APSIPA Transactions on Signal and Information Processing, 2025, 14, e27

Original Paper

Research and Standards in 3D Scene Description Technologies: A Survey

Dong-shin Lim[1†], Dong-hun Lee[1†], Dong-hwi Kim[1], Jeong-hun Hong[1], Aro Kim[1], Chae-yeong Song[1], Bosung Baek[1], Dabin Kang[1], Myeong-jin Jang[1], Jinwoo Jeong[2], Sungjei Kim[3] and Sang-hyo Park[1*]

[1] *Kyungpook National University, South Korea*
[2] *Korea Electronics Technology Institute, South Korea*
[3] *Korea University of Technology and Education, South Korea*

ABSTRACT

The rapid expansion of the metaverse drives the growing demand for effective 3D scene description technologies, as well as augmented reality (AR) and virtual reality (VR) applications. However, a noticeable disconnect remains between academic research and industry-driven standardization efforts. While academic work often focuses on semantic richness, such as through the development of 3D scene graphs, industry standards prioritize interoperability, exemplified by the MPEG graphics language transmission format (i.e., glTF) extensions. This survey seeks to bridge this divide by systematically reviewing and comparing pivotal contributions from both academia and industry related to 3D scene description technologies and standardization efforts. Our analysis highlights notable differences in methodologies, data formats, and core objectives. Key challenges include the need for unified data representations and the establishment of standardized evaluation benchmarks to support broader integration. This survey emphasizes the urgent need for closer collaboration between academia and industry and proposes potential pathways towards a unified

*Corresponding author: s.park@knu.ac.kr. †: equal contribution

Received 04 June 2025; revised 18 July 2025; accepted 26 August 2025
ISSN 2048-7703; DOI 10.1561/116.20250038

framework to accelerate the real-world adoption of advanced 3D scene description technologies.

Keywords: 3D Scene Graphs, 3D Understanding, MPEG, scene description, file format

1 Introduction

The rapid advancement of technologies such as the metaverse, augmented reality (AR), virtual reality (VR), and digital twins has led to an unprecedented demand for rich, interactive 3D content. Developing immersive and realistic virtual environments that support dynamic user interaction requires advanced methods for representing and managing complex 3D scenes. Bridging the physical and virtual worlds effectively necessitates more than visually appealing models. This capability requires a comprehensive understanding of scene objects, including their properties, spatial configurations, and semantic relationships. Recent breakthroughs in generative artificial intelligence are further transforming 3D content creation by extending capabilities beyond traditional 2D media. At the core of this evolution are technologies collectively referred to as 3D scene description, which provide structured, extensible representations of 3D environments. These technologies encompass object geometry, appearance, behaviors, and interconnections. As a result, they enable content to adapt dynamically to user interactions and contextual variations.

3D scene description incorporates both structural and semantic approaches to modeling virtual environments. One key avenue involves efforts by industry standardization bodies, such as MPEG and the Khronos Group (creators of glTF). These bodies focus on defining efficient and interoperable formats for scene components, including node hierarchies, geometries, materials, and animations [19, 49]. Their emphasis lies in facilitating real-time rendering and seamless content exchange across platforms. In contrast, academic research often prioritizes deeper semantic modeling and concentrates on the explicit representation of objects, their attributes, and inter-object relationships. Within this realm, 3D Scene Graphs have gained prominence as a framework for encoding entities and their relational structures [54, 65, 29]. Despite these differing priorities, both industry and academic efforts contribute substantially to the core objectives of 3D scene description. These objectives include organizing complex spatial data, encoding object-level semantics, and supporting efficient rendering and adaptive interaction. These capabilities are vital for powering the next generation of responsive, immersive applications in AR, VR, and the metaverse.

The increasing demand for advanced 3D scene description technologies has attracted substantial attention from both academic researchers and standardization bodies (Figure 1 provides a conceptual overview). However, their approaches and goals often diverge markedly. Academic research typically emphasizes semantic richness and intricate relational modeling. This research often involves representations like 3D Scene Graphs that are derived from sensor inputs such as point clouds [54, 65, 29] or leverages AI-driven techniques to generate advanced scene representations [36]. In contrast, industry-led standardization efforts, spearheaded by organizations such as MPEG and Khronos, prioritize robust, efficient, and interoperable formats like glTF. These formats primarily define essential components (e.g., geometry, materials, and node hierarchies) tailored for real-time rendering pipelines [19, 29].

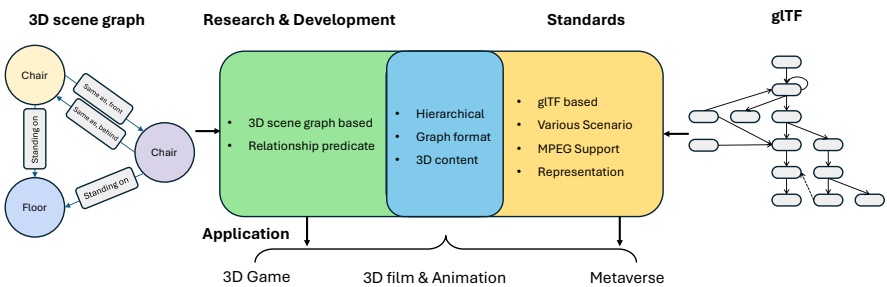

Figure 1: Overview of 3D scene description technology in research and standards.

Furthermore, industrial standards often integrate supporting technologies such as codecs and haptics, which typically fall outside the scope of academic investigations [19]. This divergence in objectives, data representations (e.g., relationship-based graphs versus rendering-optimized scene trees), and broader scope has limited meaningful interaction between academia and industry. As a result, the transition of research breakthroughs into widely adopted standards and applications remains slow. Given this gap, a comprehensive review that bridges these domains, evaluates their distinct approaches, and explores synergies is currently lacking. This article addresses this need by offering the following key contributions:

- A structured taxonomy that categorizes leading academic methods (including 3D Scene Graphs [54, 65], 3D Dense Captioning [12, 63], and Implicit Representations such as NeRF [37, 68]) alongside industry-driven standards such as MPEG's Scene description components and glTF extensions [19]. This taxonomy is illustrated conceptually in Figure 2.

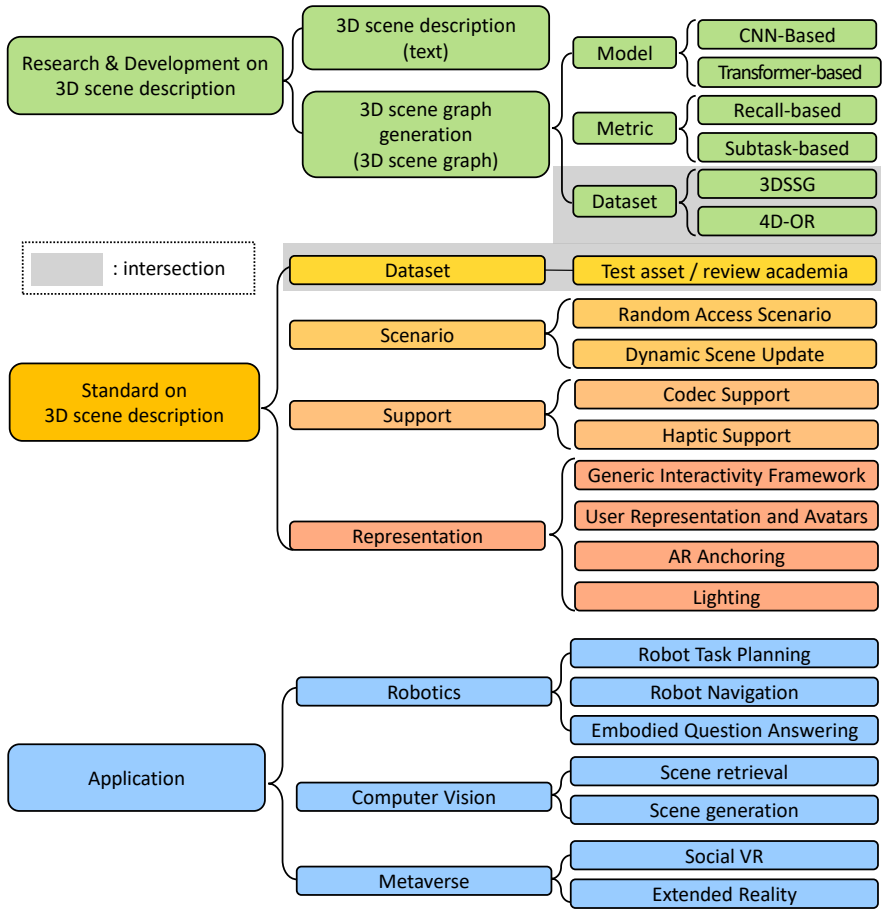

Figure 2: Taxonomy of 3D scene description technology in research, standards, and application.

- A comparative analysis highlighting the contrast in objectives (e.g., semantic depth vs. real-time efficiency), representational structures (e.g., graph-based vs. component-based models), and technical priorities within each domain.
- An identification of key barriers to integration, such as incompatible data formats, the absence of standardized datasets and benchmark applications for both research and commercial use, and the inherent difficulty of embedding rapidly evolving research into stable industrial standards.

- A call for closer collaboration between academia and industry, along with proposed future research directions designed to foster unified frameworks and facilitate the adoption of advanced 3D scene description methods within standardized pipelines.

2 Academic Research in 3D Scene Description

Academic research in 3D scene description primarily aims to achieve a comprehensive semantic understanding and support sophisticated interactions within 3D environments (Figure 1). The focus extends beyond basic geometric representations to encompass the semantics of objects, such as their attributes, functional properties, and the intricate network of inter-object relationships [54, 29]. Such semantic depth is essential to enable scene reasoning, context-aware interactions, and AI-driven content generation. This section reviews major academic advances and emphasizes core approaches such as explicit structural representations (particularly 3D Scene Graphs (3D-SGs)), text-based semantic descriptions (especially 3D Dense Captioning (3D-DC)), and emerging representations (such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3D-GS)) [38, 37, 68, 25, 15]. Table 1 offers a high-level comparison of these key techniques. We then explore foundational methodologies, notable studies, research challenges, benchmark datasets, and evaluation metrics associated with each domain.

Table 1: Comparison of key academic 3D scene description techniques.

Category	Input Modality	Representation	Strengths
3D-DC	PointCloud	BBox + Text	Rich captions
3D-SGs	PointCloud	Graph	Explicit relations
NeRF	RGB Images	Implicit MLP	Photorealistic views
3D-GS	SfM Cloud	Gaussian Splats	Real-time rendering

2.1 Text-based Descriptions: 3D Dense Captioning

3D Dense Captioning has become a prominent vision-language task in academic research. This task aims to automatically generate multiple detailed natural language descriptions that are explicitly grounded in objects within a 3D point cloud scene [63]. Unlike general scene description tasks, dense captioning requires simultaneous object localization (where 3D bounding boxes are typically used) and the generation of corresponding textual descriptions that capture object attributes and contextual relationships [12]. The input

is usually a raw 3D point cloud that encodes both geometric structure and appearance features. Most existing methods follow a standardized encoder-decoder pipeline that comprises three core stages [63]:

- Scene Encoder: Processes the input point cloud to generate object proposals and extract visual features. Common backbone architectures include PointNet++ [42] and VoteNet [40].
- Relation Module: Models the spatial and semantic relationships among detected objects to enrich context. Graph Convolutional Networks (GCNs) [12] and Transformer-based attention mechanisms [55, 4] are frequently utilized in this stage.
- Feature Decoder: Translates the enriched object features into natural language descriptions. This stage often employs sequential models, such as GRUs or Transformers, which are typically enhanced with attention mechanisms. Recent studies have also explored parallel architectures that integrate object detection and caption generation [8].

Recent advancements in this domain incorporate multimodal cues, such as semantic features from 2D images [64, 56] and contextual information beyond explicit object boundaries [66], to enhance the richness and precision of the generated captions. Benchmarking commonly uses datasets such as ScanRefer [6] and Nr3D [1], both of which are based on the ScanNet dataset [14]. Evaluation metrics include standard natural language generation scores (CIDEr [52], BLEU-4 [39], METEOR [2], and ROUGE [31]), often coupled with localization accuracy metrics based on Intersection-over-Union (IoU) thresholds. This evaluation approach reflects best practices in the field [63]. Although 3D Scene Graph captioning differs from 3D Dense generation, which focuses on structured graph representations, both tasks share the objective of deep semantic understanding of 3D scenes and often confront similar challenges in the processing of 3D data and the modeling of object relationships [63].

2.2 Explicit Graph-based Representations: 3D Scene Graphs

2.2.1 Definition, Goals, and Importance

A central focus in academic research on 3D scene understanding is the development and use of explicit graph-based representations, particularly 3D Scene Graphs (3D-SGs) [54, 29]. Unlike rendering-centric formats such as glTF, a 3D-SG is a structured semantic representation that maps the elements of a 3D scene, typically reconstructed from point clouds or meshes, onto a graph-based abstraction. In this graph, nodes represent object instances identified in the scene, and edges denote pairwise relationships among those objects [54, 65].

The primary objective of 3D-SGs is to support rich semantic interpretation beyond basic object detection or spatial configuration [29]. These graphs capture both Object-level Semantics (e.g., attributes and affordances) and Inter-Object Relationships (e.g., spatial, supportive, or comparative associations). Some models also incorporate hierarchical class labels derived from ontologies, such as WordNet [54].

The importance of 3D-SGs lies in their ability to provide a compact, structured, and semantically meaningful abstraction of complex environments. This explicit modeling is essential to enable higher-level reasoning tasks. 3D-SGs link low-level geometric data with symbolic representations and thereby offer a robust framework for AI systems to perceive, interpret, and interact with the 3D world in a more human-like manner [36, 29].

2.2.2 Input Data and Tasks

Research on 3D scene graph generation (3D-SGG) predominantly utilizes 3D point clouds as the primary input [54, 29]. These point clouds, often captured in real-world indoor settings via RGB-D sensors and reconstructed into meshes or aggregated scans [14, 53], typically include per-point data such as XYZ coordinates, RGB color, and occasionally surface normals [54].

A critical requirement for most current 3D-SGG methods is the availability of class-agnostic instance segmentation masks, where each point is pre-associated with a unique object instance ID [54, 29]. Although some methods incorporate supplementary modalities such as 2D images during the training phase [54], the standard inference process generally relies solely on the segmented point cloud. To assess model performance, the field borrows evaluation protocols from 2D scene graph generation (2D-SGG) literature [54]. The key tasks include:

- Predicate Classification (PredCls): Predicting relationships between known object pairs with ground-truth labels.
- Scene Graph Classification (SGCls): Predicting both object labels and their relationships.
- Scene Graph Generation (SGGen)/Relationship Detection (RelDet): Simultaneously detecting objects and inferring all relationships.

Among these, SGGen most closely reflects the practical challenge of achieving full-scene understanding directly from sensor data.

2.2.3 Methodologies and Architectural Evolution

The generation of 3D Scene Graphs (3D-SGs) from point clouds has primarily relied on deep learning approaches. These approaches adapt techniques

from 2D-SGG and general graph representation learning. For instance, initial efforts prominently employed graph convolutional networks (GCNs) [26].

GCN-based Approaches. One of the earliest and foundational works in this area, SGPN (Scene Graph Prediction Network) [54], established a baseline architecture to directly predict 3D-SGs from point clouds. SGPN leverages PointNet [41] to extract features from individual object instances (nodes) and spatial regions between object pairs (edges or relationships). These features are passed through a GCN to facilitate message passing. This allows for refined object representation and relationship inference [54]. However, this approach often prioritizes node classification over the modeling of relationships and potentially neglects the rich semantics of edges [65].

Subsequent research has sought to enhance this foundational framework. Scene Graph Fusion Network (SGFN) [57], for instance, incrementally constructs scene graphs from RGB-D sequences. To address underutilization of edge features, EdgeGCN (also known as SGGpoint) [65] introduces an edge-focused reasoning mechanism that explicitly models high-dimensional edge attributes and incorporates 'twinning interactions.' Meanwhile, Granular3D [18] introduces multi-granularity analysis to better handle complex and large-scale point clouds. Despite these improvements, GCN-based methods face inherent challenges such as over-smoothing [30, 36] and limited receptive fields, especially in sparse 3D environments [18]. These limitations have motivated the investigation of alternative architectures.

Transformer-based Approaches. To overcome the inherent limitations of GCNs, especially to capture long-range dependencies and global context, Transformer-based architectures [51] have been increasingly adopted for 3D-SG generation [36]. Transformers, known for their self-attention mechanism, are well-suited to model global relationships, which has prompted their adaptation for graph-structured 3D data.

An exemplary model is SGFormer [36], which utilizes Transformer layers as its architectural core. SGFormer incorporates novel modules tailored for the 3D-SGG task. These include the Graph Embedding Layer (GEL), which enables edge-aware self-attention, and the Semantic Injection Layer (SIL), which integrates external knowledge sources such as ChatGPT [43]. This design enhances the models ability to capture the global structure of 3D scenes and yields notable performance gains [36]. Nevertheless, challenges remain in the adaptation of sequential Transformers to irregular graph topologies and in the management of their computational demands [36, 29].

2.2.4 Addressing Key Challenges

Although architectural innovation has advanced 3D-SGG performance, several fundamental challenges persist. Ongoing research primarily focuses on

integrating external knowledge, addressing the long-tail distribution problem, and improving scalability and efficiency.

Knowledge Integration. Sole reliance on 3D geometric and visual features often proves insufficient to disambiguate complex semantic relationships [56, 36]. Researchers have explored various ways to inject external knowledge. These ways include approaches such as Visual-Linguistic Assisted Training (e.g., the VL-SAT scheme [56] that uses a multi-modal "oracle" model [43]), LLM-based Semantic Enhancement (e.g., in SGFormer [36] that leverages ChatGPT and CLIP embeddings), and Learned Knowledge Priors (e.g., methods that use graph auto-encoders or co-occurrence statistics [67, 29, 9]).

Long-Tail and Unbiased Learning. Similar to many real-world datasets, 3D-SGG datasets exhibit a significant long-tail distribution [56, 29, 36]. Knowledge integration strategies help mitigate this issue [56, 36]. Additionally, the employment of specialized loss functions such as Focal Loss [33], as used in SGFormer [36] and by Wald et al. [54], can help. Other techniques common in 2D unbiased SGG are potential future avenues [29, 48].

Scalability and Efficiency. The processing of large-scale 3D point clouds and reasoning over densely connected graphs pose significant computational challenges [36, 18]. Scalable and efficient solutions include Efficient Architectures such as Granular3D [18] and SGFormer's GEL [36], and efficient Point Cloud Backbones such as PointNet++ or RandLA-Net [17]. The achievement of both high accuracy and real-time performance remains an active area [18].

2.2.5 Relevant Datasets

The advancement and evaluation of 3D-SGG models depend heavily on specialized datasets that offer both 3D scene geometry and rich semantic annotations for objects and their interrelationships. Among these, the most widely used benchmark in point cloud-based 3D-SGG research is *3DSSG* [54]. It was introduced by Wald et al. [54] and constructed on top of the *3RScan* dataset [53]. 3RScan is a valuable resource that comprises ∼1500 multi-temporal scans of real-world indoor environments. These scans thereby capture scene variations over time. 3DSSG extends this foundation and incorporates comprehensive semantic scene graph annotations aligned with the reconstructed scenes [54]. Key characteristics of the 3DSSG dataset include:

- *Input Data:* It provides 3D point clouds derived from the reconstructed meshes, typically used alongside the corresponding class-agnostic instance segmentation masks available from 3RScan.
- *Rich Annotations:* It contains annotations for a large number of object instances (∼48k nodes) belonging to numerous semantic categories (∼160 object classes initially, often evaluated on subsets such as RIO27

[36]). Nodes are annotated with hierarchical class labels (using WordNet) and various attributes (static, dynamic, affordances).

- *Relationships:* It includes annotations for a diverse set of pairwise relationships (∼544k edges) covering spatial, support, and comparative types (∼26 predicate classes initially) [54].
- *Standard Benchmark:* Due to its scale and annotation richness, 3DSSG has become the standard benchmark for evaluating and comparing recent 3D-SGG methods developed for point clouds, as used in studies such as EdgeGCN [65], Granular3D [18], SGFormer [36], and VL-SAT [56].

The point cloud data in 3DSSG and similar studies often originates from large-scale reconstruction datasets such as ScanNet [14], which offers both geometric and semantic segmentation data widely used across 3D scene understanding tasks. Other datasets, such as ScanRefer [6] and Nr3D [1], although also based on ScanNet, are primarily tailored for 3D visual grounding and dense captioning, respectively [63]. For synthetic data-driven research, the SUNCG dataset [46] was employed in earlier 3D-SGG works [65].

2.2.6 Evaluation Metrics and Benchmarking

Evaluating and comparing the performance of different 3D scene graph generation models requires consistent use of well-defined evaluation metrics across standardized tasks: Predicate Classification (PredCls), Scene Graph Classification (SGCls), and the comprehensive Scene Graph Generation (SGGen) task [29]. These commonly employed metrics are summarized in Table 2.

- *Recall@K (R@K):* The most widely adopted metric for evaluating relationship triplet prediction [29, 54]. It measures the fraction of ground-truth triplets correctly predicted within the top-K predictions made by the model for a given scene. Common values for K are 20, 50, and 100. It is important to note potential variations in R@K calculation, such as whether graph constraints are applied (allowing only one predicate per object pair) and whether results are micro-averaged (across all triplets in the test set) or macro-averaged (averaging per-image recall scores) [29].
- *Mean Recall@K (mR@K):* Introduced to address the significant long-tail distribution issue prevalent in SGG datasets [29, 48]. Instead of averaging over all triplets, mR@K calculates Recall@K separately for each predicate class and then averages these per-class recall scores. This provides a less biased assessment of a model's ability to recognize both frequent and infrequent relationships [56, 36].

Table 2: Evaluation metrics for 3D scene graph generation.

Metric	Definition	Use Case
R@K	Fraction of GT triplets recovered in top-K	Overall relation recall
mR@K	Avg. Recall@K over all predicate classes	Mitigates long-tail issue
ZS R@K	Recall@K on unseen subject-predicate-object combinations	Tests generalization to novel relations
F1 Score	Harmonic mean of precision and recall	Balances precision and recall
mAP	Area under precision-recall curve	Standard detection metric

- *Zero-Shot Recall@K (ZS R@K):* Evaluates generalization capability by measuring recall only on relationship triplets (subject-predicate-object combinations) that were not observed during training [29, 35]. Particularly relevant for assessing robustness to the long-tail problem [36].
- *Accuracy@K (A@K):* Simple top-K accuracy sometimes reported for individual components such as object or predicate classification (especially in PredCls) [56, 18].
- *F1 Score:* Macro-averaged F1 score occasionally used alongside recall-based metrics due to class imbalance in predicate classification [65, 36].
- *Mean Average Precision (mAP):* While standard in object detection and some 2D-SGG benchmarks [29], mAP is less commonly reported for the primary 3DSSG dataset evaluations in the analyzed literature, potentially due to the difficulty of exhaustive relationship annotation [29].

Consistent application of these metrics, particularly R@K and mR@K, on standard benchmarks such as 3DSSG [54] allows for quantitative comparison. However, a remaining challenge is the lack of comprehensive, comparative benchmarking studies under uniform protocols and across different datasets [29].

2.3 Emerging Trends in 3D Scene Representation

Beyond explicit representations such as point clouds, meshes, and scene graphs, another significant direction in academic research explores *Implicit Neural Representations (INRs)* for modeling 3D scenes [58]. Among these, *Neural Radiance Fields (NeRF)* have garnered substantial attention [37].

2.3.1 Neural Radiance Fields (NeRF)

NeRF, introduced by Mildenhall et al. [37], represents a 3D scene implicitly via a multi-layer perceptron (MLP). This MLP maps a 3D spatial coordinate (x, y, z) and a viewing direction (θ, ϕ) to a volume density (σ) and view-dependent RGB color (c) [37, 58]. NeRF queries the MLP along camera rays and applies volume rendering techniques [24] to generate photorealistic novel views. This process uses only 2D input images with known camera poses [37].

Unlike discrete models, NeRF enables the compact and memory-efficient encoding of complex geometry and appearance and offers higher-fidelity reconstructions [58]. While initially developed for novel view synthesis, current research increasingly explores NeRFs for semantic scene understanding [28, 58]. For instance, Semantic-NeRF [68] augments the NeRF MLP to also predict semantic labels, while Panoptic-NeRF [16, 45] extends it to include instance-level segmentation. However, early methods typically operate on a per-scene basis and often struggle with boundary consistency and broader contextual understanding [28].

More recent methods, such as GP-NeRF [28], incorporate features from an advanced 2D segmentation network and leverage distillation techniques [28, 13] to address these limitations. Although the direct extraction of scene graphs from NeRF is still in its infancy, recent work has explored the integration of NeRF and scene graphs. For example, Structured-NeRF [38] employs a hierarchical scene graph to organize and optimize multiple object-level NeRFs. Similarly, SG-NeRF [10] uses scene graph optimization to enhance the robustness of NeRF-based surface reconstructions.

These studies reveal a promising synergy between implicit NeRF models and explicit graph structures. However, challenges persist in training efficiency, semantic editing, and in the extraction of structured representations for traditional scene understanding tasks. Addressing these challenges through hybrid approaches that combine the strengths of both implicit and explicit representations remains a critical direction for future research [58, 38].

2.3.2 3D Gaussian Splatting

Recent advances in 3D scene representation have introduced 3D Gaussian splatting (3D-GS) as a compelling alternative to fully implicit methods, such as NeRF [25, 15]. Unlike NeRFs volumetric ray-marching approach, 3D-GS uses an explicit representation composed of millions of learnable anisotropic 3D Gaussian distributions [25, 7]. This shift enables real-time rendering and enhanced editability and positions 3D-GS as a remarkable breakthrough in 3D reconstruction and rendering.

At its core, 3D-GS models scenes as a collection of anisotropic 3D Gaussians, each defined by a mean (position), covariance, opacity, and view-

dependent color, typically encoded with spherical harmonics [25]. Novel views are rendered through the projection of these 3D Gaussians onto a 2D image plane through a process called "splatting" [25]. The resulting 2D Gaussians are depth-sorted and composited with alpha blending. This process produces the final pixel colors. This rasterization-based approach is highly parallelizable and offers considerable speed advantages over NeRFs more computationally intensive ray-marching.

The training of a 3D-GS model involves the optimization of the parameters of each Gaussian to produce the input views accurately. This is generally achieved with stochastic gradient descent, guided by a combination of L1 loss and D-SSIM losses that compare the rendered images to the ground truth [25]. To maintain a positive semi-definite covariance matrix, it is parameterized with a learnable quaternion for rotation and a 3D vector for scale. A critical component of this training process is the adaptive control of Gaussian density, which involves iterative refinement achieved through cloning, splitting, and pruning Gaussians. Training often begins with a sparse point cloud, typically generated through Structure-from-Motion techniques [25, 7].

Recent research builds upon the foundational 3D-GS framework to explore its application in enhanced 3D scene understanding and to expand its use to new areas. Extensions include the enrichment of Gaussians with semantic, linguistic, instance-level, and spatiotemporal information. This enrichment enables tasks such as open-vocabulary querying and semantic segmentation [44, 69, 62, 5, 60]. Moreover, the integration of 3D-GS with structured representations, such as spatial MLPs or grids, has shown promise for high-fidelity human avatar modeling and dynamic scene reconstruction [59, 61, 11].

Although 3D-GS provides an explicit and richly attributed representation, its integration with structured semantic models, such as scene graphs, remains an open research challenge. While the object-centric nature of Gaussians offers a promising foundation, the extraction or embedding of complete scene graphs from optimized Gaussian sets is still under development. Nonetheless, this represents a promising direction for future work and potentially mirrors recent efforts by NeRF to utilize structured scene graph representations [38, 10].

In summary, 3D-GS provides a highly efficient and editable explicit representation for 3D scenes and delivers impressive real-time performance in novel view synthesis. Despite these strengths, several challenges remain. These include the management of high memory consumption in large-scale scenes, the achievement of accurate geometric reconstruction, the handling of complex lighting and reflective surfaces, and the enhancement of data efficiency for sparse input views [7, 15]. Ongoing research seeks to address these limitations. This research aims to develop memory-optimized methods, improve optimization algorithms, incorporate physical realism, and extend 3D-GS applications to dynamic environments and complex real-world scenarios [7, 15].

2.3.3 Benchmark and Evaluation Metrics in 3D scene representation

NeRF and 3D-GS utilize various benchmark datasets to quantitatively evaluate the performance of view synthesis that synthesizes images from new viewpoints. The characteristics of each data set used to measure performance of NeRF and 3D-GS are as follows.

- **MiP-NeRF360**[3]: A dataset proposed independently in the Mip-NeRF 360 paper and has been standardly used for performance comparison in various NeRF modification studies since then.
- **Tanks&Temples**[27]: Contains complex geometric structures of large-scale real-world scenes, and has been used to evaluate real-world performance in various NeRF studies.
- **DL3DV-10K**[34]: A large-scale real-world multi-view dataset, used for pre-training and generalization performance evaluation in Zip-NeRF and IBRNet.
- **Replica**[47]: A high-quality indoor scene dataset, often used for novel view synthesis and 3D reconstruction experiments in NeRF-VPT, Segment Anything in 3D with NeRFs, etc.
- **UrbanScene3D**[32]: Benchmark used in studies dealing with large/complex scenes at the city level.
- **ScanNet**[14]: Widely used as a standard benchmark in indoor scene 3D reconstruction and NeRF-based neural rendering studies.
- **NeRF-synthetic**[37]: A synthetic dataset proposed in the original NeRF paper, used as a standard synthetic benchmark in various NeRF and variant studies.

In terms of evaluation metrics, NeRF and 3D-GS commonly use the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), learned perceptual image patch similarity (LPIPS), multiscale SSIM (MS-SSIM) and mean squared error (MSE).

- **PSNR** : Measures the pixel-level difference between a synthetic image and a ground truth image to quantify the image quality.
- **SSIM and MS-SSIM**: Evaluate structural similarity to enable quality evaluation close to human vision.
- **LPIPS**: A deep learning-based image patch similarity index that shows a high correlation with subjective quality evaluation.

- **MSE** : The mean square error, which is the basic index that forms the basis for calculating PSNR.

Despite employing distinct rendering methodologies, NeRF and 3D-GS utilize the same benchmarks and metrics for objective performance comparison and research reproducibility. This indicates that the developmental trajectories of both technologies are not solely concentrated on enhancing the quality of novel scene synthesis, but also on fostering generalization and practical applicability across diverse environments. Furthermore, the presence of these shared criteria is crucial for ensuring consistency in performance evaluation and facilitating comparative analysis and advancement between technologies as new neural rendering techniques emerge in the future.

These datasets offer a range of scene characteristics and challenges, serving as tools for assessing the overall performance of models. However, their emphasis on rendering often results in a focus on scenes featuring a single object or centrally located objects, thereby neglecting background information. Furthermore, the majority of evaluation metrics primarily assess the presence or absence of noise based on visual quality discernible to the naked eye, leading to a deficiency in indicators that incorporate substantive semantic information.

3 Industry Standardization Efforts for 3D Scene Description

3.1 Introduction to Standardization

While academic research continues to expand the boundaries of semantic understanding and representational capabilities for 3D scenes, industry standardization plays a complementary yet critical role. Standards are essential to ensure interoperability across tools, platforms, and applications. This interoperability enables seamless exchange and rendering of 3D content within diverse ecosystems. They provide stable specifications that support real-time performance, which is particularly vital for applications such as AR, VR, and gaming. Moreover, standards promote broader industry adoption as they offer consistent development targets. Unlike academic efforts that often emphasize innovation and semantic richness, standardization prioritizes robustness, efficiency, and broad applicability in the definition of the core components of 3D scenes. This section reviews key standardization initiatives in 3D scene description, with a primary focus on the MPEG and the Khronos Groups glTF format, as outlined in publicly available documents and technical reports [19, 49]. A thorough understanding of these initiatives is crucial to align academic advancements with practical industry implementation and to unlock the full potential of next-generation 3D technologies.

3.2 Key Standardization Bodies and Formats

Several international standardization bodies are actively involved in defining formats and protocols for 3D scene description, aiming to foster an interoperable ecosystem. Among the most prominent efforts are those from the Moving Picture Experts Group (MPEG) and the Khronos Group. Key characteristics of these standardization efforts by MPEG and Khronos are summarized in Table 3.

Table 3: Overview of industry 3D scene description standards.

Standard	Format	Key Features
MPEG-I	Bin+JSON	Node hierarchy, Animation, Haptic, Codec ext
glTF	JSON+BIN	Mesh, Material, Texture, RT rendering ext

3.2.1 MPEG Scene Description (MPEG-I)

The Moving Picture Experts Group (MPEG), a working group of ISO/IEC, has significantly contributed to multimedia standards. As part of its MPEG-I (Immersive Media) initiative, MPEG developed specifications for scene description to support the growing demand for immersive content. The core standard in this effort is ISO/IEC 23090-14, Information technology Coded representation of immersive media Part 14: Scene Description [19]. A primary goal is to provide a framework that ensures consistency with traditional coded media streams. This enables integrated and synchronized immersive experiences. The standard defines fundamental constructs to build 3D scenes. These constructs feature elements such as node hierarchies, meshes, materials, cameras, and animations. It also aims for compatibility with existing scene description formats where practical [19].

Furthermore, the standard incorporates features designed to support dynamic and interactive immersive applications. These features encompass capabilities such as modular support and the ability to address random access to specific parts of the scene description data, as well as mechanisms to handle dynamic scene updates [22]. The standard explicitly considers the integration of related technologies critical for immersion, such as haptic feedback and efficient codec support (e.g., for V-PCC, MIV) for scene data [19, 23, 22]. MPEG's related activities also cover broader aspects such as the standardization of asset information handling and usage guidelines [21, 20]. Overall, MPEG's approach focuses on the provision of a comprehensive, integrated solution to represent and deliver complex, interactive immersive media experiences within the larger multimedia ecosystem.

3.2.2 Khronos Group glTF

Alongside MPEG, the Khronos Group plays a significant role in 3D asset standardization through its *glTF (Graphics Language Transmission Format)* specification [49]. Unlike MPEG's focus on integrating scene descriptions with coded media streams, *glTF* is primarily designed as an efficient, interoperable asset delivery format for 3D scenes and models, particularly targeting WebGL and other modern graphics APIs, aiming to minimize processing overhead for real-time rendering applications.

The core glTF specification defines a JSON-based structure describing the scene hierarchy, geometries, materials, textures, animations, and camera setups [49]. A key aspect is its *extensibility mechanism*, allowing for adding features beyond the core specification [50]. While Khronos manages the core specification and numerous extensions, MPEG has also actively developed glTF extensions within its MPEG-I framework, explored through dedicated Exploration Experiments (EEs) [23, 22]. Much of the standardization discussion relevant to advanced scene description functionalities revolves around these MPEG-developed extensions:

- *Lighting:* Extensions such as MPEG Texture-Based Lights and MPEG Punctual Lights add support for time-varying attributes and integrate physical and virtual light estimation, crucial for mixed reality [23].
- *Haptic Support:* The MPEG Haptic and MPEG Haptic Material extensions make it possible to define haptic objects and associate texture-based haptic data with 3D objects. This enables synchronized tactile feedback [23].
- *Generic Interactivity:* Extensions such as MPEG Scene Interactivity define a framework based on triggers and actions to manage interactive behaviors and can potentially leverage collision mesh definitions (MPEG Mesh Collision) [23].
- *Avatars:* The MPEG Node Avatar extension provides mechanisms to integrate animated user representations within the scene and links them with interaction and collision systems [23].
- *AR Anchoring:* The MPEG Anchor extension makes it possible to anchor glTF scenes to real-world elements with various reference space types for stable AR applications [23].
- *Codec Integration:* MPEG addresses the integration of volumetric media codecs such as V-PCC and MIV into glTF. This work defines processing pipeline options and specifies necessary signaling [22, 23].

Furthermore, MPEG has explored standardizing asset information descriptions and usage guidelines associated with glTF assets [21, 20] to ensure consistent interpretation. This highlights the significant effort, particularly from MPEG, to extend the core glTF format into a more comprehensive framework for complex, interactive, and immersive media applications. These MPEG-developed glTF extensions are summarized in Table 4.

Table 4: Major glTF extensions in MPEG-I.

Extension	Purpose	Key Properties
Texture-Based Lights	Dynamic lighting	Texture, light data, time params
Punctual Lights	Point light sources	Position, intensity, color
Haptic	Tactile feedback	Texture-based haptic mapping
Scene Interactivity	Interactive behaviors	Triggers, actions, collision mesh
Node Avatar	User avatars	Animation, collision linkage
Anchor	AR anchoring	Reference spaces, transforms
V-PCC Codec Ext.	Volumetric codec	Pipeline integration, signalling
MIV Codec Ext.	Mixed-media codec	Stream signalling

3.3 Comparison with Academic Approaches

Having reviewed the primary efforts in industry standardization (Sections 3.2.1 and 3.2.2) and academic research (Section 2), we now explicitly revisit the key distinctions between these two domains regarding 3D scene description. Understanding these differences is crucial for identifying challenges and opportunities for convergence. The divergence stems primarily from their distinct objectives and target applications. Table 5 provides a summary of these key distinctions.

- *Primary Goal:* Industry standards such as MPEG Scene Description and glTF prioritize interoperability, efficiency, stability, and broad adoption across platforms [19, 49]. In contrast, academic research, particularly in areas such as 3D Scene Graph Generation (3D-SGG), typically focuses on achieving deep semantic understanding, capturing complex inter-object relationships, enabling scene reasoning, and exploring novel AI-driven techniques [54, 29].
- *Core Representation:* Standardization efforts define structured formats focusing on components essential for rendering and interchange, such as node hierarchies, geometry, materials, and animations. Relationships between objects are often represented implicitly through the hierarchy

Table 5: Key differences between academic research and industry standardization.

Aspect	Academic Research	Industry Standardization
Primary Goal	Semantic depth	Interoperability & speed
Core Representation	Graph-based semantics	Component-based formats
Input Data	Point clouds & captions	Meshes, textures, streams
Evolution Pace	Fast-paced AI	Consensus-driven progress

or require dedicated extensions [19, 49]. Academic approaches such as 3D-SGs utilize explicit graph structures where nodes represent objects (often with rich attributes) and edges explicitly encode diverse semantic relationships [54, 65].

- *Data Focus and Scope:* While standards aim for broad applicability, academic scene graph generation research often concentrates on processing specific sensor data such as point clouds obtained from real-world scans [54, 18]. Furthermore, standards often need to consider integration within a larger multimedia ecosystem [19, 23], whereas academic work might investigate specific aspects in isolation.

These fundamental differences in goals and methodologies explain the current gap between cutting-edge research findings and their adoption within widely used industry standards. Bridging this gap requires mutual understanding and collaborative efforts, discussed further in Section 4.

3.4 Challenges in Standardization for Scene Description

Despite significant progress by standardization bodies such as MPEG and Khronos, developing and adopting comprehensive 3D scene description standards faces inherent challenges:

- *Pace Mismatch with Research:* Academic research, especially in AI-driven fields, evolves rapidly, often outpacing the slower, consensus-based standardization processes that prioritize stability, interoperability, and backward compatibility. This lag hinders the timely integration of state-of-the-art innovations into formal standards.
- *Semantic representation gap:* As highlighted (Section 3.3), a fundamental gap exists between semantically rich academic representations (e.g., 3D-SGs) and the rendering-focused core of standards such as glTF. Standardizing the representation and efficient transmission of complex semantic relationships and attributes remains a hurdle. Practical efforts,

such as those exploring the conversion of formats such as 3DSSG to glTF (often requiring custom extensions such as EXT_relationships), exemplify this challenge [21, 20].

- *Complexity and fragmentation:* The integration of diverse features, such as rendering, physics, haptics, compression codecs, interactivity, and semantics, into a unified, extensible standard is inherently complex. This is evident in the broad scope of MPEGs EEs [19, 23, 22]. Without careful coordination, the proliferation of specialized extensions may lead to fragmentation, impeding cross-platform interoperability.
- *Lack of standardized benchmarks:* The absence of widely accepted datasets and evaluation across academic and industrial domains hampers objective comparisons and reproducibility. This limitation obstructs the validation and potential standardization of emerging research outputs.
- *Unclear integration pathways:* Establishing agile yet reliable mechanisms for proposing, assessing, and incorporating novel research into formal standards remains a notable challenge. Effective frameworks are needed to bridge the gap between prototypes and deployable, standardized features.

Addressing these challenges requires ongoing dialogue and closer collaboration to foster a more unified ecosystem. Such collaboration can be facilitated by mechanisms including joint working groups, the development of semantic extensions, and shared benchmarks. These mechanisms are explored further in Section 4. The identified challenges, their impacts, and potential remedies are summarized in Table 6.

Table 6: Standardization challenges & remedies.

Challenge	**Impact**	**Remedy**
Pace mismatch	Slow adoption	Joint working groups
Semantic gap	Limited expressivity	glTF semantic extensions
Fragmentation	Incompatible versions	Extension consolidation
No benchmarks	Hard comparisons	Shared test suites
Slow integration	Feature delays	Clear proposal workflows

4 Discussion of Future Directions

4.1 Bridging the Academia-Industry Divide

The preceding sections have surveyed the landscape of 3D scene description technologies. This survey has revealed distinct yet complementary efforts within academic research (Section 2) and industry standardization (Section 3). Academic research has made significant strides to enhance the semantic richness of scene representations. Approaches such as 3D Scene Graphs [54, 65, 29] excel at the explicit modeling of complex object relationships and attributes, while AI-driven techniques, such as Transformer architectures [36] and knowledge integration methods [56, 36], continuously push the boundaries of scene understanding and generation capabilities. The primary focus often lies on the exploration of novel representations and algorithms for deeper semantic insight.

In contrast, industry-oriented standardization efforts, spearheaded by bodies such as MPEG and Khronos, prioritize interoperability, efficiency, and stability to enable widespread adoption and real-time application [19, 49]. Standards such as glTF provide efficient formats to transmit and render core scene components, while efforts within MPEG aim to integrate scene descriptions seamlessly within the broader multimedia ecosystem [19]. Extensibility mechanisms make it possible to add functionalities such as advanced lighting, haptics, and interactivity, but the core focus remains on the creation of robust and widely compatible building blocks [50, 23].

As highlighted in the comparison (Section 3.3), this leads to a discernible divide. Specifically, academia often prioritizes "what" can be represented semantically, while industry focuses on "how" scene data can be efficiently delivered and rendered across diverse platforms. This divergence manifests in different representational choices (explicit semantic graphs vs. rendering-focused components) and contributes to the challenges discussed previously (Section 3.4), such as the semantic gap, the pace mismatch between research and standardization, and difficulties in the establishment of unified benchmarks. While this specialization is understandable given the different mandates, the current separation limits the potential for synergy and hinders the practical realization of truly intelligent and immersive 3D experiences. To bridge this divide through collaboration is therefore essential for future progress.

4.2 The Mutual Benefits of Collaboration

Bridging the gap between academic research and industry standardization in 3D scene description, as discussed in Section 4.1, is not only desirable but also essential for accelerating the field's advancement. Closer collaboration fosters a synergistic relationship wherein the strengths of each domain effectively complement one another.

4.2.1 Benefits for Academia

- *Increased real-world relevance and impact:* Engagement with industry challenges enables academic research to address practical, real-world problems, thereby enhancing its relevance and impact. This often involves access to industry-scale datasets and application-driven use cases, enhancing the real-world applicability and impact of academic contributions beyond theoretical insights.
- *Robust validation frameworks:* Partnerships with industry offer access to standardized benchmarks and deployment environments, allowing academic innovations to be validated in a realistic setting, far surpassing the limitations of purely academic datasets.
- *Access to resources and expertise:* Academic institutions benefit from shared access to proprietary resources, development toolkits, testing platforms, and specialized domain expertise, which may otherwise be inaccessible.

4.2.2 Benefits for Industry

- *Faster innovation cycles:* Industry gains earlier and deeper access to cutting-edge research innovations in areas such as AI-driven scene understanding [36], complex relationship modeling [65], multi-granularity analysis [18], and knowledge integration [56]. This can fuel the development of next-generation products and services.
- *Improved product functionality:* Integrating richer semantic information and advanced AI capabilities derived from academic research can significantly enhance the intelligence, interactivity, and immersion of industrial applications in AR/VR, digital twins, robotics, and beyond.
- *Solutions to technical challenges:* Industry can leverage academic expertise and novel approaches to tackle complex technical hurdles encountered during the development and deployment of sophisticated 3D applications, such as handling large-scale scene understanding or complex semantic queries.
- *Future-ready standards:* Collaboration helps ensure that industry standards evolve proactively to incorporate valuable semantic representations and AI functionalities, preventing standards from becoming outdated and ensuring they meet future market needs.

In summary, academic-industry collaboration fosters a dynamic ecosystem that delivers 3D scene description solutions that are both semantically

advanced and practically deployable. This convergence is critical to unlocking the full potential of intelligent, immersive 3D technologies.

4.3 Concrete Mechanisms for Collaboration

Recognizing the mutual benefits (Section 4.2) is the first step. Translating this into tangible progress requires establishing concrete mechanisms to foster sustained interaction and integration between academic researchers and industry standardization experts. Based on the challenges identified (Section 3.4), several potential pathways can facilitate this collaboration:

- *Joint workshops and standardization forums:* Host recurring workshops co-located with major academic conferences (e.g., CVPR, ECCV, SIGGRAPH) and standardization meetings (e.g., MPEG, Khronos). These events serve as a dedicated forum where researchers can present their findings to standardization bodies, while industry representatives can share practical constraints and emerging needs with the academic community.

- *Shared benchmarks and datasets:*Collaboratively develop benchmark datasets and evaluation protocols that address both academic objectives, such as assessing semantic depth and long-tail phenomena, and industrial requirements, including scalability, computational efficiency, and applicability to real-world contexts. This alignment enables more robust and comparable evaluations, supporting the transition from academic insight to standardizable technology.

- *Standardized extension proposal pathways:* Creating a clearer and more accessible process for academic researchers to propose extensions to existing standards, such as glTF [49]. These may include comprehensive submission guidelines, mentorship schemes, or dedicated tracks within standardization working groups. Inspiration may be drawn from mechanisms like MPEGs EEs [23] and ongoing discussions around format conversions and asset definitions [21, 20].

- *Collaborative open source projects:* Promote joint open-source efforts that bridge academic and industrial ecosystems. Examples include the development of format converters between academic representation (e.g., 3DSSG) and industry standards (e.g., glTF), or the creation of reference implementations for unified scene representations and reusable software modules, as explored in MPEG initiatives [21, 20].

- *Cross-disciplinary working groups:* Establish specialized working groups that connect standardization bodies (MPEG, Khronos) with academic

communities. These groups could focus on solving targeted technical challenges, such as defining standardized representations for semantic relationships in glTF [49] or integrating AI-driven scene understanding into standardized formats.

- *Researcher exchange and residency programs:* Support short-term residencies or internships that allow academic researchers to work within industry labs or standardization organizations, and vice versa. Such programs foster mutual understanding, facilitate hands-on knowledge exchange, and align academic innovations with practical implementation pathways.

The implementation of such mechanisms can help create a continuous feedback loop. This loop enables academic innovation to inform practical standards more rapidly and ensures that standardization efforts remain relevant to the state-of-the-art in 3D scene understanding and representation.

A concrete example of collaborative efforts to bridge academic data formats and industry standards is the proposed pipeline to convert 3DSSG data to the glTF format, discussed within MPEG standardization [21, 20]. This initiative addresses the difference between glTF and 3DSSG file formats, where 3DSSG emphasizes interrelationships among objects while glTF provides a broader range of scene-related attributes and rendering technologies [21]. A standardized format that bridges these two spheres would benefit both academia and industry. This format would allow academia to develop optimized models and industry to implement them for enhanced efficiency and innovation [21]. The potential for cooperative conversions between these formats has received positive feedback regarding their applicability as test assets [21, 20].

The conversion process, illustrated in Figure 3, begins with the loading of the original 3DSSG data with custom Python scripts for data extraction. Individual OBJ files are then generated for each object within the dataset scenes. This process involves the creation of meshes and the application of texture mapping with Blender's Python API. Next, these individual OBJ files are merged into a single scene within Blender. This unified scene is then converted into a glTF file with Blender's built-in export functionality. Finally, extensions for relationship references are added to the exported glTF file. These extensions provide additional semantic information about object relationships within the 3D data representation. Figure 4 visually demonstrates key stages of this conversion process. The significance of this progression is that it illustrates a practical pathway from academic data formats to industry-ready standards. It begins with the raw, segmented point cloud (PLY) common in research, proceeds through an intermediate 3D model (OBJ), and culminates in the final, rendering-optimized glTF format. This pipeline serves as a practical demonstration that bridges academic and industry formats.

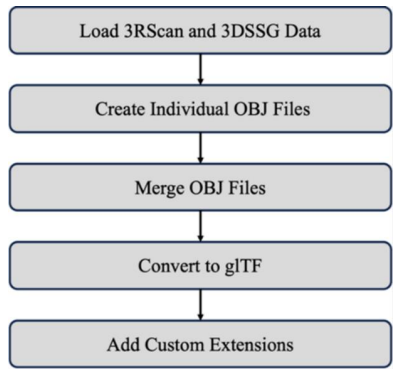

Figure 3: Process of converting 3Rscan and 3DSSG to glTF format.

4.4 Towards Unified Frameworks and Future Directions

The ultimate goal of the fostering of collaboration between academia and industry (Section 4.3) is to move towards a more unified ecosystem for 3D scene description. Such an ecosystem would ideally leverage the semantic depth and AI-driven capabilities that emerge from academic research while it also retains the interoperability, efficiency, and robustness prioritized by industry standards. While the achievement of a single, perfect "one-size-fits-all" representation might be unrealistic given the diverse application needs, the development of frameworks that allow synergistic integration of different approaches is a crucial future direction.

Future efforts, guided by collaboration, should focus on several key research and development areas to bridge the existing gaps:

- *Hybrid representations and format bridging:* Further research is essential to develop hybrid data structures or standardized formats that efficiently encode both rendering-oriented content (e.g., glTF components) and rich semantic graph information (e.g., 3D scene graphs). This may involve creating official, well-defined glTF extensions to support semantic graphs or designing novel representations that inherently unify geometric and semantic content. Initial progress in this direction includes format conversion tools, such as the 3DSSG-to-glTF pipeline explored within MPEG [21, 20]. However, broader generalization and formal standardization remain necessary.

- *Standardized semantic taxonomies:* Developing common, extensible ontologies or standardized vocabularies for object classes, attributes, and particularly relationship predicates is essential. Such taxonomies need to be rich enough for academic research yet practical for industry imple-

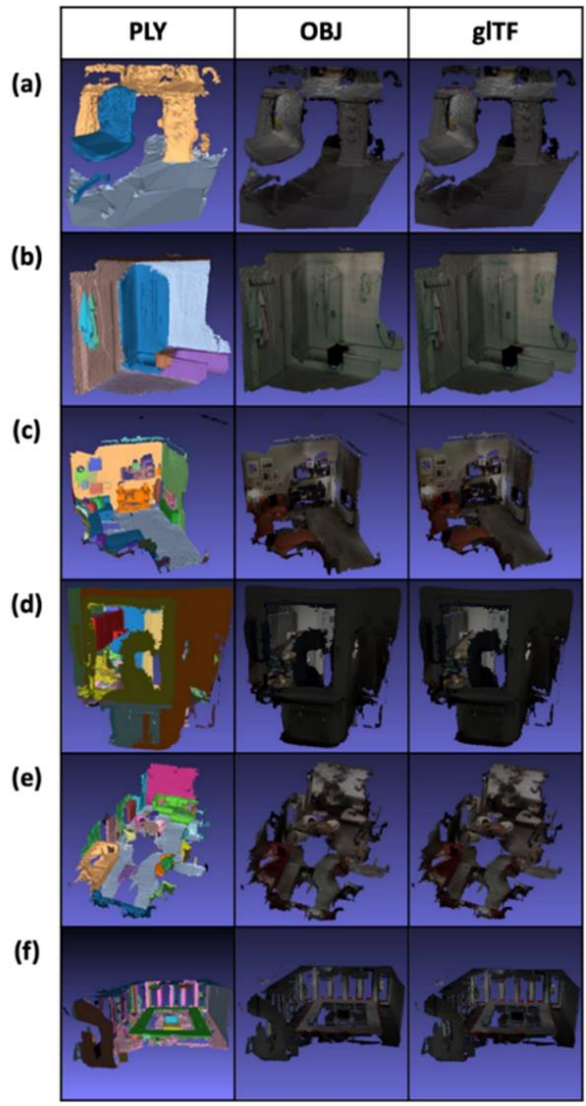

Figure 4: The 3DSSG-to-glTF conversion visualization result. We converted the obj file and segmentation ply file of 3DSSG(research data) into glTF(industry format), and confirmed that the converted glTF result looks similar to the obj file before conversion.

mentation, facilitating semantic interoperability across different datasets and applications [54, 29].

- *Cross-Domain benchmarking platforms:* The development of comprehensive benchmarking platforms is vital. These benchmarks should assess not only semantic accuracy (e.g., R@K, mR@K for scene graph generation) but also metrics critical to industrial settings, such as runtime efficiency, memory usage, and the impact on rendering quality, across diverse scene types and real-world tasks.
- *AI and standardization synergy:* There is considerable potential for AI/ML techniques to support and accelerate the standardization process itself. Possible applications include automated tools for validating format compliance, facilitating semantic data conversions, and even extracting insights to guide the creation of future standards.
- *Multi-Modal integration within standards:* Continued research is needed to enable robust integration of 3D scene descriptions with other modalities in standardized frameworks. These modalities include natural language (building on 3D Dense Captioning research [63]), haptic feedback (as explored in MPEG EEs [23]), spatial audio, and dynamic behaviors. A unified multi-modal standard would enrich the fidelity and interactivity of 3D scene representations.

Realizing a comprehensive and unified approach to 3D scene description demands ongoing collaboration and commitment to mechanisms previously outlined. By aligning efforts, academic researchers and industry professionals can ensure that innovations in 3D understanding and representation are effectively translated into the next generation of intelligent, immersive, interactive experiences.

5 Conclusion

This survey presents an in-depth review of current 3D scene description technologies and highlights the divergence between academic and industry approaches. Academic research tends to prioritize semantic and advanced AI capabilities, whereas industry-driven standardization emphasizes interoperability and real-time performance. These differing priorities contribute to critical challenges, such as inconsistent data formats and mismatched development timelines, which hinder broader technological integration. Our central conclusion is the pressing need to enhance collaboration between academia and industry. Such cooperation is essential to bridge existing gaps, and it must be grounded in joint initiatives, shared benchmarks, and the co-development of unified frameworks that combine semantic sophistication with practical usability. Strengthened partnerships between these domains will ultimately drive the development of next-generation 3D technologies and will thereby enable

more innovative, more immersive, and deeply interactive digital experiences. We hope this survey serves as a springboard for these collaborative efforts.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC(Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2025-RS-2024-00437718, 50%) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2023-00227431, 50%, Development of 3D space digital media standard technology).

References

[1] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, 422–40.

[2] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments", in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, 65–72.

[3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, 5470–9.

[4] D. Cai, L. Zhao, J. Zhang, L. Sheng, and D. Xu, "3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 16464–73.

[5] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, "Segment any 3d gaussians", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, No. 2, 2025, 1971–9.

[6] D. Z. Chen, A. X. Chang, and M. NieSSner, "Scanrefer: 3d object localization in rgb-d scans using natural language", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, 202–21.

[7] G. Chen and W. Wang, "A survey on 3d gaussian splatting", *arXiv preprint arXiv:2401.03890*, 2024, arXiv: 2401.03890.

[8] S. Chen, H. Zhu, X. Chen, Y. Lei, G. Yu, and T. Chen, "End-to-end 3d dense captioning with vote2cap-detr", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 11124–33.

[9] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-Embedded Routing Network for Scene Graph Generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Y. Chen, S. Dong, X. Wang, L. Cai, Y. Zheng, and Y. Yang, "Sg-nerf: Neural surface reconstruction with scene graph optimization", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, 188–205.

[11] Y. Chen, C. Gu, J. Jiang, X. Zhu, and L. Zhang, "Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering", *arXiv preprint arXiv:2311.18561*, 2023, arXiv: 2311.18561.

[12] Z. Chen, A. Gholami, M. NieSSner, and A. X. Chang, "Scan2cap: Context-aware dense captioning in rgb-d scans", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 3193–203.

[13] Z.-T. Chou, S.-Y. Huang, I. Liu, and Y.-C. F. Wang, "GSNeRF: Generalizable semantic neural radiance fields with enhanced 3D scene understanding", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 20806–15.

[14] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. NieSSner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 5828–39.

[15] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3d gaussian splatting as new era: A survey", *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[16] X. Fu, S. Zhang, T. Chen, Y. Lu, L. Zhu, X. Zhou, A. Geiger, and Y. Liao, "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation", in *Proceedings of the International Conference on 3D Vision (3DV)*, 2022, 1–11.

[17] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 11108–17.

[18] K. Huang, J. Yang, J. Wang, S. He, Z. Wang, H. He, Q. Zhang, and G. Lu, "Granular3D: Delving into multi-granularity 3D scene graph prediction", *Pattern Recognition*, 153, 2024, 110562.

[19] International Organization for Standardization, "Information technology – Coded representation of immersive media – Part 14: Scene Description", *Standard* No. ISO/IEC 23090-14:2022, ISO/IEC, 2022, https://www.iso.org/standard/80900.html (accessed on 04/25/2025).

[20] ISO/IEC JTC 1/SC 29/WG 3, "[SD] Guideline of format conversion from 3DSSG and 3Rscan to glTF", *Contribution/Input Document* No. m65479, ISO/IEC JTC 1/SC 29/WG 3, October 2023, https://dms.mpeg.expert/doc_end_user/documents/144_Hannover/wg11/m65479-v1-m65479-%5BSD%5DGuideline%5Callowbreak%20offormatconversiozip (accessed on 04/25/2025).

[21] ISO/IEC JTC 1/SC 29/WG 3, "[SD] Report on possible asset for scene description", *Contribution/Input Document* No. m64815, ISO/IEC JTC 1/SC 29/WG 3, October 2023, https://dms.mpeg.expert/doc_end_user/documents/144_Hannover/wg11/m64815-v3-m64815-%5Callowbreak%20%7B%5D%5Callowbreak%20%7B%5D%5Callowbreak%20%7B%5D%5Callowbreak%20%7B%5D%5Callowbreak%20report%5Callowbreak%20on%5Callowbreak%20possible%5Callowbreak%20asset%5Callowbreak%2010%5Callowbreak%20zip (accessed on 04/25/2025).

[22] ISO/IEC JTC 1/SC 29/WG 3, "Exploration Experiments for MPEG-I Scene Description", *Output Document* No. N00383, ISO/IEC JTC 1/SC 29/WG 3, October 2021, https://dms.mpeg.expert/doc_end_user/documents/136_OnLine/wg11/MDS20860_WG03_N00383.zip (accessed on 04/25/2025).

[23] ISO/IEC JTC 1/SC 29/WG 3, "Exploration Experiments for MPEG-I Scene Description", *Output Document* No. N00540, ISO/IEC JTC 1/SC 29/WG 3, April 2022, https://dms.mpeg.expert/doc_end_user/documents/138_OnLine/wg11/MDS21433_WG03_N00540.zip (accessed on 04/25/2025).

[24] J. T. Kajiya and B. P. V. Herzen, "Ray tracing volume densities", *ACM SIGGRAPH Computer Graphics*, 18(3), 1984, 165–74.

[25] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.", *ACM Trans. Graph.*, 42(4), 2023, 139:1–139:23.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks", *arXiv preprint arXiv:1609.02907*, 2016, arXiv: 1609.02907.

[27] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction", *ACM Transactions on Graphics (ToG)*, 36(4), 2017, 1–13.

[28] H. Li, D. Zhang, Y. Dai, N. Liu, L. Cheng, J. Li, J. Wang, and J. Han, "Gp-nerf: Generalized perception nerf for context-aware 3d scene under-

standing", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 21708–18.

[29] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, X. Zhao, S. A. A. Shah, and M. Bennamoun, "Scene graph generation: A comprehensive survey", *Neurocomputing*, 566, 2024, 127052.

[30] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

[31] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries", in *Text Summarization Branches Out*, 2004, 74–81.

[32] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, "Capturing, reconstructing, and simulating: the urbanscene3d dataset", in *European Conference on Computer Vision*, Springer, 2022, 93–109.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[34] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu, *et al.*, "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 22160–9.

[35] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual Relationship Detection with Language Priors", in *Proceedings of the European Conference on Computer Vision (ECCV)*, ed. B. Leibe, J. Matas, N. Sebe, and M. Welling, 2016, 852–69.

[36] C. Lv, M. Qi, X. Li, Z. Yang, and H. Ma, "Sgformer: Semantic graph transformer for point cloud-based 3d scene graph generation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 5, 2024, 4035–43.

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

[38] J. Ost, F. Mannan, N. Thuerey, J. Knodt, and F. Heide, "Neural scene graphs for dynamic scenes", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 2856–65.

[39] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, 311–8.

[40] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 9277–86.

[41] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 652–60.

[42] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space", in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, 5099–108.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision", in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, 8748–63.

[44] J.-C. Shi, M. Wang, H.-B. Duan, and S.-H. Guan, "Language embedded 3d gaussians for open-vocabulary scene understanding", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 5333–43.

[45] Y. Siddiqui, L. Porzi, S. R. Buló, N. Müller, M. NieSSner, A. Dai, and P. Kontschieder, "Panoptic lifting for 3d scene understanding with neural fields", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 9043–52.

[46] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic Scene Completion From a Single Depth Image", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, *et al.*, "The replica dataset: A digital replica of indoor spaces", *arXiv preprint arXiv:1906.05797*, 2019.

[48] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased Scene Graph Generation From Biased Training", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[49] The Khronos Group, "glTF 2.0 Specification", *Specification*, Version 2.0., The Khronos Group Inc., June 2017, https://registry.khronos.org/glTF/specs/2.0/glTF-2.0.html (accessed on 04/25/2025).

[50] The Khronos Group Inc., "glTF Extension Registry", Web Page, Official registry listing Khronos and vendor extensions for glTF., 2024, https://registry.khronos.org/glTF/extensions/ (accessed on 04/25/2025).

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,. Kaiser, and I. Polosukhin, "Attention is all you need", in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, 5998–6008.

[52] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 4566–75.

[53] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. NieSSner, "Rio: 3d object instance re-localization in changing indoor environments", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 7658–67.

[54] J. Wald, H. Dhamo, N. Navab, and F. Tombari, "Learning 3d semantic scene graphs from 3d indoor reconstructions", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 3961–70.

[55] H. Wang, C. Zhang, J. Yu, and W. Cai, "Spatiality-guided transformer for 3d dense captioning on point clouds", *arXiv preprint arXiv:2204.10688*, 2022, arXiv: 2204.10688.

[56] Z. Wang, B. Cheng, L. Zhao, D. Xu, Y. Tang, and L. Sheng, "Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 21560–9.

[57] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 7515–25.

[58] W. Xiao, R. Chierchia, R. S. Cruz, X. Li, D. Ahmedt-Aristizabal, O. Salvado, C. Fookes, and L. Lebrat, "Neural Radiance Fields for the Real World: A Survey", 2025, arXiv: 2501.13104 [cs.CV].

[59] Y. Xu, B. Chen, Z. Li, H. Zhang, L. Wang, Z. Zheng, and Y. Liu, "Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, 1931–41.

[60] Z. Yang, H. Yang, Z. Pan, and L. Zhang, "Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting", *arXiv preprint arXiv:2310.10642*, 2023, arXiv: 2310.10642.

[61] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, 20331–41.

[62] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes", in *European Conference on Computer Vision*, Springer, 2024, 162–79.

[63] T. Yu, X. Lin, S. Wang, W. Sheng, Q. Huang, and J. Yu, "A comprehensive survey of 3d dense captioning: Localizing and describing objects in 3d scenes", *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3), 2023, 1322–38.

[64] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, S. Cui, and Z. Li, "X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 8563–73.

[65] C. Zhang, J. Yu, Y. Song, and W. Cai, "Exploiting edge-oriented reasoning for 3d point-based scene graph analysis", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 9705–15.

[66] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 21769–80.

[67] S. Zhang, S. Li, A. Hao, and H. Qin, "Knowledge-inspired 3D Scene Graph Prediction in Point Cloud", in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, 2021, 18620–32.

[68] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 15838–47.

[69] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 21676–85.