLIP features in each Gaussian without compromising performance.

Another limitation is the appearance-semantic representation imbalance that occurs due to the nature of 3DGS tile-based rendering approach. More specifically, the current α-blending strategy is designed for colour information. However, adapting it directly to rendered semantic features results in ambiguous semantics since each Gaussian can only represent one semantic label. One approach [122] disentangles colour and language information at the rasteriser level. It uses a semantic-aware rasteriser with an α-blending scheme that integrates semantic information using a novel trainable attribute instead of colour opacity whilst sharing position and covariance parameters.

The third challenge is multi-view semantic inconsistency inherited from 2D predictions, resulting in conflicting supervision signals. There are multiple solutions to handle this limitation. Some approaches [118, 123] use low-dimensional parameters to retrieve uncompressed multi-view consistent CLIP features, whereas [124] utilises regularisation techniques during optimisation to ensure feature consistency, while [125] regularises the semantic features before fusing the multi-view features into a 3D contextual space. This space is then used to train a scene-specific autoencoder.

Many existing approaches assume the availability of dense input images, resulting in suboptimal reconstruction performance under sparse-view conditions. Several solutions arise to mitigate this issue. For instance, some methods [120, 126] handle this

by introducing novel views using a video generation module, while [119] uses rendered views during training, whereas [112] leverages language-aware depth distillation to handle the geometric degradation. In contrast, one notable method [118] bypasses the issue by using a pre-trained feed-forward network to obtain the Gaussian parameters.

Key Takeaways: The advancement of semantic-only integration approaches exposes multiple major weaknesses in directly embedding semantic features into the 3DGS representation. These limitations include the high memory cost when assigning CLIP features to each Gaussian, appearance-semantic representation imbalance, and multi-view semantic inconsistency. Other research efforts target under-explored directions such as dynamic scene understanding [127], optimisable feature querying approach [116], or the use of caption-based features, demonstrated in [117, 127], to obtain enhanced or temporally consistent features.

6.1.2 Semantic-Instance Integration (Fig. 7 & Table 7)

These approaches integrate both semantic and instance information into the 3DGS framework. They either follow a semantic-first scheme [130], instance-first [131–142], or jointly training both types of information [143]. Semantic-first methods can be viewed as a conceptual development of semantic-only methods with a focus on enabling instance-level understanding. On the other hand, instance-first approaches follow a common pipeline as they embed low-dimensional instance features and train them using SAM-derived segmentation masks, followed by a language-aligned semantic association. This avoids storing the high-dimensional CLIP features in each Gaussian as they are associated to the object instances after optimisation. Finally, joint methods utilise both information to ensure object-consistent language information.

Semantic-first integration approaches adopt the pipeline of semantic-only methods as they start with language modelling, followed by instance differentiation to enable instance-level understanding. One approach [130] proposes a panoptic scene understanding framework that utilises a tri-plane representation for language modelling, followed by a graph clustering-based approach for separating individual instances.

The development of instance-first methods is a response toward the difficulty of embedding and compressing CLIP features in 3DGS parameters, resulting in suboptimal performances. Early approaches [131, 132] focus on embedding low-dimensional identity parameters into the 3DGS to group same-entity Gaussians together, followed by semantic association. This is followed by many approaches that handle various limitations such as the appearance-semantic representation imbalance. They target this challenge by leveraging a more structured anchor-based representation [134, 137, 140], or through decoupling semantic from appearance opacity [139].

Another limitation is the multi-view feature inconsistency apparent in the supervision signal. Even though approaches embed identity parameters to each Gaussian, some provide an explicit contribution to enhance feature consistency. For instance, one approach [133] leverages contrastive learning and a regularisation loss, whereas [142] proposes a feature aggregation module to alleviate multi-view feature inconsistency. In contrast, [141] challenges the idea of view-dependent semantic aggregation since some views might hold meaningful information while others provide ambiguous information.

Table 7 Comparison of 3DGS-based distillation methods for open-vocabulary 3D scene understanding. The table summarises the teacher network, feature modelling, and the training configuration. The components for 4D LangSplat cover only its time-agnostic field. The (X) indicates that the module or training strategy is not being used. The methods are grouped by type: semantic-only methods are unshaded, semantic-instance methods are grey, and referring-only methods are blue. Abbreviations: Ext: Extraction; HD: High-dimensional; Geo: Geometric; Sem: Semantic; Ins: Instance; G-SAM: Grounded-SAM.

Methods	Teacher Network			Feature Modelling	Training Configuration			
	2D Vision				Fields			Joint
	Mask Source	**Feature Ext.**	**Unified Ext.**		**Geo.**	**Sem.**	**Ins.**	
LEGaussians [113]	Sliding Window	CLIP+DINO	×	Dictionary Lookup	√	√	×	√
Feature-3DGS [114]	×	×	LSeg	Student decoding	√	√	×	√
LangSplat [115]	SAM	OpenCLIP	×	Autoencoder	√	√	×	×
VaF-LangSplat [128]	SAM+SAM2	OpenCLIP	×	Autoencoder	√	√	×	√
Peng et al. [122]	SAM	OpenCLIP	×	Autoencoder	√	√	×	√
GOI [116]	×	×	APE	Dictionary Lookup	√	√	×	×
OpenMIGS [117]	CropFormer	CLIP+SBERT	×	Dictionary Lookup	√	√	×	√
N2F2 [111]	SAM	CLIP	×	Feature-field	√	√	×	×
FastLGS [123]	SAM	OpenCLIP	×	Grid Mapping	√	√	×	√
4D LangSplat [127]	SAM+DEVA [129]	OpenCLIP	×	Autoencoder	√	√	×	×
EconSG [125]	SAM	×	OpenSeg	Autoencoder	√	√	×	√
SOVGaussian [112]	SAM	CLIP	×	Autoencoder	√	√	×	√
SpContrast [119]	SAM 2	CLIP	×	Student decoding	√	√	×	√
SPC-GS [120]	SAM 2	CLIP	×	Student decoding	√	√	×	√
SLGaussian [118]	SAM+DEVA	CLIP	×	Dictionary Lookup	√	√	×	×
LangScene-X [126]	SAM 2	×	FC-CLIP	Autoencoder	√	√	×	×
CLIP-GS [124]	SAM+DEVA	CLIP	×	Semantic Labelling	√	√	×	√
PanoGS [130]	×	×	LSeg	Feature-field	√	√	×	×
OpenGaussian [131]	SAM	×	×	Codebook	√	×	√	×
Gaussian Grouping [132]	SAM+DEVA	×	×	N/A	√	×	√	√
OpenSplat3D [133]	SAM	×	×	N/A	√	×	√	√
InstanceGaussian [134]	SAM	×	×	N/A	√	×	√	√
DCSeg [135]	SAM	×	×	N/A	√	×	√	×
OVSG-SLAM [136]	SAM	×	×	N/A	√	×	√	√
ObjectGS [137]	DEVA+G-SAM	×	×	N/A	√	×	√	√
VoteSplat [138]	SAM	×	×	N/A	√	×	√	√
MaskGaussians [139]	SAM	×	×	N/A	√	×	√	×
AG2aussian [140]	SAM	×	×	N/A	√	×	√	√
LaGa [141]	SAM	×	×	N/A	√	×	√	×
OpenInsGaussian [142]	SAM	×	×	N/A	√	×	√	×
ILGS [143]	SAM+DEVA	CLIP	×	Autoencoder	√	√	√	√
ReferSplat [144]	G-SAM	×	×	N/A	√	√	×	×

As these methods develop, complementary directions address specific use cases. One approach [135] follows a modular instance-first scheme, whereas another [138] utilises instance-level centroid information as a replacement for learned instance features for a more efficient instance grouping.

Unlike semantic-first and instance-first methods, joint integration approaches train identity and language information together. One example [143] explores joint training to ensure multi-view semantic consistency. It utilises SAM-based segmentation masks to optimise the identity parameter, and CLIP features to optimise semantic features, resulting in a 3DGS representation with semantic and instance-level understanding.

Key Takeaways: Semantic-instance integration approaches address the same challenge of integrating high-dimensional CLIP features into 3DGS following a different approach. They typically integrate instance-level information instead of compressed CLIP features before providing the language-aligned knowledge afterwards. One promising direction is exploring joint integration methods as it is an under-explored strategy and refining semantic and instance features using bidirectional feature interaction to improve the segmentation performance.

6.1.3 Referring-only Integration (Table 7)

These approaches target the task of segmenting occluded objects given a text description with object-level spatial relations or attributes. This research direction is still under-explored as there is very few works targeting this task. One example is [144] as it proposes a new dataset for this task in 3DGS representation, and introduces Refer-Splat framework, with an embedded referring feature parameter in each Gaussian, capable of referring 3D Gaussian splatting segmentation.

6.2 Training-free

These methods differ from distillation-based approaches as they assign the 2D open-vocabulary semantic features to the corresponding Gaussians in an already reconstructed 3DGS. One of the main advantages of training-free segmentation approaches is that it facilitates direct segmentation without having to rely on scene-specific optimisation. These approaches commonly use SAM-based CLIP feature extraction to obtain the open-vocabulary features before diverging on how they perform the feature association. Some approaches rely on a codebook-based registration approach [107], while the others use a back-projection strategy to associate the 2D features with their corresponding Gaussians [108–110].

7 Cross-Representation Analysis

The development of representation-specific OV-3DS methods leverages the strengths of each representation while inheriting its limitations. A thorough understanding of these representations provides deeper insight into which challenges arise from the data representation itself and which stem from methodological design choices. Table 8 presents a comparative summary of the different representation types, highlighting their key characteristics, strengths, and weaknesses.

Table 8 Cross-representation analysis between point cloud, NeRF, and 3DGS.

	Point Cloud	**NeRF**	**3DGS**
Spatial Unit	Discrete explicit samples	Continuous implicit field	Overlapping explicit Gaussians
Training Paradigm	Varies	Ray-based Rendering	Tile-based Rendering
Semantic Storage	Backbone weights	Scene-specific MLP	Per-Gaussian Parameters
Key Strength	Explicit Surface Geometry	Compact storage	High-quality, fast rendering
Key Limitation	Limited Coverage	Expensive Rendering	High storage requirements, Semantic Ambiguity

The structural differences between 3D representations have distinct implications for geometric coverage, object boundaries, and instance-level understanding. The geometric coverage of point clouds largely depends on data acquisition, whereas 3DGS employs adaptive densification to capture finer details. Unlike discrete point clouds with direct point-level assignments, 3DGS may exhibit semantic and instance ambiguities due to overlapping Gaussians, as each Gaussian can influence multiple pixels during tile-based rasterisation, resulting in suboptimal object boundary delineation. In contrast, NeRFs lack explicit spatial units, making geometric precision, boundary localisation and instance-level understanding dependent on the learned implicit field.

The selection of an appropriate 3D scene representation depends on the requirements of downstream applications. Point clouds are naturally suited to surface-based understanding, making them suitable for tasks such as object detection and segmentation. Although 3DGS employs explicit primitives, they are optimised for photorealism and do not necessarily align with physical surfaces. 3DGS and NeRF-based methods are suited for appearance-oriented applications such as novel view synthesis and photorealistic rendering. In fact, 3DGS typically offers faster rendering, making it preferable for real-time applications, whereas NeRF provides more compact storage, making it preferable for storage-constrained applications.

8 Datasets and Evaluation Protocols (RQ 4)

The evaluation of open-vocabulary approaches is essential to assess the generalisability of models and ensure standardised comparisons across methods. This section discusses the relevant datasets, evaluation metrics, and the evaluation settings for OV-3DS.

8.1 OV-3DS Datasets

This section discusses the 3D scene understanding datasets followed by multi-view 2D open-vocabulary datasets used to evaluate the performance of OV-3DS networks. A summary of these datasets is provided in Tables 9 and 10 respectively.

Table 9 Commonly used indoor 3D datasets for OV-3DS. Abbreviations: Sem: Semantic; Comm: Commodity; Env: Environment; Syn: Synthetic.

Dataset	Year	# Scenes	RGB Sensors		Depth Sensors	# Sem. Classes	Env.
			Comm.	DSLR			
S3DIS [145]	2016	5	√	×	Matterport scanner	12	Real
Matterport3D [146]	2017	90	√	×	Matterport scanner	40	Real
ScanNet [147]	2017	1513	√	×	Occipital structure sensor	20	Real
Replica [148]	2019	18	√	×	Custom RGB-D rig	88	Real
HyperSim [149]	2021	461	N/A	N/A	N/A	40	Syn.
ScanNet200 [150]	2022	1513	√	×	Occipital structure sensor	200	Real
MultiScan [151]	2022	117	√	×	iPhone LiDAR	419	Real
ScanNet++ [152]	2023	460	√	√	iPhone LiDAR + Faro Focus	>1000	Real

3D Scene Understanding Datasets: We provide an overview of each 3D dataset that has been used during the performance evaluation of open-vocabulary methods.

Stanford 3D Indoor Scene (S3DIS) dataset [145] comprises 5 large-scale areas annotated with 12 semantic classes and a "clutter" class. Matterport3D dataset [146] contains 90 building-scale scenes, where each scene is 2.61 floors on average, making it suitable for long-range scene understanding. ScanNet dataset [147] includes 1513 scenes with 20 classes including an "other furniture" class. ScanNet200 dataset [150] is an extended benchmark of ScanNet, which introduces a larger vocabulary with 200 categories. Replica dataset [148] comprises 18 high-resolution scenes with 88 semantic categories. HyperSim [149] is a synthetic 3D indoor dataset that consists of 461 scenes with 40 labelled classes with semantic and instance-level annotations. MultiScan dataset [151] includes 117 scenes with 419 fine-grained semantic categories with object and part-level annotations. This diversity enables various tasks such as object/part instance segmentation and motion estimation. Finally, ScanNet++ [152] is a large-scale dataset that consists of high-resolution 3D reconstruction and DSLR images, an RGB-D video obtained from an iPhone, with semantic and instance annotations. This dataset enables a range of tasks such as 3D semantic segmentation, 3D instance segmentation, and novel view synthesis.

2D Multi-view Datasets: We discuss 2D multi-view datasets, originally proposed for novel view synthesis, and adapted for open-vocabulary segmentation.

Table 10 Commonly used multi-view 2D datasets for OV-3DS. The text queries column reports the average number of queries per scene in the dataset. Mixed indicates the presence of indoor and outdoor scenes in the dataset. Abbreviations: Env: Environment; Sens: Sensors.

Dataset	Year	# Scenes	RGB Sens.	Depth Sens.	Text Queries (queries/scene)	Env.
3D-OVS [92]	2023	10	√	×	6.1	Mixed
Dove-G [153]	2023	8	√	√	250	Indoor
LERF-OVS [115]	2024	4	√	×	16.5	Indoor
LERF-Mask [132]	2024	3	√	×	7.67	Indoor
Mip-NeRF360-OVS [116]	2024	4	√	×	4	Mixed
Mip-NeRF360-OVS [113]	2024	6	√	×	9.17	Mixed
HyperNeRF-OVS [127]	2025	6	√	×	6	Indoor
Neu3D-OVS [127]	2025	6	√	×	6	Indoor

3D-OVS dataset [92] consists of 10 scenes with segmentation mask annotations and their corresponding text queries. Dove-G dataset [153] contains 8 scenes, where each scene includes 50 language queries. They expand those queries to 250 using an LLM. LERF dataset [154] comprises 13 scenes captured using an iPhone camera with the Polycam application. It only provides 3D localisation annotations for five scenes. LERF-OVS [115] is an expanded subset of LERF dataset that provides ground-truth segmentation masks with their textual queries for four scenes. They also provide additional challenging localisation examples for better evaluation. LERF-Mask [132] is another expanded subset of LERF dataset that provides accurate segmentation masks for three scenes. There are two mip-NeRF360-OVS datasets [113, 116] that adapt mip-NeRF 360 dataset [155] for open-vocabulary evaluation by providing manually annotated segmentation masks. Finally, HyperNeRF-OVS and Neu3D-OVS datasets [127] are adapted from [156, 157] by providing manually annotated semantic segmentation masks for dynamic OV-3DS.

8.2 Evaluation Settings

8.2.1 Point-based 3D Segmentation

The performance evaluation of OV-3DS networks uses closed set 3D scene understanding datasets due to the lack of a standard 3D open-vocabulary benchmark. The evaluation protocols are as follows:

- **Closed set**: This setting is commonly used to evaluate annotation-free KD approaches, as they do not rely on ground-truth labels during training. Performance is reported across all classes in the dataset.
- **ZSL and GZSL**: These settings are used to evaluate a network trained on a subset of classes on unseen classes only (ZSL) or on seen and unseen classes (GZSL).
- **Long-tail**: This protocol groups classes according to label frequency (e.g., head, common, and tail classes) to assess performance across imbalanced category distributions, as commonly performed on the ScanNet200 dataset.
- **Oracle mask**: This setting is specific to proposal-based methods and evaluates the feature computation module using ground-truth 3D instance masks.

- **Cross-dataset**: This evaluation setting assesses the generalisation capability of a network by training it on one dataset and evaluating it on a different dataset.

8.2.2 Rendering-based 3D Segmentation

The performance evaluation of rendering-based OV-3DS networks is conducted in 2D or 3D space. They can be summarised as follows:

3D-based Evaluation: This evaluation protocol uses 3D datasets (e.g., ScanNet, Replica) to assess segmentation performance against ground-truth 3D annotations.

- **Direct Querying**: Uses a 3D position for similarity-based label assignment [93].
- **Render-then-project**: Renders multi-view features and back-projects them into 3D space for evaluation [93].

2D-based Evaluation: This evaluation protocol uses multi-view 2D datasets or the 2D views of 3D datasets to assess segmentation performance under three settings.

- **3D Object Selection**: Selects relevant 3D Gaussians using cosine similarity between language and query features, followed by rendering for 2D evaluation [131].
- **Rendering-first protocol**: Renders relevancy or semantic maps for segmentation evaluation [92, 95, 115].
- **Observer-based protocol**: Extracts 2D mask proposals using a reference image and a text query. Then, it selects 3D instances using an intersection-over-area threshold, then renders their 2D masks for evaluation [132, 133].

8.3 Pitfalls of Comparative OV-3DS Analysis

Curating a comparative performance overview of OV-3DS methods has to follow strict guidelines to ensure a fair and reproducible study. The first key factor toward ensuring a fair comparison is that each network follows a fundamentally similar pipeline. For instance, comparing distillation-based and proposal-based methods can be tricky since the former is training-based while the latter is typically training-free.

After establishing the framework's fundamental pipeline, it is crucial to ensure that they follow the same evaluation protocol. This includes the same: training/evaluation datasets, base/novel class split, foundation model, prompt engineering during inference, evaluation metric (overall, or, foreground only), ground-truth versus predicted mask proposals (applicable to proposal-based methods), and 2D versus 3D evaluation.

Rendering-optimised methods can be evaluated in either 2D or 3D space, Therefore, fair comparison requires consistent dimensionality, evaluation protocols, and 2D-specific constraints, including the scenes and views used for optimisation and evaluation, and the distinction between scene-specific and scene-agnostic networks.

9 Challenges and Future Directions

Scalability: A key component of OVL approaches is the use of foundation models, which enable understanding diverse object categories. However, the lack of large-scale 3D-text datasets forces 3D scene understanding methods to rely on 2D counterparts and computationally expensive back-projection pipelines. Therefore, constructing

large-scale 3D vision-language datasets is crucial for advancing more general and scalable OV-3DS networks. One notable approach [29] tackles this by proposing the Mosaic3D-5.6M dataset that curates multiple 3D indoor scene understanding datasets and uses an automatic annotation approach to generate 5.6 million <3D mask-text> pairs across 30k scenes. Even though this enables a strong foundation to build robust 3D VLMs, it still does not compare to its 2D counterpart which leverages web-scale datasets. This requires a deep exploration of online sources to curate 3D and 2D scene video data that are in the public domain and can be utilised for network training.

Semantic Generalisation: Transferring semantic knowledge from 2D foundation models to 3D faces various limitations due to the domain shift between modalities.

One apparent limitation in captioning-based distillation methods is the inclusion of errors and noise due to their reliance on a 2D-based fully automatic module for point-text pairs generation. This can be alleviated through exploring semi-automatic pipelines with a human-in-the-loop strategy to ensure that the generated pairs are vocabulary-consistent and granularity-consistent across the dataset.

Another fundamental challenge is the current focus of OV-2DS networks on object-level scene understanding. This makes grasping the fine-grained information of each object difficult, hindering the development toward applications such as robotic manipulation. One promising direction is tackling granular semantic ambiguity by training a network capable of understanding coarse and fine-grained object-level information [103] to develop multi-granular 3D scene understanding frameworks.

Many approaches transfer the 2D knowledge into the 3D space without taking into account the uncertainty arising from the chosen 2D network's outputs and their back-projection. These predictions are view-dependent and can be imprecise depending on the network chosen, in addition to being subject to projection errors. One potential research direction is modelling this uncertainty into the training stage of the network to ensure more robust predictions [57, 113].

Geometric and Temporal Generalisation: Most OV-3DS methods focus primarily on static environments with dense input views, while considerably less attention has been given to developing approaches capable of handling limited input views and modelling dynamic real-world 3D scenes.

Rendering-optimised OV-3DS networks struggle with sparse input views as it directly affects scene-specific optimisation, resulting in suboptimal performance. Although various 3DGS-based methods introduce different modules to address this limitation, the emergence of feed-forward networks provides an exciting avenue toward bypassing this limitation toward more general 3DGS-based OV-3DS networks [118].

Modelling temporal information within 3D scene understanding frameworks is an underexplored domain. One existing approach [127] models dynamic scenes using 4D Gaussian splatting with both time-invariant and time-varying semantic fields. Advancing this area could enable more practical scene understanding systems for real-world applications. A promising direction is exploring techniques to encode time-aware object information (e.g. scene graphs [158]) across different 3D representations to enable dynamic scene understanding using pre-acquired data or using a real-time data stream.

Computational Efficiency: Building computationally efficient methods is a crucial research avenue for deploying scene understanding frameworks in real-world scenarios.

Distillation-based networks suffer from long training periods as they use high-dimensional pixel features from foundation models as pseudo-labels. A notable direction is exploring more efficient pseudo-label representations such as using semantic probability as a distillation target to improve computational efficiency [48].

On the other hand, proposal-based segmentation methods suffer from slow inference speeds compared to point cloud-based distillation methods. Exploring computationally efficient techniques is crucial toward applications for low-resource scenarios. One direction could be leveraging detection-based feature computation [76] to speed up the overall segmentation process, whereas using VLM-based 3D object recognition [159] can assist in avoiding the entire projection process.

Many NeRF-based approaches rely on preprocessing to enhance multi-view feature consistency. An underexplored opportunity is utilising NeRFs' inherent capability to enforce this consistency during training without any preprocessing steps [100].

Evaluation and Reproducibility: Standardised benchmarks are fundamental for understanding the open-vocabulary 3D segmentation landscape. This includes representation-specific, and cross-representation evaluation benchmarks.

Point cloud-based OV-3DS methods are categorised into distillation-based and proposal-based approaches, where there is limited comparative assessment between both tracks. Thus, exploring these two categories can help establish a common benchmark to evaluate each component and address weaknesses more effectively.

The performance assessment of rendering-optimised segmentation methods suffers from the use of a limited number of scenes leading to less generalisable models. In addition, the evaluation protocol can be inconsistent as the evaluation can be done in the 2D or 3D space with various settings. Tackling this challenge can be rewarding as establishing a standard benchmark to evaluate the segmentation performance will provide a foundation for fair and reproducible comparisons.

On the whole, OV-3DS methods typically evaluate their performance by comparing it with same-representation baselines. One exciting research avenue is exploring and developing a cross-representation performance assessment that can provide empirical evidence of the benefits of each 3D representation.

The development of OV-3DS methods relies on transferring information from pre-trained 2D networks, blurring the line when evaluating the performance regarding vocabulary generalisation and visual-domain generalisation due to the domain shift that occurs in the process. An interesting research avenue is exploring how to decouple the performance assessment to target each aspect individually. The common single-dataset performance assessment approach focuses on vocabulary generalisation, whereas evaluating the network on multiple datasets on a fixed set of classes targets visual-domain generalisation. This ensures the presence of different data distributions while keeping the vocabulary constant.

10 Conclusion

This systematic review provides a comprehensive analysis of indoor OV-3DS methods across various 3D representations. These methods generally leverage 2D VLMs to distil

or adapt their knowledge for 3D scene understanding. Achieving this transfer can provide substantial advantages in spatial understanding, which can provide meaningful contributions for various industrial sectors. Throughout this review, We analytically examine OV-3DS methods, resulting in a unified taxonomy and a comparative technical analysis of the different approaches. This paper also provides a cross-paper synthesis to have a broader understanding of the main pipelines and evaluation settings utilised across the domain. We conclude this work by providing challenges and future directions to guide the community toward the advancement of OV-3DS.

Acknowledgements This research is supported by Edith Cowan University.

Declarations

Author Contributions HA has designed, conducted the study and written the paper under the supervision of SS. All authors were involved in the review and revision of the manuscript.

Data Availability This review analyses existing literature, which is appropriately cited in the manuscript. No separate dataset was generated during the study.

Funding No funding was received to assist with the preparation of this manuscript.

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

[1] Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., *et al.*: Towards open vocabulary learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 5092–5113 (2024)

[2] Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2002–2010 (2017)

[3] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186 (2019)

[4] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., *et al.*: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916 (2021). PMLR

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[6] Lei, Y., Wang, Z., Chen, F., Wang, G., Wang, P., Yang, Y.: Recent Advances in Multi-modal 3D Scene Understanding: A Comprehensive Survey and Evaluation (2023)

[7] Zhu, C., Chen, L.: A survey on open-vocabulary detection and segmentation: Past, present, and future. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 8954–8975 (2024)

[8] Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5625–5644 (2024)

[9] Sapkota, R., Roumeliotis, K.I., Cheppally, R.H., Calero, M.F., Karkee, M.: A Review of 3D Object Detection with Vision-Language Models (2025)

[10] Yu, X., Wang, Y., Sun, H.: Open-vocabulary object detection survey. Procedia Computer Science **266**, 62–71 (2025)

[11] Nguyen, T.-A.-Q., Bourki, A., Macudzinski, M., Brunel, A., Bennamoun, M.: Semantically-aware neural radiance fields for visual scene understanding: A comprehensive review. International Journal of Computer Vision **134**(3), 109 (2026)

[12] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. bmj **372** (2021)

[13] Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., *et al.*: Foundation models defining a new era in vision: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2245–2264 (2025)

[14] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., *et al.*: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision, pp. 38–55 (2024). Springer

[15] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., et al.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks (2024)

[16] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)

[17] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)

[18] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., *et al.*: Segment anything.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

[19] Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)

[20] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

[21] Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (2023)

[22] Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., *et al.*: A review of generalized zero-shot learning methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4051–4070 (2022)

[23] Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7010–7019 (2023)

[24] Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Lowis3d: Language-driven open-world instance-level 3d scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 8517–8533 (2024)

[25] Yang, J., Ding, R., Deng, W., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19823–19832 (2024)

[26] Wang, Y., Czarnecki, K.: Aide: Improving 3d open-vocabulary semantic segmentation by aligned vision-language learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2674–2685 (2025). IEEE

[27] Liu, K., Liu, Y.-J., Chen, B.: General 3d vision-language model with fast rendering and pre-training vision-language alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[28] Jiang, L., Shi, S., Schiele, B.: Open-vocabulary 3d semantic segmentation with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21284–21294 (2024)

[29] Lee, J., Park, C., Choe, J., Wang, Y.-C.F., Kautz, J., *et al.*: Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14089–14101 (2025)

[30] Wang, Y., Jia, B., Zhu, Z., Huang, S.: Masked point-entity contrast for open-vocabulary 3d scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14125–14136 (2025)

[31] Zhang, S., Zhang, S., Deng, J., Shen, Y., Ma, M., Zhang, Y.: Pgov3d: Open-vocabulary 3d semantic segmentation with partial-to-global curriculum. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 5070–5079 (2025)

[32] Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision, pp. 350–368 (2022). Springer

[33] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., *et al.*: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning, pp. 23318–23340 (2022). PMLR

[34] Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., *et al.*: Grit: A generative region-to-text transformer for object understanding. In: European Conference on Computer Vision, pp. 207–224 (2024). Springer

[35] Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., *et al.*: Grounding multimodal large language models to the world. In: The International Conference on Learning Representations (2024)

[36] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., *et al.*: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36**, 19769–19782 (2023)

[37] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306 (2024)

[38] Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., *et al.*: Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In: European Conference on Computer Vision, pp. 289–310 (2024). Springer

[39] Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., *et al.*: Osprey: Pixel understanding with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28202–28211 (2024)

[40] Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., *et al.*: What if we recaption billions of web images with llama-3? In: International Conference on Machine Learning, pp. 35957–35976 (2025). PMLR

[41] Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.:

Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 815–824 (2023)

[42] Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision, pp. 540–557 (2022). Springer

[43] Han, X.-F., Wang, C., Wang, Y., Wang, M.: Open-world 3d scene understanding with cross-modal dual consistency learning. In: Proceedings of the International Conference on Multimedia Retrieval, pp. 416–423 (2025)

[44] Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., *et al.*: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7020–7030 (2023)

[45] Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision, pp. 696–712 (2022). Springer

[46] Wang, P., Wang, Y., Li, S., Zhang, Z., Lei, Z., Zhang, L.: Open vocabulary 3d scene understanding via geometry guided self-distillation. In: European Conference on Computer Vision, pp. 442–460 (2024). Springer

[47] Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022)

[48] Shen, L., Wei, X., Pan, G., Zheng, Q., Cao, Y.: Cost-efficient open vocabulary 3d scene understanding based on semantic probability. IEEE Transactions on Image Processing (2025)

[49] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., *et al.*: Open-vocabulary semantic segmentation with mask-adapted clip. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7061–7070 (2023). IEEE

[50] Zhu, X., Zhou, H., Xing, P., Zhao, L., Xu, H., *et al.*: Open-vocabulary 3d semantic segmentation with text-to-image diffusion models. In: European Conference on Computer Vision, pp. 357–375 (2024). Springer

[51] Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2955–2966 (2023)

[52] Zhang, J., Dong, R., Ma, K.: Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In: Proceedings of the IEEE/CVF International

Conference on Computer Vision, pp. 2048–2059 (2023)

[53] Yu, F., Tian, R., Wang, Z., Wang, X., Liang, X.: Cus3d: Clip-based unsupervised 3d segmentation via object-level denoise. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2024). IEEE

[54] Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)

[55] Long, F., Yao, T., Qiu, Z., Li, L., Mei, T.: Pointclustering: Unsupervised point cloud pre-training using transformation invariance in clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21824–21834 (2023)

[56] Wu, Y., Han, X.-F., Xiao, G.: Language-driven open-vocabulary 3d semantic segmentation with knowledge distillation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3320–3324 (2024). IEEE

[57] Li, J., Saltori, C., Poiesi, F., Sebe, N.: Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 19390–19400 (2025)

[58] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)

[59] Li, Z., Lu, J., Deng, J., Chang, H., Wu, L., *et al.*: Sas: Segment any 3d scene with integrated 2d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8306–8318 (2025)

[60] Wang, Z., Wang, Y., Yu, X., Zhou, J., Lu, J.: Xmask3d: Cross-modal mask reasoning for open vocabulary 3d semantic segmentation. Advances in Neural Information Processing Systems **37**, 74922–74944 (2024)

[61] Wang, J., Wang, Z., Miyazaki, T., Fan, Y., Omachi, S.: Tamc: Textual alignment and masked consistency for open-vocabulary 3d scene understanding. Sensors **24**(19), 6166 (2024)

[62] Li, R., Zhang, Z., He, C., Ma, Z., Patel, V.M., Zhang, L.: Dense multimodal alignment for open-vocabulary 3d scene understanding. In: European Conference on Computer Vision, pp. 416–434 (2024). Springer

[63] He, Q., Peng, J., Jiang, Z., Wu, K., Ji, X., *et al.*: Unim-ov3d: uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation. In: Proceedings of the Thirty-Third International Joint Conference on Artificial

Intelligence, pp. 812–820 (2024)

[64] Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. In: European Conference on Computer Vision, pp. 131–147 (2024). Springer

[65] Yu, Q., He, J., Deng, X., Shen, X., Chen, L.-C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems **36**, 32215–32234 (2023)

[66] Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., et al.: Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following (2023)

[67] Yan, M., Zhang, J., Zhu, Y., Wang, H.: Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28274–28284 (2024)

[68] Yi, T., Wang, S., Zhang, Z., Wang, C., Yan, D., et al.: Ov-bis: Open-vocabulary boundary guide zero-shot 3d instance segmentation. IEEE Transactions on Multimedia (2025)

[69] Kim, J., Park, Y., Yoon, H.-J., Zhang, B.-T.: Ov-map: Open-vocabulary zero-shot 3d instance segmentation map for robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 13780–13786 (2024). IEEE

[70] Zhou, M., Wu, X., He, C., Wang, R., Chen, X.: Freemask3d: Zero-shot point cloud instance segmentation without 3d training. IEEE Robotics and Automation Letters (2025)

[71] Wang, C., Chen, H., Ma, J., Han, T., Chen, Y.: Voxelflow: 2d semantic mask-guided voxel flow for open-vocabulary 3d instance segmentation. In: 2025 International Conference on Cyberworlds (CW), pp. 01–08 (2025). IEEE

[72] Qi, L., Kuen, J., Shen, T., Gu, J., Li, W., *et al.*: High quality entity segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4047–4056 (2023)

[73] Mohiuddin, R., Prakhya, S.M., Collins, F., Liu, Z., Borrmann, A.: Opensu3d: Open world 3d scene understanding using foundation models. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 13560–13566 (2025). IEEE

[74] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on

Computer Vision, pp. 11975–11986 (2023)

[75] Takmaz, A., Fedele, E., Sumner, R., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. Advances in Neural Information Processing Systems **36**, 68367–68390 (2023)

[76] Boudjoghra, M.E.A., Dai, A., Lahoud, J., Cholakkal, H., Anwer, R.M., *et al.*: Open-YOLO 3d: Towards fast and accurate open-vocabulary 3d instance segmentation. In: The International Conference on Learning Representations (2025)

[77] Takmaz, A., Delitzas, A., Sumner, R.W., Engelmann, F., Wald, J., Tombari, F.: Search3d: Hierarchical open-vocabulary 3d segmentation. IEEE Robotics and Automation Letters (2025)

[78] Huang, Z., Wu, X., Chen, X., Zhao, H., Zhu, L., Lasenby, J.: Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In: European Conference on Computer Vision, pp. 169–185 (2024). Springer

[79] Xu, W., Shi, C., Tu, S., Zhou, X., Liang, D., Bai, X.: A unified framework for 3d scene understanding. Advances in Neural Information Processing Systems **37**, 59468–59490 (2024)

[80] Lee, S., Zhao, Y., Lee, G.H.: Segment any 3d object with language. In: The International Conference on Learning Representations (2025)

[81] Li, H., Su, J., Zhou, D., Cao, M.: Marins3d: An open-vocabulary 3d instance segmentation model with mask refinement. Neurocomputing, 131018 (2025)

[82] Ton, T., Hong, J.W., Eom, S., Shim, J.Y., Kim, J., Yoo, C.D.: Zero-shot dual-path integration framework for open-vocabulary 3d instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7598–7607 (2024)

[83] Nguyen, P., Ngo, T.D., Kalogerakis, E., Gan, C., Tran, A., *et al.*: Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4018–4028 (2024)

[84] Zhou, M., He, C., Wang, R., Chen, X.: Ov3d-cg: Open-vocabulary 3d instance segmentation with contextual guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5305–5314 (2025)

[85] Jung, S., Zheng, J., Zhang, K., Qiao, N., Chen, A.Y., *et al.*: Details matter for indoor open-vocabulary 3d instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9637 (2025)

[86] Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 8216–8223 (2023). IEEE

[87] Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15838–15847 (2021)

[88] Blomqvist, K., Ott, L., Chung, J.J., Siegwart, R.: Baking in the feature: Accelerating volumetric segmentation by rendering feature maps. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 7629–7634 (2023). IEEE

[89] Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. Advances in neural information processing systems **35**, 23311–23330 (2022)

[90] Tsagkas, N., {Mac Aodha}, O., Lu, C.: Vl-fields: Towards language-grounded neural implicit spatial representations. In: Workshop on Representations, Abstractions, and Priors for Robot Learning Workshop at International Conference on Robotics and Automation (2023)

[91] Blomqvist, K., Milano, F., Chung, J.J., Ott, L., Siegwart, R.: Neural implicit vision-language feature fields. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1313–1318 (2023). IEEE

[92] Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., *et al.*: Weakly supervised 3d open-vocabulary segmentation. Advances in Neural Information Processing Systems **36**, 53433–53456 (2023)

[93] Engelmann, F., Manhardt, F., Niemeyer, M., Tateno, K., Tombari, F.: OpenneRF: Open set 3d neural scene segmentation with pixel-wise features and rendered novel views. In: The International Conference on Learning Representations (2024)

[94] Wang, Y., Chen, H., Lee, G.H.: Gov-nesf: Generalizable open-vocabulary neural semantic fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20443–20453 (2024)

[95] Liao, G., Zhou, K., Bao, Z., Liu, K., Li, Q.: Ov-nerf: Open-vocabulary neural radiance fields with vision and language foundation models for 3d semantic understanding. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[96] Jiang, H., Xu, Y., Zeng, Y., Xu, H., Zhang, W., *et al.*: Openocc: Open vocabulary

3d scene reconstruction via occupancy representation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1145–1152 (2024). IEEE

[97] Zhang, H., Li, F., Ahuja, N.: Open-nerf: Towards open vocabulary nerf decomposition. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3444–3453 (2024). IEEE Computer Society

[98] Hu, W., Parodi, N., Zepp, M., Feldmann, I., Schreer, O., Eisert, P.: Spnerf: Open vocabulary 3d neural scene segmentation with superpoints. In: Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 669–676. SCITEPRESS - Science and Technology Publications, ??? (2025)

[99] Lee, H., Yun, Y., Bae, J., Kim, S., Uh, Y.: Rethinking open-vocabulary segmentation of radiance fields in 3d space. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 4491–4498 (2025)

[100] Xue, F., Xu, W., Zhong, G., Ming, A., Sebe, N.: Cues3d: Unleashing the power of sole nerf for consistent and unique instances in open-vocabulary 3d panoptic segmentation. Information Fusion **122**, 103164 (2025)

[101] Chen, H., Blomqvist, K., Milano, F., Siegwart, R.: Panoptic vision-language feature fields. IEEE Robotics and Automation Letters **9**(3), 2144–2151 (2024)

[102] Yu, X., Liu, Y., Han, C., Mao, S., Zhou, S., *et al.*: Panopticrecon: Leverage open-vocabulary instance segmentation for zero-shot panoptic reconstruction. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 12947–12954 (2024). IEEE

[103] Deng, Y., Wang, J., Zhao, J., Dou, J., Yang, Y., Yue, Y.: Openobj: Open-vocabulary object-level neural radiance fields with fine-grained understanding. IEEE Robotics and Automation Letters **10**(1), 652–659 (2024)

[104] Wang, M., Zhang, Y., Wang, X., Zhang, Z., Li, Z.: Le-object: Language embedded object-level neural radiance fields for open-vocabulary scene. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 3183–3189 (2025). IEEE

[105] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16901–16911 (2024)

[106] Kong, X., Liu, S., Taher, M., Davison, A.J.: vmap: Vectorised object mapping for neural field slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 952–961 (2023)

[107] Jun-Seong, K., Kim, G., Yu-Ji, K., Wang, Y.-C.F., Choe, J., Oh, T.-H.: Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14137–14146 (2025)

[108] Dai, S., Qu, Y., Li, Z., Li, X., Zhang, S., Cao, L.: Training-free hierarchical scene understanding for gaussian splatting with superpoint graphs. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 3673–3682 (2025)

[109] Chacko, R., Häni, N., Khaliullin, E., Sun, L., Lee, D.: Lifting by gaussians: A simple, fast and flexible method for 3d instance segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3497–3507 (2025). IEEE

[110] Marrie, J., Ménégaux, R., Arbel, M., Larlus, D., Mairal, J.: Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7440–7450 (2025)

[111] Bhalgat, Y., Laina, I., Henriques, J.F., Zisserman, A., Vedaldi, A.: N2f2: Hierarchical scene understanding with nested neural feature fields. In: European Conference on Computer Vision, pp. 197–214 (2024). Springer

[112] Ling, P., Tan, T., Lin, J., Yang, W.: Sovgaussian: Sparse-view 3d gaussian splatting for open-vocabulary scene understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 5343–5351 (2025)

[113] Shi, J.-C., Wang, M., Duan, H.-B., Guan, S.-H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5333–5343 (2024)

[114] Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., *et al.*: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21676–21685 (2024)

[115] Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20051–20060 (2024)

[116] Qu, Y., Dai, S., Li, X., Lin, J., Cao, L., *et al.*: Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 5328–5337 (2024)

[117] Zhao, J., Wang, J., Deng, Y., Yue, Y.: Openmigs: Multi-granularity information-preserving open-vocabulary 3d gaussian splatting. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 21143–21150 (2025). IEEE

[118] Chen, K., Dai, B., Qin, M., Zhang, D., Li, P., *et al.*: Slgaussian: Fast language gaussian splatting in sparse views. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 3047–3056 (2025)

[119] Liao, G., Wang, A., Chen, M., Fang, Z.: Sparse-view 3d open-vocabulary gaussian splatting via collaborative contrastive learning. In: 2025 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2025). IEEE

[120] Liao, G., Li, Q., Bao, Z., Qiu, G., Liu, K.: Spc-gs: Gaussian splatting with semantic-prompt consistency for indoor open-world free-view synthesis from sparse inputs. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 11264–11274 (2025)

[121] Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12479–12488 (2023)

[122] Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., *et al.*: 3d vision-language gaussian splatting. In: The International Conference on Learning Representations (2025)

[123] Ji, Y., Zhu, H., Tang, J., Liu, W., Zhang, Z., *et al.*: Fastlgs: Speeding up language embedded gaussians with feature grid mapping. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 3922–3930 (2025)

[124] Liao, G., Li, J., Bao, Z., Ye, X., Li, Q., Liu, K.: Clip-gs: Clip-informed gaussian splatting for view-consistent 3d indoor semantic understanding. ACM Transactions on Multimedia Computing, Communications and Applications **21**(8), 1–24 (2025)

[125] Zhang, C., Lee, G.H.: econSG: Efficient and multi-view consistent open-vocabulary 3d semantic gaussians. In: The International Conference on Learning Representations (2025)

[126] Liu, F., Li, H., Chi, J., Wang, H., Yang, M., *et al.*: Langscene-x: Reconstruct generalizable 3d language-embedded scenes with trimap video diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 29010–29020 (2025)

[127] Li, W., Zhou, R., Zhou, J., Song, Y., Herter, J., *et al.*: 4d langsplat: 4d language gaussian splatting via multimodal large language models. In: Proceedings of the

Computer Vision and Pattern Recognition Conference, pp. 22001–22011 (2025)

[128] Li, C., Yang, X., Yang, W., Li, X.: Vaf-langsplat: Voxel-aware fusion language gaussian splatting. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 4952–4961 (2025)

[129] Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.-Y.: Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1316–1326 (2023)

[130] Zhai, H., Li, H., Li, Z., Pan, X., He, Y., Zhang, G.: Panogs: Gaussian-based panoptic segmentation for 3d open vocabulary scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14114–14124 (2025)

[131] Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., *et al.*: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. Advances in Neural Information Processing Systems **37**, 19114–19138 (2024)

[132] Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European Conference on Computer Vision, pp. 162–179 (2024). Springer

[133] Piekenbrinck, J., Schmidt, C., Hermans, A., Vaskevicius, N., Linder, T., Leibe, B.: Opensplat3d: Open-vocabulary 3d instance segmentation using gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 5246–5255 (2025)

[134] Li, H., Wu, Y., Meng, J., Gao, Q., Zhang, Z., *et al.*: Instancegaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14078–14088 (2025)

[135] Wiedmann, L., Wiehe, L., Rozenberszki, D.: Dcseg: Decoupled 3d open-set segmentation using gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 5217–5226 (2025)

[136] Liu, Z., Li, S., Huang, G., Wu, Y.: Ovsg-slam: Open-vocabulary semantic gaussian splatting slam. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 12804–12811 (2025). IEEE

[137] Zhu, R., Yu, M., Xu, L., Jiang, L., Li, Y., *et al.*: Objectgs: Object-aware scene reconstruction and scene understanding via gaussian splatting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8350–8360 (2025)

[138] Jiang, M., Jia, S., Gu, J., Lu, X., Zhu, G., *et al.*: Votesplat: Hough voting

gaussian splatting for 3d scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6456–6465 (2025)

[139] Sun, W., Li, Y., Jiao, J.: Maskgaussians: Advancing open-vocabulary 3d scene understanding with dynamic masking. In: Proceedings of the International Workshop on Application-driven Point Cloud Processing and 3D Vision, pp. 56–64 (2025)

[140] Wang, Z., Li, M., Tu, C.: Ag2aussian: Anchor-graph structured gaussian splatting for instance-level 3d scene understanding and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 26806–26816 (2025)

[141] Cen, J., Zhou, X., Fang, J., Wen, C., Xie, L., *et al.*: Tackling view-dependent semantics in 3d language gaussian splatting. In: International Conference on Machine Learning, pp. 7013–7034 (2025). PMLR

[142] Huang, T., Chen, R., Hu, D., Huang, F., Gong, M., Liu, T.: Openinsgaussian: Open-vocabulary instance gaussian segmentation with context-aware cross-view fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6341–6350 (2025)

[143] Jang, S., Kim, W.: Identity-aware language gaussian splatting for open-vocabulary 3d semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20467–20476 (2025)

[144] He, S., Jie, G., Wang, C., Zhou, Y., Hu, S., *et al.*: Refersplat: Referring segmentation in 3d gaussian splatting. In: International Conference on Machine Learning, pp. 22456–22467 (2025). PMLR

[145] Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., *et al.*: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534–1543 (2016)

[146] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., *et al.*: Matterport3d: Learning from rgb-d data in indoor environments. In: International Conference on 3D Vision (3DV), pp. 667–676 (2017). IEEE Computer Society

[147] 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)

[148] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., et al.: The Replica Dataset: A Digital Replica of Indoor Spaces (2019)

[149] Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., *et*

al.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10912–10922 (2021)

[150] Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: European Conference on Computer Vision, pp. 125–141 (2022). Springer

[151] Mao, Y., Zhang, Y., Jiang, H., Chang, A., Savva, M.: Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. Advances in Neural Information Processing Systems **35**, 9058–9071 (2022)

[152] Yeshwanth, C., Liu, Y.-C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12–22 (2023)

[153] Chang, H., Boyalakuntla, K., Lu, S., Cai, S., Jing, E.P., *et al.*: Context-aware entity grounding with open-vocabulary 3d scene graphs. In: Conference on Robot Learning, pp. 1950–1974 (2023). PMLR

[154] Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19729–19739 (2023)

[155] Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5470–5479 (2022)

[156] Park, K., Sinha, U., Hedman, P., Barron, J.T., *et al.*: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics (TOG) **40**(6), 1–12 (2021)

[157] Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., *et al.*: Neural 3d video synthesis from multi-view video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5521–5531 (2022)

[158] Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F.: Neural scene graphs for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2856–2865 (2021)

[159] Xu, Y., Luo, Z., Wang, Q., Kamat, V., Menassa, C.: Point2graph: An end-to-end point cloud-based 3d open-vocabulary scene graph for robot navigation. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2853–2860 (2025). IEEE