SOVGaussian: Sparse-View 3D Gaussian Splatting for Open-Vocabulary Scene Understanding

Peng Ling, Tiao Tan, Jiaqi Lin, Wenming Yang*

Shenzhen International Graduate School, Tsinghua University
{lingp23, tt23, linjq22}@mails.tsinghua.edu.cn, yang.wenming@sz.tsinghua.edu.cn

Abstract

Modeling 3D open-vocabulary language fields is challenging yet highly anticipated. Despite great progress, existing approaches heavily rely on a large number of training views to construct language-embedded 3D scenes, which is unfortunately impractical in real-world scenarios. This paper introduces SOVGaussian, the first method for few-shot novel view open-vocabulary language querying. We introduce a depth-constrained neural language field to mitigate the geometry degradation caused by overfitting training views. Rather than straightforwardly using dense depth maps for loosely accurate supervision, Language-Aware Depth Distillation (LAD) based on open-vocabulary object masks is proposed, ensuring intra-object geometric accuracy within the language field. To further refine the language-geometry consistency of the language field, we propose a novel Language-Guided Outlier Pruning (LOP) strategy, which identifies floating 3D Gaussian primitives overfitting training views based on their language-grouped densities. Our comprehensive experiments demonstrate that SOVGaussian is able to reconstruct a superior scene representation from few-shot images, outperforming existing state-of-the-art methods and achieving significantly better performance on novel view language querying and synthesis.

Repository — https://github.com/Brucess/SOVGaussian

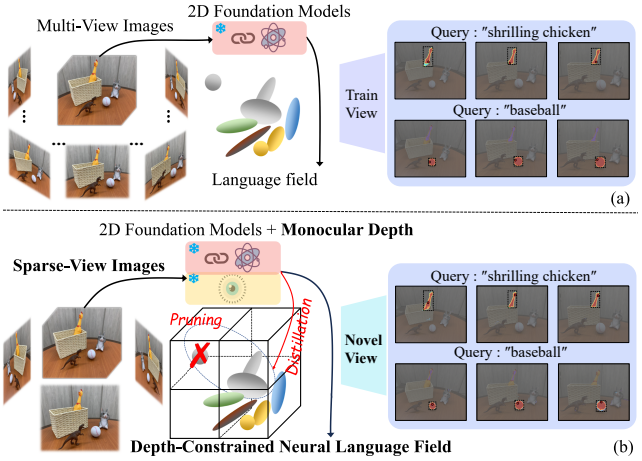

Figure 1: Comparing existing approaches (a) with our method (b): Existing methods input full views to 2D foundational models for language field training, and test open-vocabulary querying accuracy on the training views. In contrast, our method is trained on 3 input views, introduces a depth-constrained neural language field, and leverages the intrinsic correlation between language and geometry for depth distillation and primitive pruning, achieving high-quality open-vocabulary querying on novel views.

Introduction

Language-embedded scene understanding and reconstruction are promising tasks for computer vision, enabling the users interact with 3D virtual world via open-ended language in a high-quality manner. Recently, Neural Radiance Field (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) have emerged as promising method of scene reconstruction from multi-view input images. Previous works have attempted to embed language into neural representation(Liu et al. 2023) or 3D Gaussian (Qin et al. 2024; Shi et al. 2024), leveraging the image and language embeddings from 2D foundational visual-language models such as CLIP (Radford et al. 2021) and DINO (Caron et al. 2021) to interact with users using open-vocabulary queries. Language-embedded 3D scene representations, especially those based on 3DGS that achieve faster rendering speed and comparable image quality, have significantly enhanced applicability in downstream tasks such as embodied intelligence (Fan et al. 2024), autonomous robotic navigation (Zhao et al. 2024), and autonomous driving (Vobecky et al. 2024; Huang et al. 2024).

Nevertheless, existing methods concentrate on developing pipelines leveraging dense input views and do not emphasize the ability to render language features in novel views. Imposing increasingly stringent constraints on input data results in higher operational costs in practical applications, which is unacceptable in many cases, such as in autonomous vehicles or mobile robots where cameras operate at only a few Hertz (Caesar et al. 2020). However, when the number of views is significantly reduced (e.g., only 3 images), the 3DGS-based language field deteriorates into an ill-posed structure that overfits the input views, causing background

*Corresponding authors.

collapse and excessive floaters on both rendered language feature maps and RGB images (Xiong et al. 2023), failing to support language queries in novel views. An intuitive solution is to supplement additional geometric constraints as cues to mitigate the significant scene degradation caused by the few-shot input images (Turkulainen et al. 2024). Inspired by the success of the previous depth-constrained NeRFs (Guo et al. 2024; Wang et al. 2023) and methods of 3D Gaussians (Li et al. 2024), depth supervision from pretrained monucular depth estimator provides substantial value. Despite the considerable success of monocular depth estimation methods in recent years (Yang et al. 2024; Ke et al. 2024; Shao et al. 2024), monocular depth estimation is theoretically under-determined and cannot generalize accurate depth information across different datasets and scenes, undermining its effectiveness on reshaping geometry. Benefiting from our language information, we can exploit the intrinsic correlation between language features and geometry to impose additional regularization on the scene representation.

In this paper, we introduce 3D Gaussian-based SOVGaussian to tackle the aforementioned challenges. Diverging from existing approaches, our method exclusively leverages sparse-view input for training while ensuring precise open-vocabulary querying and high-fidelity synthesis in novel views. A comprehensive comparison between our approach and existing methods is presented in Figure 1. In SOVGaussian, we integrate dense monocular depth estimation maps to construct a constrained neural language field, effectively mitigating the degradation caused by overfitting training views. Given the numerical inaccuracy of this cue, we eschew using it as hard constraint. Instead, we propose a Language-Aware Depth Distillation module (LAD), which explicitly constrains the position optimization of 3D Gaussian primitives by distilling relative depth within open-vocabulary object masks, ensuring that the geometry can be correctly supervised within the object scale. Exploiting the intrinsic correlation between language feature and geometry, we further propose a Language-Guided Outlier Pruning Strategy (LOP), which identifies 3D Gaussian primitives overfitting training views based on their language-grouped densities. Consequently, this module effectively eliminates floating primitives that induce artifacts in novel views. In summary, our contributions include:

- We present SOVGaussian, to the best of our knowledge, is the first method designed for few-shot novel view open-vocabulary language querying.
- We introduce a depth-constrained neural language field, thereby improving the geometry reshaping and language optimization. To resolve the ambiguity caused by numerical inaccuracies of monocular depth cue, we propose a depth distillation module based on open-vocabulary object masks.
- Leveraging the intrinsic correlation between scene geometry and language features, we further propose an outlier pruning strategy to eliminate floating 3D Gaussian primitives.
- Experimental results demonstrate that our method outperforms existing state-of-the-art methods, achieving up to a 56.9% improvement in mIoU compared to LangSplat on the 3DOVS dataset and up to a 36% improvement on the DTU dataset.

Related Work

3D Scene Open Vocabulary Understanding

Progress in 3D object detection and segmentation (Nguyen et al. 2024; Lu et al. 2023; Takmaz et al. 2023; Cao et al. 2024) highlights the effectiveness of integrating point clouds with features for scene understanding. However, these methods focus on analyzing pre-existing point cloud representations, neglecting reconstruction. Early efforts to integrate reconstruction with open-vocabulary scene understanding primarily leverage Neural Radiance Fields (NeRF). LERF (Kerr et al. 2023) embeds multi-scale CLIP features into NeRF for consistent 3D scene understanding, while 3DOVS (Liu et al. 2023) aligns CLIP and DINO features via Relevancy-Distribution and Feature-Distribution Alignment losses to enable 3D open-vocabulary querying. However, NeRF-based methods are limited by implicit 3D representations, resulting in costly training.

3D open vocabulary understanding based on 3DGS aims to improve the quality of queried object segmentation and significantly reduce the time required for optimizing scene representation. Feature 3DGS (Zhou et al. 2024) outlines a high-dimensional semantic feature rendering process and speed-up approach leveraging a parallel N-dimensional Gaussian rasterizer and a convolutional speed-up module, enabling efficient feature field distillation guided by 2D foundation models. A quantization scheme proposed by LEGaussians (Shi et al. 2024) compresses semantic features into a compact feature space, and the spatial frequency of semantic features is reduced based on learned uncertainty values. LangSplat (Qin et al. 2024) learns hierarchical semantic language features from SAM (Kirillov et al. 2023) in the scene-specific latent space through a scene-wise autoencoder to reduce memory cost and point ambiguity issues. However, these 3DGS-based works do not consider the generalization capability for novel views and require a large number of images and camera views to train the field, which is unavailable under stringent data conditions in practical applications.

Novel View Synthesis from Sparse Training Views

The task of novel view synthesis utilizes a small number of training views to train 3D scene representations, aiming to achieve high-quality novel view synthesis. Existing works using NeRF (Deng et al. 2022; Roessle et al. 2022) and 3DGS as the 3D representation model introduce depth cues to explicitly regularize the geometry of 3D Gaussians, achieving more realistic novel view RGB image synthesis. A optimization strategy (Chung, Oh, and Lee 2024) improves the quality of 3D scene reconstruction with limited image input through depth guidance and unsupervised smooth constraints. Other approaches (Li et al. 2024; Zhang et al. 2024; Paliwal et al. 2025) attempt to address the issue of geometric degradation in 3D Gaussian Splatting by introducing

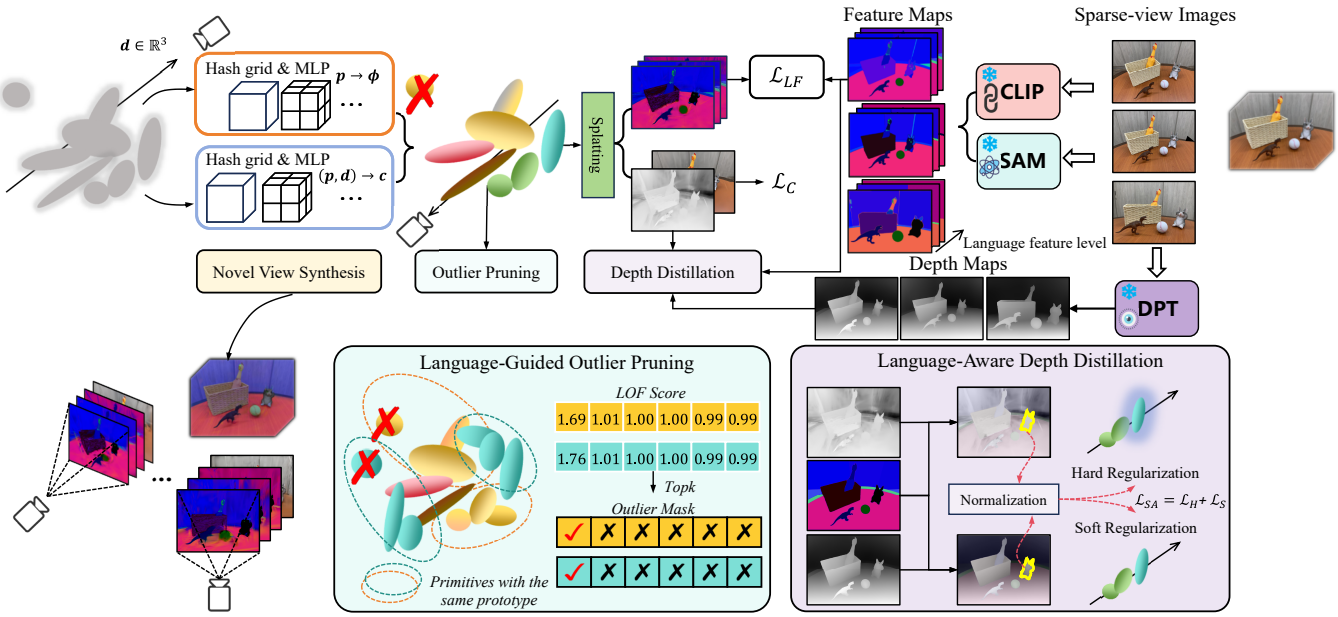

Figure 2: We present the pipeline of our SOVGaussian. From the left: SOVGaussian starts with random initialization. Language and color features of each 3D Gaussian primitive are queried using the learnable position. From the right: The sparse-view (i.e., 3 views) RGB input images are processed by CLIP and SAM to extract dense language feature maps, and by a monocular depth estimator to obtain dense depth maps. During optimization, the learnable parameters of the primitives are optimized using language feature loss, color loss, and depth loss through Language-Aware Depth Distillation (LAD). Additionally, Language-Guided Outlier Pruning (LOP) is used to remove outliers that exhibit language-geometry inconsistencies.

regularization and normalization. These works provide valuable insights into explicitly regularizing the geometry of 3D Gaussians. However, the constructed 3D scene representations are limited to synthesizing RGB images. Our paper focuses on constructing a 3D language field from sparse-view input, which leverages the intrinsic correlation between language features and geometry to achieve high-quality open-vocabulary querying and synthesis in novel views.

Method

Previous open-vocabulary scene understanding methods are limited by the necessity of all training views of a scene to train the scene representation. Our proposed method aims to use sparse-view RGB images (i.e., 3 views) as training inputs to obtain a 3DGS-based scene representation with advanced generalization capability for novel views. Figure 2 provides an overview of our method. In this section, we first introduce our depth-constrained neural language field. Subsequently, we will introduce a Language-Aware Depth Distillation module and a Language-Guided Outlier Pruning strategy, tightly coupling the intrinsic correlation between language feature and geometry to reshape the language field.

Depth-Constrained Neural Language Field

To construct a 3D language field from sparse-view input images $\{\boldsymbol{I}_i \mid i = 1, 2, ...N_I, \boldsymbol{I}_i \in \mathbb{R}^{3\times H\times W}\}$ (e.g., $N_I = 3$), a pipeline that extracts dense, pixel-level language features from 2D images is required. We utilize SAM and CLIP to construct hierarchical dense language features from sparse input images. Mathematically, the obtained dense language features:

$$\boldsymbol{H}_i^l(x) = f_{CLIP}(\boldsymbol{I}_i \boxdot f_{SAM}^l(x)) \tag{1}$$

where $f_{SAM}^l(x)$ is the l semantic level (i.e., "subpart", "part", "whole") mask region at pixel x generated by SAM, and $f_{CLIP}(\cdot)$ represents image encoding by CLIP.

In existing methods requiring full-view data, the training process emphasizes overfitting to the input images. Therefore, the optimal language embedding is a low-dimensional representation derived from high-dimensional discrete features through one-to-one dimensionality reduction quantization (Shi et al. 2024; Wu et al. 2024). However, this quantized representation is undesirable in the case of sparse training views because the low-dimensional index itself does not contain semantics. In the scenario of sparse training views, the inevitable semantic inconsistencies in training images can lead to completely incorrect index generation during novel view synthesis. To prevent this issue, rather than discretizing original CLIP feature, we use an autoencoder (Qin et al. 2024) to reduce the dimensionality of the original high-dimensional (i.e., 512-dimensional) CLIP feature space into a low-dimensional (i.e., 3-dimensional) Hilbert space:

$$\boldsymbol{E}_i^l = f_{autoencoder}(\boldsymbol{H}_i^l) \tag{2}$$

where $\boldsymbol{E}_i^l \in \mathbb{R}^{3\times H\times W}$ is the final language embeddings for $\boldsymbol{I}_i$.

We next use monocular depth cue to explicitly supervise the language field and color field reconstrction, thereby reducing erroneous primitives used for overfitting the training

views. We directly generalize the monocular depth estimator DPT (Ranftl, Bochkovskiy, and Koltun 2021) for depth supervision, without introducing additional manually annotated data. To tackle the issue of numerical inaccuracies at the global scale of the estimated monocular depth, we further propose a Language-Aware Depth Distillation module (LAD); see the next subsection.

Simple spatial domain fitting method (e.g., SH coefficients in Vanilla 3DGS) is not robust and can lead to severe color distortion under sparse view conditions (Liu et al. 2024; Chen et al. 2024). This may be due to the underdetermined optimization problem caused by the sparse-view constraints. Previous works (Lee et al. 2024; Jiang et al. 2024) prove that neural renderer contributes to enhancing color generalization for novel views, which inspires us. Instead of simply embed language features in primitives like previous works, we exploit hash grids (Müller et al. 2022) to query each primitive's language and color features. We input positions into hash grids to query color features and language features. Formally, the view-dependent color and view-independent language features of the primitive at position p_j are:

$$c_j = \mathbf{MLP}_C(h_{color}(p_j;\theta), d) \tag{3}$$

$$\phi_j^l = \mathbf{MLP}_L(h^l_{language}(p_j;\theta)) \tag{4}$$

where θ denotes the parameters of the neural field, $h_{\cdot}(\cdot;\cdot)$ is tensor indexing operation, and d is the view direction. c_j and ϕ_j^l are the color and the l-th language feature of the j-th primitive. The objective function of neural language field optimization is:

$$\mathcal{L}_{LF} = \sum_l \sum_i \|\boldsymbol{E}_i^l - \Phi_i^l\|_2 \tag{5}$$

where $\Phi_i^l \in \mathbb{R}^{3\times H\times W}$ is the rendered language feature map through rasterizer. The grid-based structure of the hash grids allows for the optimization of shared multi-resolution corner features when optimizing each primitive, enabling different primitives to deform coherently during optimization and enhancing the regional consistency of the language field.

Language-Aware Depth Distillation

Monocular depth estimators theoretically lack the capabilities to recover depth with accurate numerical values, leading to conflicts in the optimization, undermining its effectiveness on reshaping geometry. Even the estimated relative depth maps can exhibit inaccuracies at the global scale. As depicted in Figure 2, the estimated relative depth maps lack precision. For novel view querying, incorrectly positioned Gaussian primitives can become artifacts like floaters, significantly degrading integrity of object mask, especially on detail-rich object surfaces. This is detrimental to our object-centric open-vocabulary querying.

To solve this problem, we propose Language-Aware Depth Distillation (LAD) module to make the position correction process of the primitives more robust and detail-focused. We believe that the depth ordering within local regions of the estimated relative depth maps is accurate, but the numerical values may not be. We distill the depth map by segmenting it according to the object masks output by SAM and normalize the relative depth map within each mask:

$$\boldsymbol{D}^l_{SA}(x) = \frac{\boldsymbol{D}(x) - min(\boldsymbol{D}(f^l_{SAM}(x)))}{max(\boldsymbol{D}(f^l_{SAM}(x))) - min(\boldsymbol{D}(f^l_{SAM}(x)))} \tag{6}$$

where $\boldsymbol{D} \in \mathbb{R}^{1\times H\times W}$ is the depth map. Since then, rendered depth map and pseudo ground truth both are modify to the same object scale. By focusing on the relative depth at the object scale, we mitigate the numerical inaccuracies of the estimated global relative depth map within the region mask.

To enable the depth loss to directly correct the positions of primitives during backpropagation, we employed hard depth regularization and soft depth regularization (Li et al. 2024). Specifically,

$$\boldsymbol{D}_{\text{hard}}(x) = \sum_{j\in N_{\mathcal{G}}} \mu(1-\mu)^{j-1}\mathcal{G}_j^{2D}(x)\,\|p_j - o\|_2 \tag{7}$$

where $\boldsymbol{D}_{\text{hard}} \in \mathbb{R}^{1\times H\times W}$ is the rendered depth map, μ is hard opacity that is close to 1 (e.g., 0.99), o denotes the camera center, and $\mathcal{G}_j^{2D}$ is the projected 2D Gaussian from 3D Gaussian $\mathcal{G}_j$ of the j-th primitive . Similarly,

$$\boldsymbol{D}_{\text{soft}}(x) = \sum_{j\in N_{\mathcal{G}}} \alpha(1-\alpha)^{j-1}\mathcal{G}_j^{2D}(x)\,\|p_j - o\|_2 \tag{8}$$

where α is the original opacity of $\mathcal{G}_j$.

The total loss function $\mathcal{L}_D$ of depth constraint is defined by:

$$\mathcal{L}_{SA} = \sum_l \left\|\boldsymbol{D}^l_{gt,SA} - \boldsymbol{D}^l_{hard,SA}\right\|_2 + \left\|\boldsymbol{D}^l_{gt,SA} - \boldsymbol{D}^l_{soft,SA}\right\|_2 \tag{9}$$

$$\mathcal{L}_{GL} = \|\boldsymbol{D}_{gt,GL} - \boldsymbol{D}_{hard,GL}\|_2 + \|\boldsymbol{D}_{gt,GL} - \boldsymbol{D}_{soft,GL}\|_2 \tag{10}$$

$$\mathcal{L}_D = \lambda\mathcal{L}_{SA} + \mathcal{L}_{GL} \tag{11}$$

where $\boldsymbol{D}_{\cdot,GL}$ is global nolmalization.

Language-Guided Outlier Pruning

During optimization, some 3D Gaussian primitives appear in non-occupied areas to fit the training views. Sparse input views do not provide sufficient constraints for the optimization process to remove these outlier primitives. These floating primitives significantly hinder novel view synthesis, forming artifacts. Previous few-shot 3D Gaussian splatting methods, limited by 2D supervisions, could not eliminate these erroneous primitives. Proposed Language-Guided Outlier Pruning (LOP) strategy incorporates language features and densities of primitives to identify these floating outliers, constraining the geometry of the 3D Gaussian field. Specifically, we first obtain the prototypes of the ground truth language embeddings for all primitives:

$$\mathcal{PRO}^l = \{\boldsymbol{E}_i^l(x) \mid 1 \le x \le HW, 1 \le i \le N_I\} \tag{12}$$

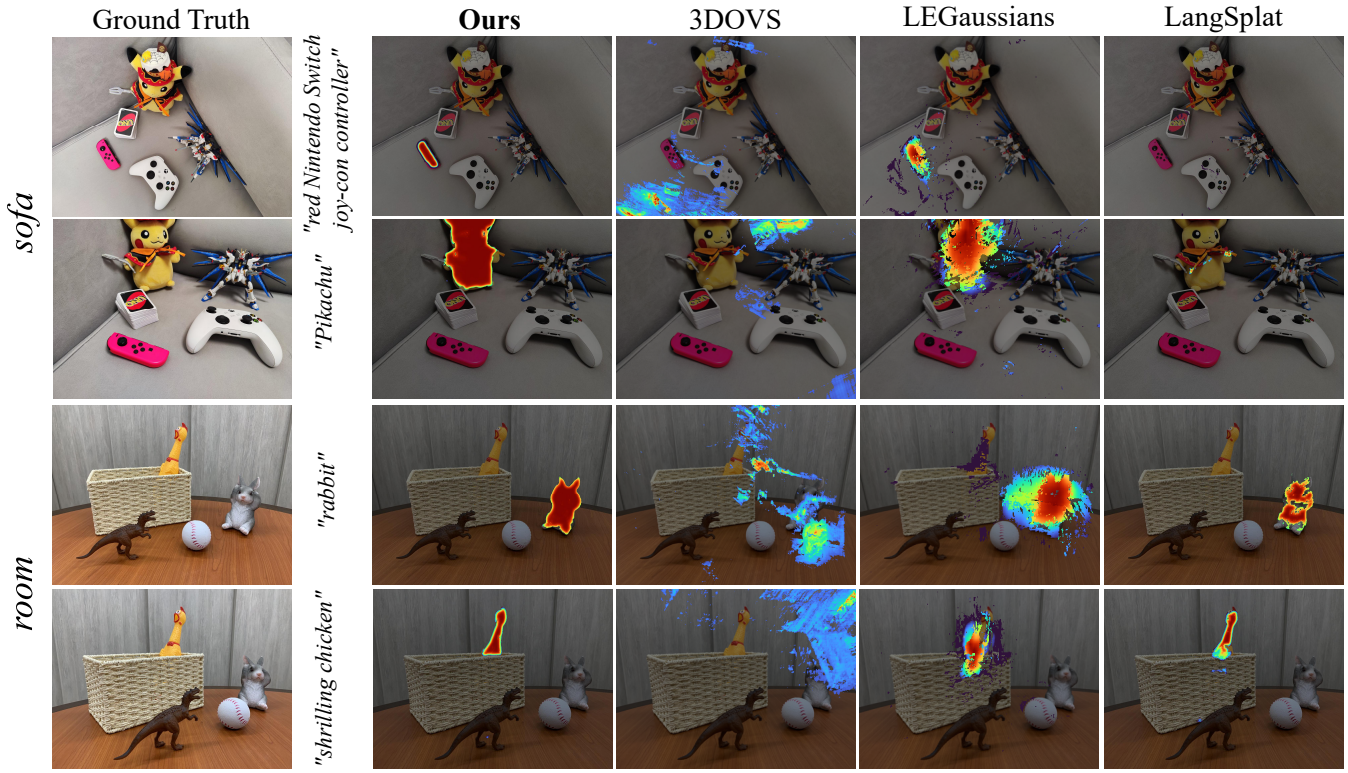

Figure 3: Qualitative comparisons of open-vocabulary querying accuracy in novel views on the 3DOVS dataset. We visualized the relevance maps of two novel views for each of the "sofa" and "room" scenes.

where $\boldsymbol{E}_i^l(x) \in \mathbb{R}^3$, all unique $\boldsymbol{E}_i^l(x)$ constitute $\mathcal{PRO}^l = \{\mathbf{P}_k^l\}$ representing the set of l-level language features. Then, we retrieve all primitives in the 3D Gaussian field and match the prototype corresponding to each of them:

$$\mathcal{I}_j^l = \arg\min_k \|\mathbf{P}_k^l - \phi_j^l\|_2 \tag{13}$$

where $\mathcal{I}_j^l$ is the corresonding prototype index of the j-th primitive.

Our goal is to identify outliers within these primitives by treating those with the same prototype as a basic group. Considering that they exhibit a multi-center distribution in the spatial domain, a suitable and significant characteristic for distinguishing floating outliers from normal primitives is the density of primitive because primitives in the same group should have similar densities. The LOP calculates the Local Outlier Factor (LOF) of each primitive to determine whether it is an outlier:

$$LOF(\mathcal{G}_m) = \frac{\sum_{\mathcal{G}_u \in \text{KNN}(\mathcal{G}_m)} LRD(\mathcal{G}_u)}{\gamma LRD(\mathcal{G}_m)}, \quad \text{s.t.} \quad \mathcal{I}_u^l = \mathcal{I}_m^l \tag{14}$$

where KNN$(\cdot)$ denotes γ-nearest neighbors, and γ is a hyper-parameter. The local reachability density is:

$$LRD(\mathcal{G}_m) = \frac{\gamma}{\sum_{\mathcal{G}_u \in \text{KNN}(\mathcal{G}_m)} \|p_m - p_u\|_2} \tag{15}$$

If $\mathcal{G}_m$ is in a sparser region compared to its neighbors, the LOF score will be greater than 1, indicating that p is an outlier. We set an outlier ratio τ to compute the outlier mask:

$$\mathcal{M}_{outlier} = \arg\text{topk}(LOF(\mathcal{G}_m), \tau) \tag{16}$$

With the outlier mask, we prune the floating primitives to prevent them from being used for overfitting the training views and from deteriorating the generalization of novel views.

Optimizing Details

Our model is trained end-to-end, with the overall loss divided into three components: color loss $\mathcal{L}_C$, language feature loss $\mathcal{L}_{LF}$, and depth loss $\mathcal{L}_D$:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{LF} + \mathcal{L}_D \tag{17}$$

where $\mathcal{L}_C = \mathcal{L}_1 + \mathcal{L}_{D-SSIM}$ is the basic color supervision (Kerbl et al. 2023). The $\mathcal{L}_{LF}$ is designed to only backpropagate to the hash grids features, without generating gradients for other optimizable parameters of the primitives. We set the interval for LOP to 1000 iterations, thus the total time overhead introduced by it is negligible. We empirically set γ to 20, $\tau = 5\%$ for LOP, and $\lambda = 0.1$ for the loss function.

Experiments

Settings

Datasets We evaluate our method on the 3DOVS (Liu et al. 2023) and DTU datasets (Aanæs et al. 2016). The

Methods	sofa		bed		room		lawn		bench		**Overall**	
	mIOU	mAcc	mIOU	mAcc	mIOU	mAcc	mIOU	mAcc	mIOU	mAcc	mIOU	mAcc
3DOVS	19.9	30.0	13.1	20.0	23.3	33.3	13.3	26.7	10.8	28.6	16.1	27.7
LEGaussians	24.2	40.0	25.9	50.0	26.1	53.3	34.7	60.0	51.2	68.6	32.4	54.4
LangSplat	39.0	83.3	30.2	86.7	54.9	93.3	31.2	66.7	54.6	97.1	42.0	85.4
Ours	**69.3**	**99.2**	**48.4**	**90.0**	**72.4**	**99.2**	**75.9**	**98.3**	**63.7**	**98.6**	**65.9**	**97.1**

Table 1: Performance of semantic querying in novel views on the 3DOVS dataset. We report the mIoU↑ scores (%) and localization mAcc↑.

3DOVS dataset comprises a collection of scenes featuring long-tail objects captured from different views, designed for open-vocabulary scene understanding and object localization. We follow the LangSplat evaluation protocol and conduct assessments on the same scenes for mean IoU and localization accuracy. The resolution is set at 1440 $\times$ 1080, consistent with the original LangSplat. We further employ the DTU dataset, which is widely used for evaluating methods for sparse-view input. Given that the original DTU dataset does not provide open-vocabulary semantic annotations, we select 7 complex scenes and manually annotate them to facilitate model evaluation. The resolution is maintained at 1440 $\times$ 1080 to control for variables.

Baselines We compare our method against several state-of-the-art models, including LangSplat, LEGaussians, and 3DOVS. Different from their vanilla pipelines that use all views (i.e., 35 for 3DOVS and 49 for DTU) for training, we use only 3 views and evaluate generalization on novel views. To ensure fair comparison, all methods are trained following the same sparse-view protocol as ours, using the same 3 input views, camera poses, and test views. It is worth mentioning that the baselines are based on publicly available source codes and papers, with modifications made to the dataset splits to enable sparse-view training. We further control hyperparameters such as learning rate and density increment percentage to enhance the baselines' performance.

Implementation Details Our approach is based on 3DGS (Kerbl et al. 2023) and implemented by PyTorch. We modifiy the rasterization pipeline to simultaneously obtain rendered depth maps and language feature maps. Our method does not require separate training for scene geometry and the language feature field. Instead, it achieves a single training process for a scene. We train for 20,000 iterations on the 3DOVS dataset and 6,000 iterations on the DTU dataset using a single RTX 3090, requiring approximately 1 hour and 25 minutes, respectively, using around 4GB of memory. To address low co-visibility issues that may cause COLMAP (Schönberger et al. 2016) to fail, we initialize with a random point cloud. View selection follows uniform sampling for 3DOVS and the protocol in (Li et al. 2024) for DTU.

Comparisons on The 3DOVS Dataset

Quantitative Results Tabel 1 presents the cross-metric comparisons of our method with other approaches on the 3DOVS dataset, covering the localization accuracy and mIOU for open-vocabulary queries. Our observations show that SOVGaussian achieves an overall mIOU of 65.9%, significantly outperforming the latest methods, with a margin of 23.9% mIOU (**+56.9%**) over LangSplat and 33.5% mIOU (**+103.4%**) over LEGaussians. The baselines primarily focus on overfitting to the training views, while they fail to leverage sparse training views to construct a geometrically consistent language field, leading to poor performance on novel views. Our method is designed to align with practical application conditions, emphasizing generalization to novel views. By incorporating the proposed LAD and LOP, we enhance the geometric quality of the language field and suppress floaters that cause artifacts in novel views, thereby improving the model's generalization.

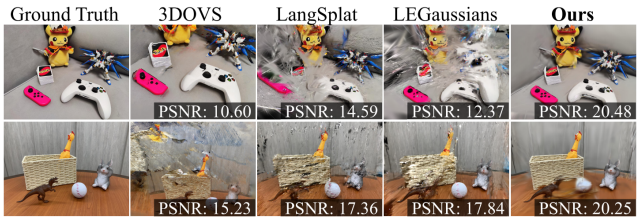

Figure 4: Visual quality comparison of novel view synthesis results on 3DOVS dataset. Our method demonstrates significantly better novel view synthesis capability.

Qualitative Results We visualize the qualitative results in Figure 3 and Figure 4. Our observations show that SOVGaussian achieves more precise and less holey query masks in novel views, maintaining clear object boundaries. Specifically, 3DOVS almost fails to locate the queried object's position, likely caused by the significant degradation of NeRF-based implicit representations under sparse-view constraints. Moreover, LEGaussians and LangSplat produce query masks with lower precision and more holes, attributable to their constructed language fields exhibiting chaotic geometric structures and being severely affected by floaters.

Comparisons on The DTU Dataset

Quantitative Results To validate the generalizability of our method across different datasets, we numerically compare our method against the baselines on the DTU dataset, as shown in Table 2. Once again, our method achieves an overall superior mIOU of 71.7% than other approaches.

menthods	mIOU							
	21	31	38	41	55	103	110	**Overall**
3DOVS	27.5	8.8	39.6	36.6	19.4	F	18.7	21.5
LEGaussians	38.4	20.4	44.1	18.4	1.9	0.5	25.0	21.2
Langsplat	38.4	56.6	61.0	58.3	45.0	68.8	40.7	52.7
Ours	**46.5**	**67.2**	**72.6**	**90.2**	**68.9**	**83.9**	**72.4**	**71.7**

Table 2: The semantic querying mIOU↑ in novel views of 7 scenes. Annotations are used solely for evaluation and are not visible during training. 3DOVS fails in scan103 since it is unable to query objects that only appear in partial views.

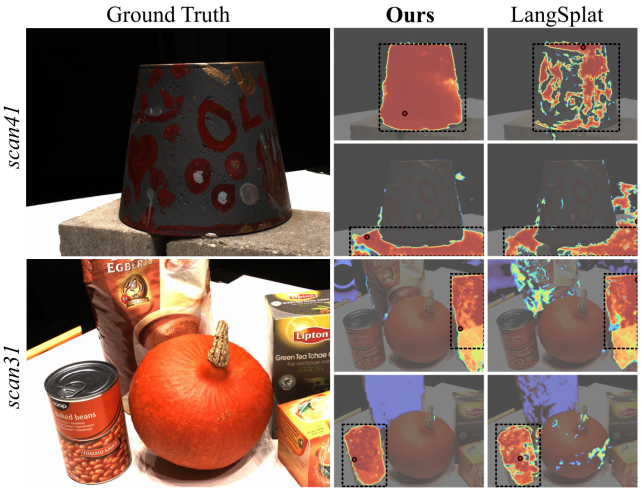

Figure 5: Qualitative comparison of our method and LangSplat. Top to bottom: query words "metal bucket with painted patterns", "concrete block", "cardboard box of Lipton green tea", "metal can of baked beans". The red points indicate the model's predictions, while the black dashed bounding boxes represent the annotations.

Qualitative Results Figure 5 illustrates the qualitative results of our method. In the scan41 scene, we query "metal bucket with painted patterns." Our method responds with a complete object mask, featuring clear object edges. The peak response is located in the area with painted patterns, indicating that our method effectively constructs a language field with robust generalization capabilities for novel views.

Ablation Study

Here, we conduct ablations on the 3DOVS dataset to evaluate the performance increment contributed by each component, including open-vocabulary querying accuracy and synthesis quality from novel views. The quantitative results are presented in Table 3. In Case #3, without using the depth cue, all metrics are significantly lower than those of the full model. Case #1 demonstrates the effectiveness of the LOP by ablating it individually. In Case #2, we observe performance improvements attributed to the LAD's enhancement of scene geometry. In Case #4, replacing the hash grids of language features with 3-dimensional embeddings results in relatively unchanged novel view synthesis capabilities but a decline in querying accuracy.

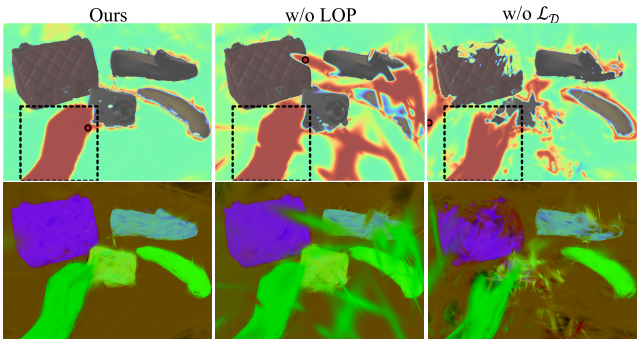

Figure 6: We visualized the query relevancy for "hand" in the novel views (top part) and the rendered feature maps of the highest responsive level (bottom part), with all of them at the "part" level.

Case		mIOU↑	mAcc↑	PSNR↑	SSIM↑
#1	w/o LOP	61.8	95.3	16.57	0.4661
#2	w/o LAD	63.4	95.5	17.14	0.4743
#3	w/o $\mathcal{L}_D$	46.2	88.2	15.00	0.4281
#4	w/o ϕ_j	64.5	95.3	17.52	0.4854
#5	Ours	**65.9**	**97.1**	**17.63**	**0.4872**

Table 3: Quantitative results of ablation experiments in novel views.

In Figure 6, we further visualize the effects of the depth cue and LOP. When the depth cue is ablated, the degradation of the language field's geometry leads to rendered novel view with severe object language features distortion. However, thanks to the LOP, artifacts are suppressed. When the LOP is disabled, the rendered novel view language feature maps exhibit numerous floaters, significantly impacting querying accuracy, although the depth cue enhances the integrity of the object features.

Conclusion and Limitation

To the best of our knowledge, we are the first to tackle open-vocabulary scene understanding from sparse-view input. We propose SOVGaussian, a method that constructs a 3D language field from few-shot inputs, enabling precise open-vocabulary querying and high-fidelity novel view synthesis. By integrating a monocular depth estimator and hash grids, SOVGaussian builds a depth-constrained neural language field, enhancing geometry reshaping and language optimization. We further introduce Language-Aware Depth Distillation (LAD) to address depth inaccuracy and propose Language-Guided Outlier Pruning (LOP) to improve geometric precision by mitigating floaters caused by overfitting. Experiments demonstrate SOVGaussian's superior novel view generalization over state-of-the-art methods. However, SOVGaussian relies on static primitives and excludes dynamic scenes.

Acknowledgments

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen (No. KJZD20231023094700001).

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-Scale Data for Multiple-View Stereopsis. *International Journal of Computer Vision*, 1–16.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cao, Y.; Yihan, Z.; Xu, H.; and Xu, D. 2024. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems*, 36.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, Y.; Xu, H.; Zheng, C.; Zhuang, B.; Pollefeys, M.; Geiger, A.; Cham, T.-J.; and Cai, J. 2024. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*.

Chung, J.; Oh, J.; and Lee, K. M. 2024. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Fan, L.; Zhou, J.; Xing, X.; and Wu, Y. 2024. Active Open-Vocabulary Recognition: Let Intelligent Moving Mitigate CLIP Limitations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16394–16403.

Guo, S.; Wang, Q.; Gao, Y.; Xie, R.; and Song, L. 2024. Depth-Guided Robust and Fast Point Cloud Fusion NeRF for Sparse Input Views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1976–1984.

Huang, Y.; Zheng, W.; Zhang, B.; Zhou, J.; and Lu, J. 2024. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19946–19956.

Jiang, Y.; Tu, J.; Liu, Y.; Gao, X.; Long, X.; Wang, W.; and Ma, Y. 2024. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5322–5332.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9492–9502.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19729–19739.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lee, J. C.; Rho, D.; Sun, X.; Ko, J. H.; and Park, E. 2024. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21719–21728.

Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20775–20785.

Liu, K.; Zhan, F.; Zhang, J.; Xu, M.; Yu, Y.; El Saddik, A.; Theobalt, C.; Xing, E.; and Lu, S. 2023. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36: 53433–53456.

Liu, T.; Wang, G.; Hu, S.; Shen, L.; Ye, X.; Zang, Y.; Cao, Z.; Li, W.; and Liu, Z. 2024. Fast Generalizable Gaussian Splatting Reconstruction from Multi-View Stereo. *arXiv preprint arXiv:2405.12218*.

Lu, Y.; Xu, C.; Wei, X.; Xie, X.; Tomizuka, M.; Keutzer, K.; and Zhang, S. 2023. Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1190–1199.

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

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

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

Paliwal, A.; Ye, W.; Xiong, J.; Kotovenko, D.; Ranjan, R.; Chandra, V.; and Kalantari, N. K. 2025. Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. In *European Conference on Computer Vision*, 19–37. Springer.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12892–12901.

Schönberger, J. L.; Zheng, E.; Pollefeys, M.; and Frahm, J.-M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Shao, S.; Pei, Z.; Wu, X.; Liu, Z.; Chen, W.; and Li, Z. 2024. Iebins: Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing Systems*, 36.

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

Takmaz, A.; Fedele, E.; Sumner, R. W.; Pollefeys, M.; Tombari, F.; and Engelmann, F. 2023. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*.

Turkulainen, M.; Ren, X.; Melekhov, I.; Seiskari, O.; Rahtu, E.; and Kannala, J. 2024. DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing. *arXiv preprint arXiv:2403.17822*.

Vobecky, A.; Siméoni, O.; Hurych, D.; Gidaris, S.; Bursuc, A.; Pérez, P.; and Sivic, J. 2024. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36.

Wang, G.; Chen, Z.; Loy, C. C.; and Liu, Z. 2023. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9065–9076.

Wu, Y.; Meng, J.; Li, H.; Wu, C.; Shi, Y.; Cheng, X.; Zhao, C.; Feng, H.; Ding, E.; Wang, J.; et al. 2024. OpenGaussian: Towards Point-Level 3D Gaussian-based Open Vocabulary Understanding. *arXiv preprint arXiv:2406.02058*.

Xiong, H.; Muttukuru, S.; Upadhyay, R.; Chari, P.; and Kadambi, A. 2023. Sparsegs: Real-time 360 {\deg} sparse view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00206*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Zhang, J.; Li, J.; Yu, X.; Huang, L.; Gu, L.; Zheng, J.; and Bai, X. 2024. CoR-GS: Sparse-View 3D Gaussian Splatting via Co-Regularization. *arXiv preprint arXiv:2405.12110*.

Zhao, G.; Li, G.; Chen, W.; and Yu, Y. 2024. OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and StructurEd Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16296–16306.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2024. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21676–21685.