

LESV: Language Embedded Sparse Voxel Fusion for Open-Vocabulary 3D Scene Understanding

Fusang Wang^{1,2}, Nathan Piasco¹, Moussab Bennehar¹,
Luis Roldão¹, Dzmitry Tsishkou¹, and Fabien Moutarde²

¹ Noah's Ark, Huawei Paris Research Center, France

² CAOR, Mines-Paris PSL, France

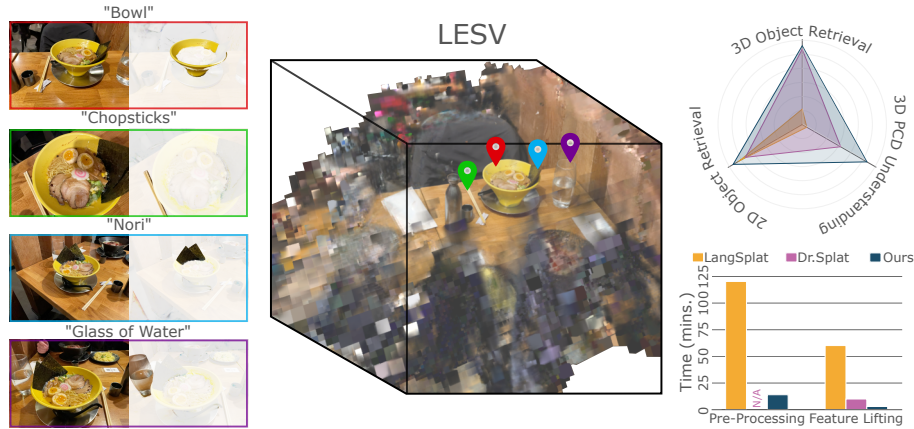


Fig. 1: By registering dense, language-aligned features from the AM-RADIO foundation model onto an explicit Sparse Voxel Representation, LESV enables precise, deterministic localization of complex, fine-grained queries directly in 3D. This structured volume fusion facilitates general-purpose scene understanding, establishing a new state-of-the-art in open-vocabulary point cloud understanding and achieving on par performance on 3D object retrieval and 2D object localization, while drastically reducing feature lifting and data preprocessing time.

Abstract. Recent advancements in open-vocabulary 3D scene understanding heavily rely on 3D Gaussian Splatting (3DGS) to register vision-language features into 3D space. However, we identify two critical limitations in these approaches: the *spatial ambiguity* arising from unstructured, overlapping Gaussians which necessitates probabilistic feature registration, and the *multi-level semantic ambiguity* caused by pooling features over object-level masks, which dilutes fine-grained details. To address these challenges, we present a novel framework that leverages Sparse Voxel Rasterization (SVRaster) as a structured, disjoint geometry representation. By regularizing SVRaster with monocular depth and normal priors, we establish a stable geometric foundation. This enables a deterministic, confidence-aware feature registration process and suppresses the semantic bleeding artifact common in 3DGS. Furthermore, we resolve

multi-level ambiguity by exploiting the emerging dense alignment properties of the **AM-RADIO** foundation model, avoiding the computational overhead of hierarchical training methods. Our approach achieves state-of-the-art performance on Open Vocabulary Point Cloud Understanding, and highly competitive results on 3D Object Retrieval benchmarks.

1 Introduction

The ability to query and interact with 3D environments using open-vocabulary natural language is a cornerstone of next-generation robotics and augmented reality [1, 8, 9, 17, 18]. While early approaches leveraged implicit Neural Radiance Fields (NeRFs) [19], the community has recently shifted toward explicit 3D Gaussian Splatting (3DGS) [27] to enable real-time rendering of feature fields. Pioneering methods such as LangSplat [16, 22] and LeGaussians [27] view 3D scene understanding as a view-synthesis problem: they lift language embeddings like CLIP [24] into 3D Gaussians by supervising rendered 2D feature maps against pre-extracted priors. However, this “render-and-compare” paradigm is computationally expensive and decoupled from direct spatial reasoning in 3D [30]. To address this, recent works [6, 11] propose *direct feature registration*. Instead of optimizing features via 2D supervision, these methods directly lift 2D vision-language features onto 3D primitives using inverse volume rendering. While this shift reduces the feature-lifting time from hours to minutes, we identify two fundamental limitations that prevent current registration pipelines from achieving high-fidelity retrieval.

Spatial Ambiguity in Unstructured Geometry. 3DGS represents scenes using a collection of unstructured, anisotropic Gaussians that significantly overlap. Consequently, determining which Gaussian should be assigned a specific high-dimensional feature becomes a complex, heuristic-driven task [11]. This inherent ambiguity results in semantic bleeding and ray-like spillovers at the object boundaries [6], preventing the precise delineation required for point-level 3D interaction.

Multi-Level Semantic Ambiguity. A single point in 3D space often resides within multiple semantic hierarchies simultaneously, a challenge known as *point ambiguity* [22]. For instance, a spatial coordinate on a bear’s nose should ideally respond to queries for its sub-part (‘nose’), parent component (‘head’), and global category (‘bear’). A widely adopted solution is to resolve this by leveraging Segment Anything Model (SAM) [14] hierarchies to train multiple independent feature fields [22]. While effective, such an approach imposes massive pre-processing overhead, given redundant optimization and registration passes for every semantic hierarchy.

To resolve the aforementioned challenges, we propose a paradigm shift in both scene representation and feature encoding for 3D language fields. For spatial ambiguity, we move beyond the probabilistic nature of Gaussians by adopting Sparse Voxel Rasterization (SVRaster) [28]. This leverages both an explicit voxel geometry and a well-defined surface, enabling deterministic feature registration

and avoiding severe memory bottlenecks. For multi-level semantic ambiguity, we build upon the emergent properties of Agglomerative Vision Foundation Model, specifically AM-RADIO [25]. Unlike CLIP’s global alignment, AM-RADIO is trained to distill both fine-grained spatial signals from DINOv2 [20], SAM [14] and high-level language semantics from SigLIP [32], effectively allowing spatial tokens to inherit global content while retaining local semantics [1]. By aligning the spatial token with the language feature space, we generate a semantic field that naturally supports multi-scale queries in a single pass. Our contributions can be summarized as follows:

- We exploit Sparse Voxel Rasterization (SVRaster) [28] for 3D feature fields. Its explicit, disjoint voxel geometry, facilitates deterministic feature registration, circumventing the memory bottlenecks from inverse volume rendering.
- By relying on agglomerative vision foundation models [25], we resolve the multi-level ambiguity inherent in hierarchical methods. This enables precise and fine-grained querying while drastically reducing feature preprocessing time.
- Through extensive experiments on 2D and 3D benchmarks, we demonstrate that LESV establishes a new state-of-the-art in open-vocabulary point cloud understanding [7] and achieves competitive performance on 3D object retrieval and 2D localization [13].

2 Related Work

2.1 Language-Embedded 3D Scene understanding

The integration of vision-language priors [23, 32] into 3D reconstruction has evolved from implicit neural representations [19] to explicit, real-time primitives like 3D Gaussian Splatting (3DGS) [12, 27]. Early approaches such as LERF [13] grounded CLIP features within Neural Radiance Fields (NeRFs). However, relying on volumetric ray-marching to query language embeddings is computationally expensive and requires multi-scale rendering during inference to handle different semantic levels [22]. To address the computational bottlenecks of NeRF-based language fields, recent research [22, 26, 27, 33] has pivoted toward 3DGS. These frameworks optimize 3D feature fields by explicitly supervising rendered 2D features against pre-computed feature maps. Because directly splatting these high-dimensional feature vectors (e.g., 512-D for CLIP) incurs prohibitive memory costs and precludes real-time performance, existing frameworks employ dimensionality reduction strategies. For instance, LangSplat [22] and Feature3DGS [33] compress features into a low-dimensional latent space via scene-specific autoencoders. More recently, voxel-based representations have begun to emerge as a promising alternative backbone representation for language field construction [10], motivating our adoption of Sparse Voxel Rasterization as a structured foundation for feature lifting.

2.2 Direct Feature Registration with 3DGS

While feature distillation has proven effective for novel view synthesis, it is poorly suited to direct 3D queries: features stored in individual Gaussians are typically only meaningful after alpha-blending in 2D, biasing the representation toward rendering rather than 3D understanding [3, 11, 30]. To address this, recent methods pursue different strategies. Some approaches leverage multi-view consensus during feature distillation [3], others employ contrastive learning with SAM masks to group multi-view 2D semantics into coherent 3D objects [4, 30]. However, these approaches still require an additional optimization stage after the 3D representation is trained. An alternative is direct feature registration, which lifts 2D features onto 3D Gaussians via inverse volume rendering [5, 6, 11]. For example, Dr. Splat [11] registers embeddings by mapping per-pixel CLIP features onto the top- K Gaussians intersected by each viewing ray, enabling more 3D-native retrieval. Despite the fast lifting speed and improved 3D-native precision, the registration remains probabilistic and ray-dependent, which can dilute semantics and produce spillover false positives during 3D queries [6]. ProFuse [6] mitigates these artifacts by aggregating 2D masks into 3D context proposals prior to registration and concentrating feature assignments onto a small set of Gaussians. However, the fundamental ambiguity induced by alpha-blending is not fully resolved. This motivates the use of structured geometric backbones, which support more deterministic and decoupled feature mapping.

3 Method

3.1 Preliminary: Sparse Voxel Rasterization (SVRaster)

Unlike the unstructured ellipsoids used in 3D Gaussian Splatting (3DGS), SVRaster represents the scene as a set of sparse voxels $\mathcal{V} = \{v_1, \dots, v_N\}$ organized within a high-resolution hierarchical grid [28]. Each voxel v_i is parameterized by its spatial center $\mathbf{x}_i \in \mathbb{R}^3$, a set of density values defined at its eight corners $\boldsymbol{\sigma}_i = \{\sigma_0, \dots, \sigma_7\}$, and a view-dependent color represented by Spherical Harmonic (SH) coefficients.

To render the scene, SVRaster projects voxels onto the image plane and sorts them using View-Dependent Morton Codes. This strict depth ordering explicitly prevents the popping artifacts prevalent in unstructured 3DGS architectures. The final pixel color C and depth D are computed by accumulating the contributions of sorted voxels $\mathcal{N}$ intersecting the pixel ray:

$$C = \sum_{j \in \mathcal{N}} T_j \alpha_j \mathbf{c}_j, \quad D = \sum_{j \in \mathcal{N}} T_j \alpha_j z_j, \quad \text{with} \quad T_j = \prod_{m=1}^{j-1} (1 - \alpha_m), \quad (1)$$

where z_j is the distance from the camera center to the j -th voxel center. The opacity α_j is derived from the volume density $\sigma(\mathbf{x}_j)$ and the ray-voxel intersection interval δ_j as $\alpha_j = 1 - \exp(-\sigma(\mathbf{x}_j)\delta_j)$. The density $\sigma(\mathbf{x}_j)$ is computed via trilinear interpolation of the corner densities $\boldsymbol{\sigma}_j$, ensuring a continuous density field within the explicit grid.

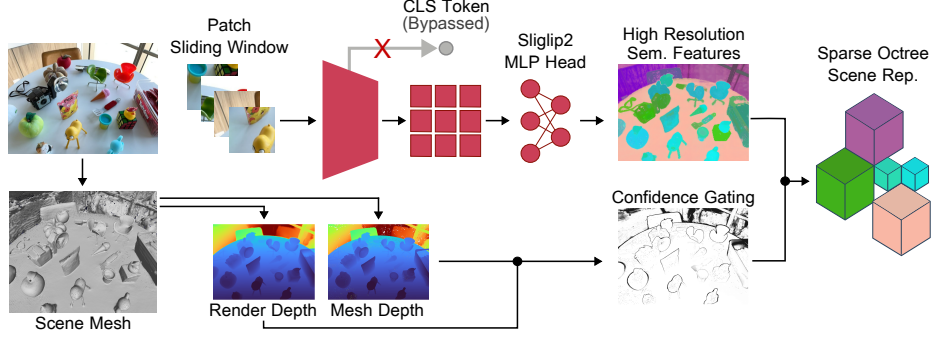


Fig. 2: Overview of the LESV Architecture. **Top:** Input images are processed via a sliding window through AM-RADIO and a SigLIP-2 MLP to extract high-resolution language aligned features. **Bottom:** Comparing rendered depth against explicit *SVRaster* mesh depth generates a geometric confidence map, which selectively gates the feature projection, ensuring high-fidelity volume fusion (**Right**).

Base Volume Fusion. The structured voxel representation of *SVRaster* facilitates a deterministic mapping between 2D pixels and 3D primitives. For a given voxel v_i , a robust global feature $\mathbf{F}_i$ can be aggregated from multi-view 2D feature maps $\mathcal{F}_k$ via weighted volume fusion:

$$\mathbf{F}_i = \frac{\sum_{k \in \Omega_i} w_{i,k} \mathbf{f}_{i,k}}{\sum_{k \in \Omega_i} w_{i,k} + \epsilon}, \quad \text{where} \quad \mathbf{f}_{i,k} = \text{Bilinear}(\mathcal{F}_k, \pi_k(\mathbf{x}_i)). \quad (2)$$

Here, Ω_i denotes the set of visible views and $\pi_k(\cdot)$ is the camera projection. To ensure geometric adherence, the weight $w_{i,k}$ is defined by a Gaussian kernel based on the depth discrepancy: $w_{i,k} = \exp(-|z_{i,k} - \mathbf{D}_k|^2 / 2\beta^2)$, which prioritizes features sampled near the rendered surface $\mathbf{D}_k$. The term β represents the bandwidth of the Gaussian kernel, which explicitly controls the sensitivity of the fusion weights to spatial depth discrepancies.

3.2 Geometric-Aware Robust Volume Fusion

Geometric Refinement via Scene Constraints. While the explicit grid structure of *SVRaster* conceptually simplifies the lifting of high-dimensional features, achieving high-fidelity 3D feature fields remains non-trivial. The vanilla *SVRaster* optimization prioritizes photometric fidelity over structural integrity, frequently resulting in fragmented or "hollow" underlying geometry. These structural impurities produce noisy surface renderings that severely disrupt the volumetric fusion. To optimize the underlying explicit geometry of *SVRaster*, we incorporate 2D monocular priors into the training objective.

Patch-wise Depth Consistency. Directly penalizing absolute depth differences between multi-view renderings and monocular priors often destabilizes optimization due to inherent global scale and shift ambiguities. To mitigate this without compromising local topology, we strategically adopt a patch-level structural

constraint, $\mathcal{L}_{\text{patch}}$, following [15]. For each localized patch $p \in \mathcal{P}$, the loss is formulated as:

$$\mathcal{L}_{\text{patch}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left\| \left(\frac{\mathbf{D}_{\text{ren}}(p) - \mu_{\text{ren}}}{\sigma_{\text{ren}}} \right) - \left(\frac{\mathbf{D}_{\text{prior}}(p) - \mu_{\text{prior}}}{\sigma_{\text{prior}}} \right) \right\|_2^2, \quad (3)$$

where μ and σ denote the mean and standard deviation of depth values within patch p . This normalized formulation is advantageous to inherit high-fidelity relative structures like edges and local curvature from the monocular prior.

Normal Consistency Supervision. While patch-wise depth regularizes zero-order spatial positioning, voxel-based feature aggregation remains highly sensitive to first-order surface variations, particularly at grazing angles and object boundaries where semantic leakage is most severe. To suppress these artifacts, we incorporate an analytic normal constraint, $\mathcal{L}_{\text{norm}}$, following [31]:

$$\mathcal{L}_{\text{norm}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} (1 - \mathbf{N}_{\text{ren}}(p) \cdot \mathbf{N}_{\text{prior}}(p)), \quad (4)$$

where $(\cdot)$ denotes the vector dot product.

Robust Volume Fusion via Multi-Level TSDF Confidence. Despite monocular geometric regularization, the rendered depths in SVRaster can still exhibit view-dependent inconsistencies. Since the fusion weight uses an exponential kernel, errors in the rendered depth $\mathbf{D}_{\text{ren},k}$ directly distort the aggregation weights. To resolve this, we leverage an explicit voxel grid advantage: extracting a 3D-consistent depth prior from the underlying Truncated Signed Distance Function (TSDF) field within the grid. This can be exploited to compute an inherently consistent mesh, naturally aligned with the global 3D scene representation and free from single-view rendering noise.

Standard mesh extraction operating on a single resolution tier typically forces a sub-optimal geometric trade-off: extraction from coarse voxels removes high-frequency structural details, whereas fine voxels yield incomplete geometries riddled with holes due to sparsely populated or unobserved TSDF values. We mitigate this by fusing TSDF information across multiple resolution levels, producing a mesh that preserves fine topological details without compromising global surface continuity (cf. supplementary). We then project this multi-level mesh into each view k to obtain a reference depth $\mathbf{D}_{\text{mesh},k}$ and formulate a continuous geometric confidence map, $w_{\text{conf},k}$, defined by the absolute depth discrepancy:

$$w_{\text{conf},k} = \exp \left(- \frac{|\mathbf{D}_{\text{mesh},k} - \mathbf{D}_{\text{ren},k}|}{2\sigma} \right), \quad (5)$$

where σ controls the decay. This confidence establishes a principled trade-off between spatial proximity and geometric accuracy. Finally, we inject this confidence into the volumetric fusion process by dynamically modulating the original

spatial weight $w_{s,i,k}$ from Eq. 2 and replacing it by our robust aggregation weight used to aggregate the 3D feature $\mathbf{F}_i$ across all visible views:

$$w_{i,k} = w_{s,i,k} \cdot w_{\text{conf},i,k}, \quad (6)$$

Scalable Batch Fusion. Direct feature registration in 3D Gaussian Splatting (3DGS) architectures faces severe memory bottlenecks (e.g., Dr. Splat [11] requires over 256GB of RAM for dense scenes). Because 3DGS defines surfaces implicitly via volumetric accumulation, aggregating 2D features onto 3D primitives necessitates global depth-sorting of all Gaussians to compute view-dependent alpha-blending weights. In contrast, the explicit grid structure of SVRaster inherently resolves this bottleneck. As demonstrated in our fusion formulation, the spatial distance from any voxel to the physical surface is deterministic. Therefore, our volume fusion process computes feature weights based solely on local geometric proximity and confidence, entirely bypassing the need for view-dependent global sorting. This spatial independence ensures voxels remain strictly disjoint during fusion. Consequently, we can partition the grid and execute multi-view feature projection in isolated spatial batches, enabling highly scalable, $\mathcal{O}(1)$ memory processing regardless of scene scale.

3.3 Resolving Multi-Level Ambiguity via Agglomerative Features

Emergent Dense Alignment in AM-RADIO. Rather than relying on the standard SAM+CLIP pipeline, we adopt AM-RADIO [25] as our feature backbone, which distills multiple teacher signals — CLIP [23], SAM [14], and DI-NOv2 [20] — into a single unified backbone. Given an input image I , the RADIO backbone Φ_{RADIO} produces a sequence of tokens consisting of a global class token $\mathbf{t}_{cls}$ and a set of spatial patch tokens $\mathbf{T}_{patch} \in \mathbb{R}^{N \times D}$:

$$[\mathbf{t}_{cls}, \mathbf{T}_{patch}] = \Phi_{\text{RADIO}}(I). \quad (7)$$

In standard vision-language alignment, a global semantic vector is obtained by projecting the CLS token through a specific language-aligned MLP head, denoted as $h_{\text{lang}}(\cdot)$. However, recent works have observed an emergent property in agglomerative models: the dense patch tokens preserve strong alignment with the language space even without explicit dense supervision [1]. Instead of relying solely on the global $\mathbf{t}_{cls}$, we directly project the spatial patch tokens through the language head MLP to generate a dense, language-aligned feature map $\mathbf{F}_{\text{lang}} \in \mathbb{R}^{H \times W \times C}$:

$$\mathbf{F}_{\text{lang}} = h_{\text{lang}}(\mathbf{T}_{patch}). \quad (8)$$

By utilizing these patch tokens, which inherently encode both local texture and global semantics, $\mathbf{F}_{\text{lang}}$ allows the generated feature field to simultaneously capture multi-scale concepts without the computational overhead of explicit hierarchical SAM-based mask pooling.

Despite these advantages, the direct application of patch-level language features introduces spatial resolution bottlenecks, particularly when segmenting small objects. Because the AM-RADIO architecture is optimized for a native input resolution of 512×512 (with a strict upper bound of 4096), naively upscaling the input image to recover spatial granularity is architecturally sub-optimal and computationally prohibitive.

Self-Correcting Sliding Window Upsampling. To circumvent these architectural constraints and extract dense semantic features at the native image resolution, we introduce a self-correcting sliding window upsampling strategy. While learning-based deep upsampling architectures (e.g., AnyUp [29]) can arbitrarily increase spatial resolution, empirical evaluations indicate that their attention-based spatial mixing heavily distorts the strict high-dimensional semantic space required for zero-shot 3D retrieval (cf. supplementary).

To bypass this semantic degradation and maintain strict feature-space fidelity, we adopt a sliding window inference approach combined with spatial Gaussian attenuation. We partition the high-resolution input into overlapping crops $\{C_1, \dots, C_M\}$. To reconstruct a continuous, high-resolution feature map $\mathbf{F}_{\text{high}}$ while explicitly mitigating hard boundary discontinuities between adjacent crops, we aggregate the localized predictions using a Gaussian kernel. For a given pixel coordinate $\mathbf{u}$ in the full-resolution image coordinate frame, the aggregated feature is defined as:

$$\mathbf{F}_{\text{high}}(\mathbf{u}) = \frac{\sum_{k=1}^M G(\mathbf{u} - \mathbf{c}_k) \cdot \mathbf{F}_k(\mathbf{u} - \mathbf{c}_k)}{\sum_{k=1}^M G(\mathbf{u} - \mathbf{c}_k)} \quad (9)$$

where $\mathbf{c}_k$ denotes the top-left spatial anchor of crop k , $\mathbf{F}_k$ represents the bilinearly upsampled AM-RADIO feature map for the crop, and $G(\cdot)$ is a 2D Gaussian kernel that assigns smoothly decaying weights near the crop boundaries.

Self-Correlating Aggregation. Although Gaussian weighting smoothly blends overlapping regions, the raw dense features extracted from AM-RADIO often retain high-frequency noise. To resolve this, we integrate the Self-Correction Recursive Attention (SCRA) and Self-Correlating Global Aggregation (SCGA) mechanisms from RADSeg [2]. These modules replace standard dot-product attention with a thresholded cosine-similarity formulation that explicitly masks out negatively correlated token pairs. By acting as a strict semantic filter, this hard-thresholding approach effectively eliminates the discordant noise injected during the window fusion process. Consequently, the final dense feature map, $\mathbf{F}_{\text{high}}$, retains semantic fidelity and sharp object boundaries prior to integration into our 3D volumetric fusion pipeline.

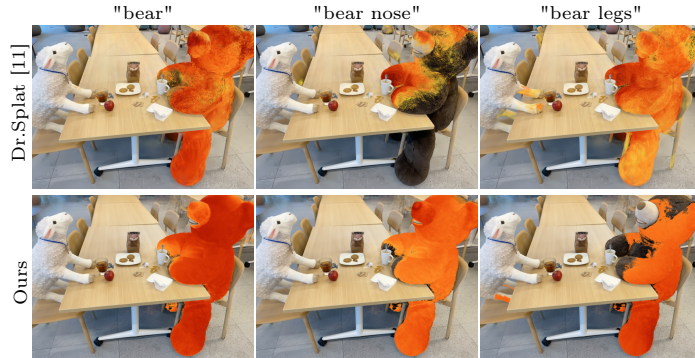


Fig. 3: Multi-Level Semantic Ambiguity. Comparison of querying on different semantic level, from a global ("bear") to sub-parts ("bear nose", "bear legs"). Our method (bottom) dynamically concentrates semantic activation onto the targeted regions.

4 Experiments

4.1 Implementation Details.

All experiments and performance analyses were conducted on a single workstation equipped with a high-end GPU (featuring compute and VRAM capabilities equivalent to an NVIDIA RTX 5090). Notably, because the original open-source implementation of Dr. Splat [11] requires prohibitive system memory allocations (exceeding 256GB of RAM on dense scenes), we integrated optimized memory management routines to evaluate it locally. This allowed us to successfully replicate their officially reported baseline metrics within standard hardware limits, ensuring a rigorously fair comparison.

4.2 Open-Vocabulary Object Retrieval

Evaluation Protocol and Metrics. To assess our model’s spatial querying capabilities, we evaluate open-vocabulary object retrieval on the LERF [13] dataset across both 3D and 2D settings. The fundamental distinction between these two tasks lies in the domain where semantic thresholding is applied. For **3D object retrieval**, we follow the evaluation protocol of Dr. Splat [11], where all methods are evaluated under a single object-level SAM mask setting. Objects are retrieved by thresholding cosine similarities directly in 3D space, without any mean filtering applied to the rendered masks during metric calculation. Since AM-RADIO features exhibit a different value distribution compared to standard SAM+CLIP paradigms, we adopt normalized cosine similarity for evaluated methods to align feature spaces prior to 3D thresholding, ensuring a fair and consistent comparison across methods. For **2D object retrieval**, we follow the evaluation strategy of LangSplat [22], which first renders the relevancy map onto the 2D image plane

from a specified camera viewpoint, and subsequently applies thresholding and mean filtering in 2D space to yield the final results.

3D Object Retrieval. Quantitative results are presented in Table 1. LESV achieves competitive state-of-the-art results on this benchmark, attaining the second-best average mIoU (56.11%) and the best Acc@25 (85.21%), outperforming the registration-based baseline Dr. Splat. Notably, dense language features from AM-RADIO effectively resolve multi-level semantic ambiguities, yielding substantial improvements on scenes dominated by complex sub-object queries (e.g., the *ramen* scene). Figure 3 further qualitatively illustrates our method’s superiority in handling multi-level linguistic ambiguities. We additionally report LaGa [4] under a different evaluation setting (marked [†]) using multi-level SAM masks, as hierarchical contrastive learning is a core contribution of the work. Although LaGa [4] achieves the best mIoU under this setting, this comes at a significantly higher preprocessing and training cost (~ 190 mins vs. ~ 20 mins for LESV).

Figure 4 demonstrates that our method effectively eliminates the spillover artifacts prevalent in 3DGS, as evidenced by the “plate” and “sink” queries. The integration of dense AM-RADIO features robustly resolves multi-level ambiguities, successfully isolating fine-grained sub-parts, such as the “corn” within the ramen. Furthermore, we achieve precise segmentation of extremely small objects, such as the “pirate hat”, which the baseline struggles to capture. Interestingly, we observe a highly nuanced semantic behavior enabled by dense feature formulation: when querying the “refrigerator”, our methods segments the appliance’s surface while excluding the unrelated photos attached to its door. Although this strict semantic decoupling is penalized by the coarse ground-truth annotations of the LERF benchmark, it highlights a fine-grained understanding of scene enabled by the method.

Generalization to 2D Localization. While distillation-based methods like LangSplat fail dramatically on the aforementioned pure 3D object retrieval task (averaging only 10.97% mIoU), deterministic 3D registration provides a much more robust mechanism for scene understanding that naturally generalizes back to 2D pro-

Method	ramen		figurines		teatime		waldo_kitchen		Average	
	mIoU	Acc@25	mIoU	Acc@25	mIoU	Acc@25	mIoU	Acc@25	mIoU	Acc@25
LangSplat [22]	5.96	9.86	6.05	8.93	18.89	23.73	12.96	18.18	10.97	15.18
OpenGaussian [30]	31.01	42.25	39.29	55.36	<u>60.44</u>	<u>76.27</u>	22.70	31.82	38.36	51.43
Dr. Splat [11]	<u>37.49</u>	<u>69.01</u>	61.73	<u>83.93</u>	59.45	72.88	52.00	81.81	<u>52.67</u>	<u>76.91</u>
LaGa [4]	31.60	46.48	49.01	83.93	58.73	89.83	38.80	63.64	44.54	70.97
LaGa [†] (multi-level) [4]	47.95	70.42	55.87	80.35	69.53	89.83	61.92	90.09	58.82	82.88
LESV (ours)	53.34	81.69	<u>55.87</u>	87.50	71.40	89.83	<u>43.84</u>	81.81	56.11	85.21

Table 1: Open-vocabulary 3D object retrieval on the LERF dataset, reporting mIoU (%) and Acc@25 (%). **Bold** and underlined denote the best and second-best results respectively. Results for LangSplat [22] and LaGa [4] are re-implemented under the same evaluation protocol as Dr. Splat [11]. LaGa[†] (multi-level) is evaluated under a different setting using multi-level SAM masks and is included for reference.

Method	ramen		figurines		teatime		waldo_kitchen		Average	
	mIoU	Loc	mIoU	Loc	mIoU	Loc	mIoU	Loc	mIoU	Loc
GAGS [21]	46.80	69.00	53.60	78.60	60.30	<u>88.10</u>	55.80	<u>90.90</u>	54.13	81.65
LangSplat [22]	51.20	73.20	44.70	80.40	65.10	<u>88.10</u>	44.50	95.50	51.38	84.30
LangSplatV2 [16]	<u>51.80</u>	<u>74.70</u>	<u>56.40</u>	<u>82.10</u>	<u>72.20</u>	93.20	<u>59.10</u>	86.40	<u>59.88</u>	84.10
Occam's LGS [5]	51.00	74.70	58.60	80.40	70.20	93.20	65.30	81.80	61.28	82.53
LESV (ours)	63.40	88.73	48.41	85.71	69.30	84.75	46.24	77.27	56.84	<u>84.12</u>

Table 2: Open-vocabulary 2D object retrieval and localization results on the LERF dataset. We report mIoU (%) for retrieval and Localization Accuracy (Loc %). **Bold** indicates the best result, underline indicates the second best. Note that the result is achieved without multi-level SAM mask generation but simple pass to Radio-v3



Fig. 4: Qualitative Comparison of 3D Object Retrieval. Visual results on the LERF dataset. Compared to the baseline Dr. Splat, LESV effectively eliminates spillover artifacts ("plate", "sink"), precisely segments fine-grained sub-parts ("corn"), and captures extremely small objects ("pirate hat"). Notably, our method exhibits fine-grained semantic decoupling by accurately isolating the "refrigerator" surface while excluding attached photos.

jection tasks. As shown in Table 2, despite not being optimized or distilled for this setting, LESV achieves localization and retrieval performance on par with dedicated 2D methods such as LangSplat-v2 [16] and Occam's LGS [5], strictly matching the average localization accuracy of LangSplat-v2 (84.12% mAcc vs. 84.10% mAcc) and outperforming it on complex hierarchical scenes like *ramen* (63.84% mIoU vs. 51.87% mIoU). Although Occam's LGS and LangSplat-v2 attain higher overall retrieval mIoU (61.28% and 59.88%, respectively), we emphasize that these methods are explicitly trained for the 2D setting, whereas LESV achieves comparable performance through generalization from its 3D-native feature lifting.

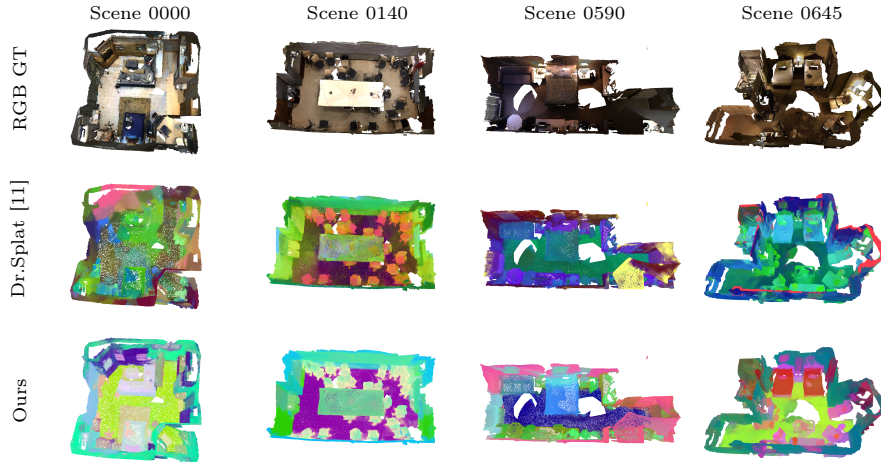


Fig. 5: Qualitative Comparison on ScanNet. Visual results of feature-transferred point clouds across diverse scenes. Compared to the baseline Dr. Splat (middle row), our method (bottom row) significantly reduces semantic bleeding at object intersections with more accurate spatial boundaries.

4.3 Open-Vocabulary Point Cloud Understanding

Evaluation Protocol. We evaluate open-vocabulary point cloud understanding on the ScanNet [7] dataset across 19, 15, and 10-class configurations. To fairly evaluate distinct representations, we adapt the feature-transfer protocol from ProFuse [6]. For 3DGS baselines, we candidate nearest $K = 64$ Gaussians short-listed by Euclidean distance and weighted by distance weighting, formulated as $\exp(-\frac{1}{2}d^2)$, and Gaussian opacity. We omit Mahalanobis gating, as we empirically observed that it degrades performance in scenes dominated by small Gaussians. For *SVRaster*, we simplify to a direct aggregation of the $K = 8$ nearest voxel features using the identical distance weighting.

Quantitative Results. Quantitative evaluations are summarized in Table 3. LESV establishes a new SoTA, significantly outperforming the leading 3DGS-based base-

Method	19 Classes		15 Classes		10 Classes	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
LangSplat [22]	3.78	9.11	5.35	13.20	8.40	22.06
OpenGaussian [30]	24.73	41.54	30.13	48.25	38.29	55.19
LaGa [4]	<u>32.50</u>	<u>49.10</u>	35.50	53.50	42.60	63.20
Dr.Splat [11]	31.66	48.64	<u>37.59</u>	<u>56.54</u>	<u>44.87</u>	<u>64.27</u>
LESV (ours)	53.22	70.41	54.78	73.62	65.25	82.95

Table 3: Open-vocabulary 3D semantic segmentation results on the ScanNet dataset across different class granularities (19, 15, and 10 classes)

line, Dr. Splat. Specifically, on the highly challenging 19-class split, our method achieves a +21.56% absolute improvement in mIoU (53.22% vs. 31.66%) and a +21.77% leap in mAcc (70.41% vs. 48.64%). Furthermore, this substantial performance gap remains consistent across both the 15-class (+17.19% mIoU) and 10-class (+20.38% mIoU) settings.

Qualitative Analysis. To contextualize these quantitative gains, we provide visual comparisons of the feature-transferred ScanNet point clouds in Figure 5. Qualitatively, LESV exhibits significantly reduced "semantic bleeding" at object intersections compared to baseline methods and preserves more accurate spatial boundaries across varying levels of semantic granularity.

4.4 Ablation Study

To validate our architectural design, we conduct an ablation study across two complementary axes. In Table 4, we isolate the geometric contribution by fixing both LESV and Dr. Splat to identical SAM+CLIP features. The baseline *SVRaster* yields sub-optimal geometry. Introducing patch-wise depth and surface normal supervision substantially refines structural fidelity. Integrating multi-level TSDF depth confidence further stabilizes results by down-weighting view-inconsistent projections, pushing LESV to outperform Dr. Splat [11] on ScanNet (+0.40 mIoU, +2.03 mAcc). This validates that both the monocular cue integration and our proposed depth confidence gating are crucial components in making Sparse Voxels outperform the 3DGS counterpart.

In Table 5, we disentangle the effect of the feature backbone by evaluating all combinations of representation and encoder. Upgrading Dr. Splat from SAM+CLIP to SAM+AM-RADIO yields only marginal gains, as its 3DGS registration cannot scale to dense per-pixel features, forcing spatially rich AM-RADIO tokens into mask-level pooled embeddings and destroying fine-grained local semantics. In contrast, LESV’s deterministic fusion directly registers dense patch-level features per voxel without mask pooling, fully preserving AM-RADIO’s spatial richness. This confirms that the performance gains arise from our dense feature registration paradigm rather than the encoder choice alone.

Configuration	LERF (3D Retr.)		ScanNet (PCD)	
	mIoU	Acc@25	mIoU	mAcc
Dr. Splat [11]	52.76	76.91	31.66	48.64
SVRaster (vanilla)	50.80	70.63	30.62	48.90
+ Mono. Sup.	53.32	71.56	31.37	49.78
+ Mono. Sup. + Depth. Conf.	53.37	72.91	32.06	50.67

Table 4: Architecture ablation with SAM+CLIP features.

Method	Feature	LERF (3D Retr.)		ScanNet (PCD)	
		mIoU	Acc@25	mIoU	mAcc
Dr. Splat [11]	CLIP	52.76	76.91	31.66	48.64
LESV (Ours)	CLIP	53.37	72.91	32.06	50.67
Dr. Splat [11]	AM-RADIO	53.42	70.60	37.57	47.80
LESV (Ours)	AM-RADIO	56.11	85.21	53.22	70.41

Table 5: Disentangling representation vs. backbone features.

4.5 Computational Efficiency

To demonstrate practical scalability, we compare peak memory usage and computational overhead in Tables 6 and 7, respectively. Dr. Splat’s official implementation exceeds 256 GB RAM on dense scenes, revealing a structural limitation: mask-level proposal caching causes memory to scale non-linearly with image count. To obtain tractable baseline metrics, we optimized their implementation via chunking and FP16 precision with equivalent performance, resulting in the values reported in Table 6. Scaling this paradigm to dense, per-pixel features inevitably causes catastrophic memory explosion, making dense feature registration infeasible for Dr. Splat. In contrast, LESV’s explicit voxel grid and deterministic fusion enable spatial batching, bounding peak memory at ~ 25 GB (~ 14 GB fusion, ~ 11 GB feature extraction) regardless of scene scale.

	ramen figurines teatime waldo kitchen				Average
Num. Images	131	299	177	187	198.5
Dr. Splat [11]	34	176	165	152	131.75
LESV (ours)	25	20	23	22	22.5

Table 6: Peak memory analysis.

Regarding computational speed (Table 7), optimization-based pipelines like LangSplat [22] or LaGa [4] suffer from substantial preprocessing bottlenecks (~ 120 minutes per scene) due to the necessity of extracting multi-level SAM masks and CLIP features. OpenGaussian [30] similarly follows this pipeline but limits extraction to single-level mask features, reducing overhead by $3\times$. Occam’s LGS [5] proposes an efficient way to register features for 2D tasks, but it still necessitates multi-level registration and feature preprocessing. By leveraging dense features from a single forward pass of the AM-RADIO foundation model, LESV drastically reduces per-scene preprocessing time by over $8\times$ to approximately 14 minutes, while our deterministic fusion achieves a feature lifting time of merely ~ 3 minutes. Dr. Splat [11] bypasses per-scene preprocessing via a generalized pre-trained codebook built on the SAM+CLIP pipeline of LangSplat [22], and is therefore denoted “-”.

5 Conclusion

In this paper, we presented LESV, an efficient and accurate framework for open-vocabulary 3D scene understanding. By transitioning to an explicit Sparse Voxel

Method	Preprocessing	Geometry	Feature Lifting
LangSplat [22]	~120 mins	~20 mins	~60 mins
OpenGaussian [30]	~40 mins	~20 mins	~40 mins
LaGa [4] (multi-level)	~120 mins	~20 mins	~70 mins
Occam’s LGS [5]	~120 mins	~20 mins	~3 mins
Dr. Splat [11]	-	~20 mins	~10 mins
LESV (ours)	~14 mins	~15 mins	~3 mins

Table 7: Computational efficiency comparison on a scene of ~ 300 images.

Rasterization (SVRaster) backbone, we successfully mitigate the spatial ambiguity and semantic bleeding artifacts inherent in probabilistic 3DGS registration methods. Our methodology introduces two primary innovations for 3D language fields. First, by leveraging geometry grounded lifting with confidence gating rather than purely visual-guided feature lifting, we enable a robust, view-consistent and efficient feature registration framework, achieving an average feature lifting time of ~ 3 minutes on the LERF dataset. Second, by directly lifting dense, language-aligned features from the AM-RADIO foundation model, we resolve complex part-to-whole semantic ambiguities without the prohibitive computational overhead of hierarchical 2D mask-pooling, reducing per-scene preprocessing time by over $\times 8$ compared to LangSplat [22].

Empirical evaluations across the LERF and ScanNet benchmarks demonstrate that LESV establishes a new SoTA in open-vocabulary point cloud understanding, achieving a substantial gain of $+21.56\%$ mIoU over the leading baseline. For 3D object retrieval, LESV achieves competitive performance against methods that rely on costly multi-level mask preprocessing, while reducing total processing time by 90%. Furthermore, LESV naturally generalizes to 2D object retrieval and localization tasks, achieving performance comparable to SoTA methods optimized specifically for 2D.

6 Limitation

A primary bottleneck of LESV lies in the faithful upsampling of Vision Foundation Model (VFM) features. The effective resolution of our upsampled feature maps is currently limited to 256 (cf. supplementary), which falls short of the spatial resolution provided by SAM masks. This resolution gap hinders performance on scenes containing small or densely packed objects, such as the *figurines* scene. Developing methods for high-fidelity VFM feature upsampling that preserve semantic consistency remains an open and active research challenge. Furthermore, as with all direct registration methods, LESV does not support real-time feature rendering into novel views. While our deterministic fusion excels at 3D-native tasks and generalizes well to 2D projection, a more compact feature representation is needed to enable real-time rendering. Sparse coefficient fields, as proposed in LangSplat-v2 [16], represent a promising direction for future work.

References

1. Alama, O., Bhattacharya, A., He, H., Kim, S., Qiu, Y., Wang, W., Ho, C., Keetha, N.V., Scherer, S.A.: RayFronts: Open-set semantic ray frontiers for online scene understanding and exploration. In: International Conf. on Intelligent Robots and Systems (IROS) (2025)
2. Alama, O., Jariwala, D., Bhattacharya, A., Kim, S., Wang, W., Scherer, S.: RAD-Seg: Unleashing parameter and compute efficient zero-shot open-vocabulary segmentation using agglomerative models. arXiv:2511.19704 (2025)
3. Arafa, A., Stricker, D.: Beyond averages: Open-vocabulary 3d scene understanding with gaussian splatting and bag of embeddings. arXiv preprint arXiv:2509.12938 (2025)
4. Cen, J., Zhou, X., Fang, J., Wen, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Tackling view-dependent semantics in 3d language gaussian splatting. arXiv preprint arXiv:2505.24746 (2025)
5. Cheng, J., Zaech, J.N., Van Gool, L., Paudel, D.P.: Occam’s lgs: An efficient approach for language gaussian splatting. arXiv preprint arXiv:2412.01807 (2024)
6. Chiou, Y.J., Cheng, W.T., Yang, Y.F.: ProFuse: Efficient cross-view context fusion for open-vocabulary 3D gaussian splatting. arXiv:2601.04754 (2026)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2017)
8. 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)
9. Halacheva, A.M., Zaech, J.N., Wang, X., Paudel, D.P., Van Gool, L.: Gaussianvlm: Scene-centric 3d vision-language models using language-aligned gaussian splats for embodied reasoning and beyond. IEEE Robotics and Automation Letters (2025)
10. Huang, S.Y., Choe, J., Wang, Y.C.F., Sun, C.: Openvoxel: Training-free grouping and captioning voxels for open-vocabulary 3d scene understanding. arXiv preprint arXiv:2601.09575 (2026)
11. Jun-Seong, K., Kim, G., Kim, Y.J., Wang, Y.C.F., Choe, J., Oh, T.H.: Dr. Splat: Directly referring 3D gaussian splatting via direct language embedding registration. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2025)
12. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Trans. on Graphics (TOG) (2023)
13. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: LERF: Language embedded radiance fields. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
15. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
16. Li, W., Zhao, Y., Qin, M., Liu, Y., Cai, Y., Gan, C., Pfister, H.: LangSplatV2: High-dimensional 3D language gaussian splatting with 450+ fps. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)

17. Li, Y., Ma, Q., Yang, R., Li, H., Ma, M., Ren, B., Popovic, N., Sebe, N., Konukoglu, E., Gevers, T., et al.: Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2025)
18. Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., El Saddik, A., Theobalt, C., Xing, E., Lu, S.: Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems* **36**, 53433–53456 (2023)
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: Proc. of the European Conf. on Computer Vision (ECCV) (2020)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
21. Peng, Y., Wang, H., Liu, Y., Wen, C., Dong, Z., Yang, B.: GAGS: Granularity-aware feature distillation for language gaussian splatting. In: Proc. of the Conf. on Artificial Intelligence (AAAI) (2026)
22. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: LangSplat: 3D language gaussian splatting. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. of the International Conf. on Machine Learning (ICML) (2021)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
25. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
26. Shen, H., Ni, J., Chen, Y., Li, W., Pei, M., Huang, S.: Trace3d: Consistent segmentation lifting via gaussian instance tracing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6656–6666 (2025)
27. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3D gaussians for open-vocabulary scene understanding. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
28. Sun, C., Choe, J., Loop, C., Ma, W.C., Wang, Y.C.F.: Sparse voxels rasterization: Real-time high-fidelity radiance field rendering. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2025)
29. Wimmer, T., Truong, P., Rakotosaona, M.J., Oechsle, M., Tombari, F., Schiele, B., Lenssen, J.E.: Anyup: Universal feature upsampling. In: International Conf. on Learning Representations (ICLR) Workshop (2026)
30. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., Zhang, J.: OpenGaussian: Towards point-level 3D gaussian-based open vocabulary understanding. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
31. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

32. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2023)
33. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3DGS: Supercharging 3D gaussian splatting to enable distilled feature fields. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)