SPC-GS: Gaussian Splatting with Semantic-Prompt Consistency for Indoor Open-World Free-view Synthesis from Sparse Inputs

Guibiao Liao[1,2] Qing Li[2] Zhenyu Bao[1,2] Guoping Qiu[3] Kanglin Liu[2*]
[1]School of Electronic and Computer Engineering, Peking University
[2]Pengcheng Laboratory, [3]University of Nottingham

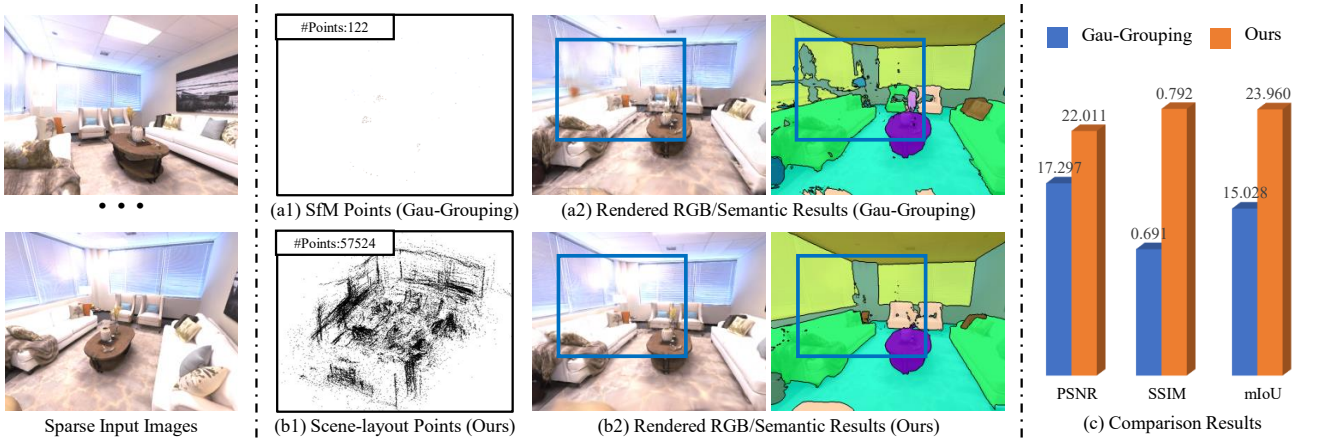

Figure 1. Visual comparisons of open-world free-view synthesis from sparse inputs (12 training views). The previous 3DGS-based method, Gau-Grouping [50], utilizes sparse SfM points for Gaussian initialization and limited training views for supervision, leading to inferior rendering results. In contrast, our approach leverages scene-layout-based points for instructive initialization, and cooperates with semantic-prompt consistency constraints (detailed in Sec. 3.3), yielding superior reconstruction and segmentation results.

Abstract

3D Gaussian Splatting-based indoor open-world free-view synthesis approaches have shown significant performance with dense input images. However, they exhibit poor performance when confronted with sparse inputs, primarily due to the sparse distribution of Gaussian points and insufficient view supervision. To relieve these challenges, we propose SPC-GS, leveraging Scene-layout-based Gaussian Initialization (SGI) and Semantic-Prompt Consistency (SPC) Regularization for open-world free view synthesis with sparse inputs. Specifically, SGI provides a dense, scene-layout-based Gaussian distribution by utilizing view-changed images generated from the video generation model and view-constraint Gaussian points densification. Additionally, SPC mitigates limited view supervision by employing semantic-prompt-based consistency constraints developed by SAM2. This approach leverages available semantics from training views, serving as instructive prompts, to optimize visually overlapping regions in novel views with 2D and 3D consistency constraints. Extensive experiments demonstrate the superior performance of SPC-GS across Replica and ScanNet benchmarks. Notably, our SPC-GS achieves a 3.06 dB gain in PSNR for reconstruction quality and a 7.3% improvement in mIoU for open-world semantic segmentation. Project website at: https://gbliao.github.io/SPC-GS.github.io.

1. Introduction

3D Gaussian Splatting-based open-world free-view synthesis seeks to construct semantic Gaussian radiance fields using a set of 3D Gaussian primitives enriched with color and semantic attributes. This task enables the free novel view synthesis of both realistic appearances and open-vocabulary semantic understanding (*also referred to as 3D open-vocabulary semantic segmentation*) of the scenes, with significant implications for augmented/virtual reality, robot navigation, and manipulation [16, 18, 26, 28, 30, 58]. Existing methods [35, 50, 59] generally require dense, large sets of input views to train the semantic Gaussian radiance

* Corresponding author

field for comprehensive scene coverage. However, when only sparse views are available, these approaches exhibit substantial declines in reconstruction and segmentation accuracy. This is mainly attributed to two key limitations.

1) Sparsely Distributed Gaussian Points: With sparse input views, the initialized Gaussian points via Structure-from-Motion (SfM) in 3DGS [19] are inherently few and sparsely distributed (Fig. 1 (a1)). Such sparse point distribution imposes challenges in optimizing 3D Gaussians to depict complex indoor environments, leading to suboptimal reconstruction quality and, consequently, impaired segmentation accuracy (Fig. 1 (a2)). **2) Insufficient View Constraints**: Optimizing 3D Gaussian parameters with inadequate view supervision often converges to local minima that overfit training views [25, 60]. This results in semantic ambiguity and poor performance in under-observed regions when rendering from novel viewpoints, due to limited reliable semantic integration across different views.

To address the above challenges, we propose **SPC-GS**, a novel framework that incorporates **Scene-layout-based Gaussian Initialization (SGI)** and **Semantic-Prompt Consistency (SPC)** regularization.

To seek denser points for improved Gaussian initialization, SGI produces a structured scene-layout Gaussian distribution in two steps. First, SGI employs a video generation model to create view-changed images proximal to training images. These generated images, combined with training views, enrich feature matching in SfM, yielding denser SfM points that promote scene-layout Gaussian generation. Second, these denser SfM points undergo view-constraint Gaussian densification and outlier removal to generate structured scene-layout points. These points provide an instructive initialization that faithfully captures indoor scene geometry (*e.g.* furniture, walls, ceilings), enhancing 3DGS training and scene representation, thereby improving segmentation accuracy (Fig. 1 (b1) and (b2)).

Building on dense scene-layout Gaussian initialization, SPC enhances supervision through view-to-view semantic consistency constraints. The core of SPC lies in utilizing informative semantics signals from known training views, serving as *semantic prompts*, to optimize online-rendered novel (pseudo) views, establishing pseudo-multi-view constraints through training-to-pseudo semantic consistency. To achieve it, we utilize the powerful Segment Anything Model 2 (SAM2) [38] to build region mask correspondences across training and pseudo views. Using these correspondences, we leverage the semantics prompts provided by training views to enforce semantic consistency constraints on the corresponding regions of pseudo views at 2D and 3D spaces. This approach substantially augments supervision signals, leading to more coherent segmentation on novel views under sparse input conditions.

Prior works on 3D open-vocabulary segmentation or sparse-input free-view synthesis mainly focus on *object-centric* or *face-forward* scenarios, such as LERF [20], and 3DOVS [31], where objects are centrally positioned, and views are sampled in an "outside-in" manner, resulting in significant view overlap. In contrast, this study focuses on sparse-input open-world free-view synthesis within *indoor scenes*, particularly relevant for robotics and augmented/virtual reality applications. To this end, we utilize the widely recognized Replica [42] and ScanNet [8] datasets, which feature representative indoor scenes characterized by an "inside-out" viewing configuration. This setup leads to less overlap between certain views, posing a substantial challenge for sparse-input open-world free-view synthesis. The contributions of this work are as follows:

- We propose SPC-GS, an early and novel attempt at jointly reconstructing and understanding open-world indoor scenes from sparse-input images.
- We introduce Scene-layout-based Gaussian Initialization (SGI) to tackle the challenge of sparsely distributed Gaussians, offering a dense and instructive scene-layout point initialization that promotes semantic Gaussian learning.
- We present Semantic-Prompt Consistency (SPC) regularization to alleviate inadequate view supervision by enforcing semantic consistency constraints on unseen pseudo views through prompts derived from training views, enhancing coherent segmentation.
- Extensive experiments on sparse-input scenes show the superiority of our approach in free-view synthesis of realistic appearances and open-vocabulary segmentation.

2. Related Work

Free-view Synthesis via Radiance Fields. In recent years, radiance field-based free-view synthesis has made remarkable advancements. A Notable example is the Neural Radiance Field (NeRF) [13, 32], which leverages a coordinate-based multi-layer perceptron (MLP) to implicitly model 3D scenes, enabling novel view synthesis via volume rendering [10]. Another innovative approach is 3D Gaussian Splatting (3DGS) [19, 48], which utilizes a collection of 3D Gaussians to represent the scene and incorporates a fast tile-based rasterizer technique, allowing real-time rendering.

Although current methods achieve impressive results in dense-view scenarios [4, 12, 15, 27, 29, 33, 54], their performance degrades significantly with sparse inputs [17, 60].

Sparse-input Free-view Synthesis. Recent efforts have tried to address the sparse-input view synthesis and they can broadly categorized into cross-scene pre-training methods and per-scene optimization methods.

Pre-training methods learn scene priors by pre-training on large-scale datasets spanning multiple scenes. For instance, pixelNeRF [52] conditions a NeRF on CNN features from input images during volume rendering. Similarly, several approaches [3, 45] utilize different NeRF conditions

to facilitate cross-scene pre-training. Inspired by the success of 3DGS, recent methods such as pixelSplat [2] and MVSplat [6] integrate efficient CUDA-based rasterization with 3D Gaussian representations for pre-training. However, these works demand substantial training resources and large-scale datasets.

Per-scene optimization methods aim to rapidly reconstruct a 3D scene by introducing effective regularizations. Early NeRF-based methods enhance sparse-input view synthesis through depth-based [34, 44] and frequency range regularization [49]. Recent 3DGS-based works focus on improving rendering efficiency. [7] designs depth regularization loss using monocular depth. DNGaussian [25] and FSGS [60] enhance sparse Gaussian representations through global-local depth normalization and correlation depth distribution losses. CoR-GS [56] adopts color co-regularization loss using rendered pseudo views.

Instead, we focus on enhancing view synthesis and open-world 3D indoor semantic understanding from sparse inputs following the per-scene optimization route.

3D Open-world Understanding via Radiance Fields. Recent studies have explored integrating 2D visual foundation models (*e.g*. DINO [1] and CLIP [24, 37]) with radiance fields to facilitate 3D open-world scene understanding. These approaches can be broadly divided into NeRF-based and 3DGS-based methods.

NeRF-based. Early methods [11, 43] distill visual features from DINO into semantic radiance fields. More recent language-embedded neural field methods [20, 22] integrate CLIP's visual features into semantic fields, enabling language-driven 3D semantic understanding. Additionally, 3DOVS [31] further improves semantic accuracy through the proposed Relevancy-Distribution Alignment loss.

3DGS-based. Alternative to NeRFs, recent works explore embedding CLIP features into 3D Gaussians for more efficient semantic rendering. Feature 3DGS [59] directly attaches CLIP embeddings to 3D Gaussians, while FMGS [61] combines Gaussians with multi-resolution hash encodings to form the semantic Gaussian field. [41] embeds quantized CLIP features into Gaussians to reduce storage demand. Furthermore, LangSplat [35] learns low-dimensional semantic Gaussian embeddings and uses a deep neural network for post-process upsampling to obtain semantic representation. Gau-Grouping [50] introduces a 2D object identity loss within training views to optimize Gaussians, while this method requires dense view supervision. Note that the 3D interactive segmentation methods, such as [14, 21, 39, 51], lie outside the scope of language-driven open-world segmentation, as they rely on 2D point prompts to generate object masks but lack semantic meaning.

While previous approaches rely on dense input images, our work focuses on GS-based 3D open-world scene understanding in sparse-input indoor scenarios.

3. Methodology

Fig. 2 illustrates our framework, which consists of two primary components. First, Scene-layout-based Gaussian Initialization (SGI) (Section 3.2) constructs a dense, instructive scene-layout Gaussian initialization that improves Gaussian representation, thus supporting enhanced semantic learning. Then, using the initialized scene-layout Gaussians, Semantic-Prompt Consistency (SPC) Regularization (Section 3.3) introduces 2D and 3D semantic consistency constraints, creating effective pseudo-multi-view optimization for semantic learning under sparse-input conditions.

3.1. Preliminary

Problem Definition. Our approach aims to leverage sparse input images to reconstruct 3D indoor scenes and enable language-driven semantic perception. Concretely, given L posed sparse input images $I = \{I_i\}_{i=1}^{L} \in \mathbb{R}^{L\times H\times W\times 3}$ and language descriptions, our goal is to generate both realistic appearance and open-world semantic understanding results for specified novel viewpoints. To achieve this, we exploit 3D Gaussians [19] for scene representation, and utilize CLIP models [24, 37] to provide semantic supervision.

Preliminary. 3D Gaussian Splatting (3DGS) [19] represents a 3D scene using a collection of explicit 3D Gaussian primitives $\{\mathbf{g}_i = (\mathbf{p}_i, \mathbf{c}_i, o_i, \mathbf{s}_i, \mathbf{q}_i)\}_{i=1}^{N_K}$, where each primitive comprises a 3D position $\mathbf{p} \in \mathbb{R}^3$, color $\mathbf{c} \in \mathbb{R}^3$, opacity $o \in \mathbb{R}$, scaling factor $\mathbf{s} \in \mathbb{R}^3$, and rotation quaternion $\mathbf{q} \in \mathbb{R}^4$. For pixel color $\hat{I}(x_p)$ rendering, 3DGS employs α-blending point-based rendering [23], formulated as:

$$\hat{I}(x_p) = \sum_{i\in\mathcal{N}} T_i \alpha_i \mathbf{c}_i, \tag{1}$$

where $\mathcal{N}$ represents Gaussian points overlapping pixel x_p, $\alpha_i = o_i G_i^{proj}$ with G_i^{proj} as the i-th Gaussian's 2D projection, and transmittance $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$.

For pixel semantic feature $\hat{\mathbf{F}}(x_p)$ rendering, following [35, 59], we attach a learnable semantic parameter $\mathbf{f}_i \in \mathbb{R}^D$ to each Gaussian and use the above α-blending:

$$\hat{\mathbf{F}}(x_p) = \sum_{i\in\mathcal{N}} T_i \alpha_i \mathbf{f}_i. \tag{2}$$

Following [31], we use a CLIP text encoder to generate textual feature $\mathbf{T} \in \mathbb{R}^{M\times 512}$ for M classes and obtain rendered segmentation logits $\hat{\mathbf{S}}$ as follows: $\hat{\mathbf{S}} = \cos\langle \omega_f(\hat{\mathbf{F}}), \mathbf{T}\rangle$, where cos denotes cosine similarity and ω_f is a 1×1 convolution layer for feature dimension alignment.

We incorporate reconstruction and semantic losses to train the 3D Gaussians. The color reconstruction loss, denoted as $\mathcal{L}_C$, follows the formulation in 3DGS [19], and the semantic loss, $\mathcal{L}_S$, is defined as:

$$\mathcal{L}_S = \mathcal{L}_{cos}(\omega_f(\hat{\mathbf{F}}), \mathbf{F}) + \mathcal{L}_{ce}(\omega_s(\hat{\mathbf{S}}), \mathbf{S}), \tag{3}$$

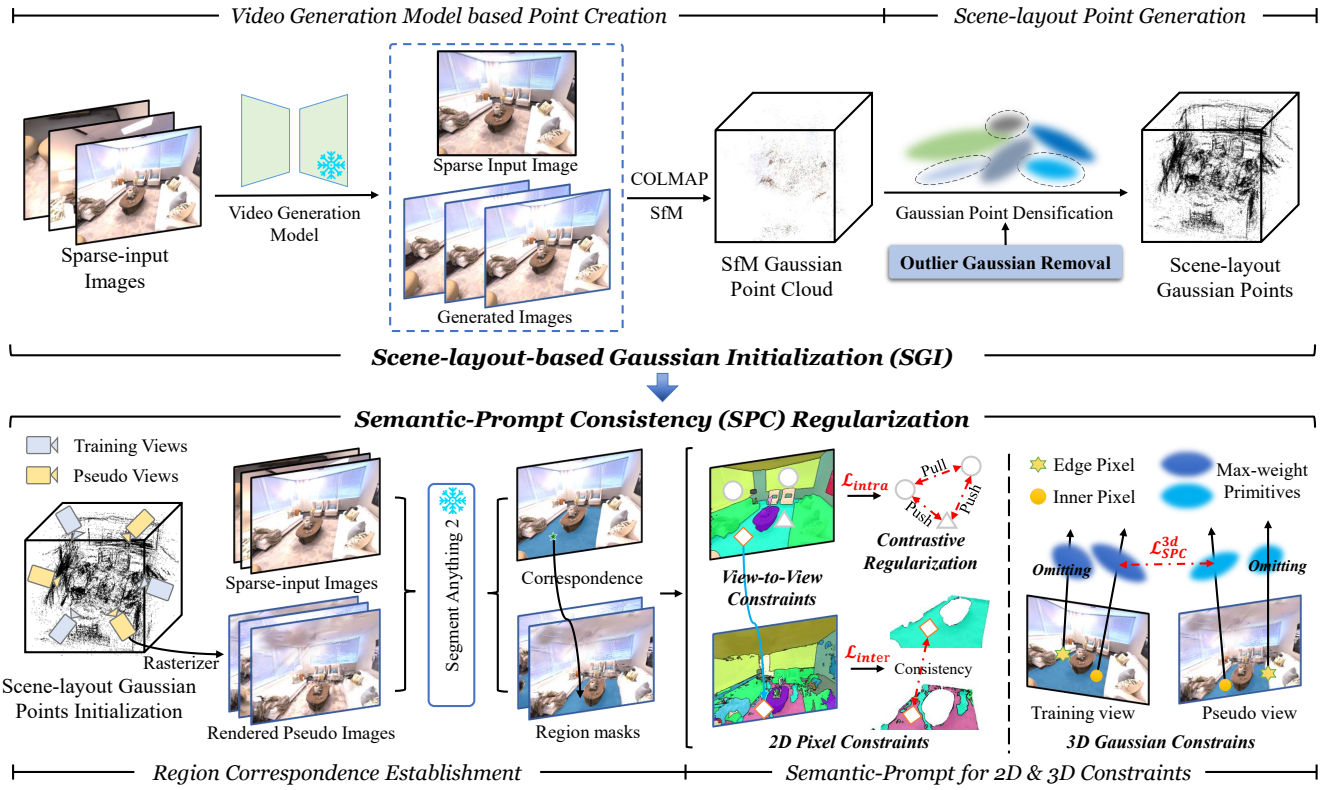

Figure 2. **Framework of SPC-GS.** (a) We first generate adjacent images of each training image using the video generation model [47]. These generated images, combined with the original training images, produce denser initialized SfM points. These points are then optimized to create a scene-layout Gaussian distribution via Gaussian densification and outlier removal. (b) Building on the scene-layout Gaussian initialization, SPC leverages semantic information from training views as instructive semantic prompts to optimize adjacent rendered pseudo views, establishing semantic consistency constraints that enhance overall sparse-input semantic understanding of 3D scenes.

where $\mathbf{F}$ and $\mathbf{S}$ are CLIP-derived semantic features and relevancy maps, respectively. ω_s is a 1×1 convolution layer, and $\mathcal{L}_{cos}$ and $\mathcal{L}_{ce}$ denote cosine and cross-entropy losses.

3.2. Scene-layout-based Gaussian Initialization

We propose Scene-layout-based Gaussian Initialization (SGI) to produce a dense and conducive point distribution for initialization with the following insights and designs.

VGM-based Point Creation. In static indoor scenes, viewpoint variations correspond to camera movements. Motivated by this, we employ a video generation model (VGM) to create additional views beyond the sparse training set through image-to-video generation. These generated views augment the available scene information and, when combined with the original training views, enrich feature matching during SfM, yielding denser SfM points for generating scene-layout Gaussians.

Specifically, for each training image I_i, we generate eight adjacent views $\{\tilde{I}_i^j\}_{j=1}^{8}$ using the image-to-video generation mode of the powerful video generation model MotionCtrl [47]. This is accomplished by inputting the training images along with eight standard camera motion trajectories [47] into MotionCtrl. These trajectories include *pan up, pan down, pan left, pan right, zoom in, zoom out, clockwise rotation*, and *anticlockwise rotation*. Each trajectory generates a sequence of consecutive video frames. From each sequence, we select a single frame temporally adjacent to the input training image as the valid generated image since it exhibits superior visual quality. These generated images are then combined with the original training images for SfM processing, yielding denser SfM points for Gaussian initialization (*see supplementary file*). Despite being denser than SfM points derived from sparse-input training images, these points remain limited in capturing the geometric layout and details of indoor scenes.

Scene-layout Point Generation. We further generate scene-layout point distributions to enhance 3DGS training and scene representation through the following steps.

i) View-constraints 3D Gaussian Densification. Using the denser SfM point initialization, we follow Gaussian densification [19] and using our view constraints from Eq. (9) to generate a more densely distributed, instructive scene-layout Gaussian point set $\{\mathbf{g}_i = (\mathbf{p}_i, \mathbf{c}_i, \mathbf{f}_i, o_i, \mathbf{s}_i, \mathbf{q}_i)\}_{i=1}^{N_{K'}}$, as shown in Fig. 2. These denser $N_{K'}$ points provide an

improved spatial layout of indoor scenes, and enhanced attribute representations (*e.g.* $\mathbf{c}_i$, $\mathbf{f}_i$), making them effective priors for Gaussian initialization.

ii) Outlier Gaussian Primitive Removal (OGR). During Gaussian densification, new Gaussian primitives are sampled based on a 3D Gaussian probability distribution, which may generate outlier primitives positioned far from the actual scene layout. These outliers, when used in subsequent initialization, complicate the optimization process. To mitigate this, we present a density-based removal strategy to remove those sparse outlier points. Specifically, Gaussian primitives with fewer than $\mathcal{P}$ neighboring points within a spherical radius r are periodically eliminated during training. Empirical experiments show that setting $\mathcal{P}$ to 5 and r to 1 strikes a balance between effective outlier removal and the preservation of essential scene elements.

By using instructive scene-layout-based points for initialization, the 3DGS training and scene representation can be enhanced, consequently improving segmentation performance under the sparse-input condition.

3.3. Semantic-Prompt Consistency Regularization

We introduce Semantic-Prompt Consistency (SPC) regularization to exploit additional supervision signals. The key idea of SPC is to leverage semantics from training views as prompts, to impose effective semantic consistency constraints on corresponding regions of pseudo views.

Region Correspondence Establishment. To establish correspondences between training and pseudo views, we draw inspiration from the powerful zero-shot temporal region proposal extraction capability of Segment Anything Model 2 (SAM2) [38] and leverage it to achieve this.

Specifically, we introduce Iterative Stochastic Prompting (ISP) to efficiently build correspondence. During training, ISP renders pseudo (unobserved) views using the online training SPC-GS. The view locations are sampled from the two nearest known training views by adding random noise following [60]. Then, ISP treats the current training image and its associated pseudo-views as a viewpoint-changed video sequence, and selects a stochastic point coordinate in the training image, as the 2D point prompt for SAM2, generating corresponding region masks across both the training and pseudo views, which can be formulated as:

$$\mathbf{M}_i, \{\mathbf{M}_i^*\}_{p=1}^{P} = \mathbf{ISP}(I_i, \{I_i^*\}_{p=1}^{P}), \tag{4}$$

where I_i^* is the rendered pseudo image and $\mathbf{M}_i^*$ is the corresponding region mask within I_i^*. P is set to 2 for computational efficiency. As training progress proceeds, the stochastic sampling mechanism encourages a uniform distribution of points across the entire image space, facilitating the efficient region mask correspondence construction for diverse regions of the whole image, as shown in Fig. 3.

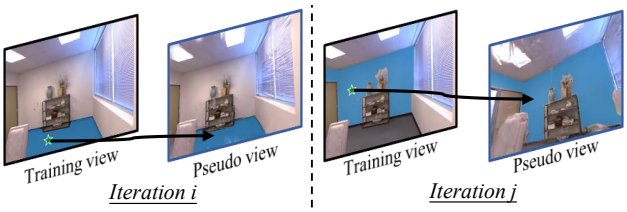

Figure 3. Illustration of Iterative Stochastic Prompting (ISP). During training, ISP randomly samples point coordinates (★) from the training view, serving as the point prompt for SAM2 to produce region masks (highlighted in blue) across training and pseudo views.

Semantic-Prompt for 2D Constrains. Building on the established region mask correspondences, we enforce effective constraints within 2D views to enhance Gaussian learning with sparse inputs.

i) View-to-View Regularization. Leveraging the semantics from training views as prompts, we enforce training-to-pseudo semantic consistency constraints in feature and relevancy map aspects, which is formulated as:

$$\mathcal{L}_{inter} = \sum_{p=1}^{P} || \frac{\hat{\mathbf{F}}_p^* \odot \mathbf{M}_p^*}{\sum \mathbf{M}_p^*} - \frac{\hat{\mathbf{F}} \odot \mathbf{M}}{\sum \mathbf{M}} ||_2 - \sum_{p=1}^{P} \mathcal{R}[\Phi(\hat{\mathbf{S}} \odot \mathbf{M})] \log(\hat{\mathbf{S}}_p^* \odot \mathbf{M}_p^*), \tag{5}$$

where $\hat{\mathbf{F}}^*$ and $\hat{\mathbf{S}}^*$ represent the semantic feature and the relevancy map derived from pseudo views. The function Φ is the semantic uniformity operation: $\Phi(\hat{\mathbf{S}} \odot \mathbf{M}) = \mathbf{argmax}_c \sum_{i \in \mathbf{M}} \mathbb{1}\{\mathbf{argmax}(\omega_s(\hat{\mathbf{S}}_i)) = c\}$, which integrates SAM2's masks with a maximum voting mechanism to select the dominant class c within the region mask $\mathbf{M}$, unifying the region's semantics and enforcing consistent region-wise semantic supervision. $\mathcal{R}[\cdot]$ denotes the dimensional reshaping operation to align the dimension shape of the region mask $\mathbf{M}^*$ from the pseudo view. In $\mathcal{L}_{inter}$, we utilize self-rendered semantic $(\hat{\mathbf{F}}, \hat{\mathbf{S}})$ from training views, generated by the trained SPC-GS, rather than 2D CLIP semantics. This is motivated by the superior coherence and reliability of 3D representations over their 2D counterparts [55] (*see supplementary file for experimental validation*).

ii) Contrastive Regularization. Beyond inter-view, we introduce semantic-guided contrastive learning regularization within intra-training-views to further enhance supervision. This is based on the principle that features of semantically consistent objects should be similar.

Given N feature samples, each sample $\mathbf{z}$ is derived from $\mathbf{z} = W_\psi(\hat{\mathbf{F}})$, where W_ψ is a 1×1 convolution transformation layer. Let $i \in U_z \equiv \{1...N\}$ be the index of a sample, and $T_z(i) \equiv U_z \backslash \{i\}$ represents the index set excluding index i. We designate features with the same rendered semantic label $\tilde{\mathbf{q}}$ as positive samples, where $\tilde{\mathbf{q}} =$

$\Phi(\hat{\mathbf{S}} \odot \{\mathbf{M}^k\}_{k=1}^{\eta})$. Here, the $\{\mathbf{M}^k\}_{k=1}^{\eta}$ denote η masks generated by SAM2's automatic mask generator in the entire image, and Φ is the function that unifies semantics within each mask. Accordingly, for the i-th sample, positive sample indexes are defined as: $P_z(i) \equiv \{p \in T_z(i) : \tilde{\mathbf{q}}_p = \tilde{\mathbf{q}}_i\}$, and negative sample indexes: $T_z(i) \backslash P_z(i)$. Similar to SimCLR [5], we employ contrastive learning loss to maximize agreement between positive samples. This method brings P_z-indexed positive samples closer in feature space, imposing effective contrastive constraints on sparse-view Gaussian learning, which can be formulated as:

$$\mathcal{L}_{intra} = -\sum_{i \in U_z} \frac{1}{\mathcal{M}_z(i)} \sum_{p \in P_z(i)} \log \left[\frac{e^{((\mathbf{z}_i \cdot \mathbf{z}_p)/\tau)}}{\sum_{j \in T_z(i)} e^{((\mathbf{z}_i \cdot \mathbf{z}_j)/\tau)}} \right], \tag{6}$$

where $\mathcal{M}_z(i)$ denotes the number of elements in $P_z(i)$. The symbol $\cdot$ refers to the inner product operation. τ is the temperature parameter, set to 0.2 in the experiment.

To obtain N samples, we employ a patch-wise sampling strategy by randomly extracting a patch of size $h \times w$ in each training iteration, where $N = h \times w$. Empirically, h and w are both set to 128. In summary, the semantic prompt for 2D constraints can be formulated as:

$$\mathcal{L}_{SPC}^{2d} = \mathcal{L}_{inter} + \mathcal{L}_{intra}. \tag{7}$$

Semantic-Prompt for 3D Constrains. In addition to 2D constraints, we extend semantic-prompt to impose effective 3D constraints on 3D Gaussian primitives. Considering that Gaussian primitives with maximum weight dominate the rendering results (Eq. (2)), we encourage *semantic feature similarity* between the maximally-weighted 3D Gaussians from two corresponding regions. Specifically, for each pixel within the region mask, we identify the Gaussian primitive with maximum weight along the ray cast into 3D space. Then, the semantic features of these maximally-weighted Gaussians from training views are used to constrain their counterparts in pseudo views.

However, since Gaussians related to region boundaries often span objects with differing semantics, directly applying the above constraints may complicate optimization. To mitigate this, we introduce a *Region Boundary Erosion* technique, which removes ambiguous primitives corresponding to region boundaries, which is defined as:

$$\mathcal{L}_{SPC}^{3d} = \mathcal{L}_{kl}(\mathcal{F} || \mathcal{F}^*) = \sum_{j=1}^{n'} \bar{\mathbf{f}} \log(\frac{\bar{\mathbf{f}}}{\mathbf{f}_j^*}), \tag{8}$$

where $\mathcal{F}$ denotes the mean feature representation of 3D Gaussian primitives derived from mask $\mathbf{M}'$ in the training view, defined as $\bar{\mathbf{f}} = (\mathbf{f} \odot \mathbf{M}') / \sum \mathbf{M}'$. Here, $\mathbf{M}'$ refers to the eroded mask after region boundary erosion, which is achieved through morphological erosion. Specifically, a pixel $\mathbf{M}'(x, y)$ is retained as "True" only if all pixels within its 5×5 neighborhood in the input original mask $\mathbf{M}$ are "True"; otherwise, it is set to "False". This operation effectively shrinks the boundaries of the region, causing them to contract inward. The set $\mathcal{F}^* = \{\mathbf{f}_1^*, \mathbf{f}_2^*, ..., \mathbf{f}_{n'}^*\}$ comprises n' Gaussian primitives obtained from eroded masks $\{\mathbf{M}^{*\prime}\}_{p=1}^{P}$ in pseudo views.

To summarize, the overall loss is given by:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_S + \mathcal{L}_{GS} + \mathcal{L}_S^{3d} + \lambda(\mathcal{L}_{SPC}^{2d} + \mathcal{L}_{SPC}^{3d}), \tag{9}$$

where $\mathcal{L}_{GS}$ is the semantic loss for generated images from video generation model: $\mathcal{L}_{GS} = \mathcal{L}_{cos}(\omega_f(\hat{\mathbf{F}}^g), \mathbf{F}^g) + \mathcal{L}_{ce}(\omega_s(\hat{\mathbf{S}}^g), \mathbf{S}^g)$. These generated RGB images are not used for color constraints due to their error-prone color pixels that degrade overall results (*see supplementary file*). $\mathcal{L}_S^{3d}$ is the 3D local adaptive regularization, which promotes semantic similarity among neighboring Gaussians: $\mathcal{L}_S^{3d} = \frac{1}{\theta k} \sum_{i=1}^{\theta} \sum_{j=1}^{k} \omega_{ij}(||\mathbf{f}_i - \mathbf{f}_j||_2)$, where $\theta = 800$ denotes the number of Gaussian randomly sampled per iteration, and $k = 5$ specifies the number of nearest neighbors for each sampled Gaussian in 3D space. $\omega_{ij} = \exp(-||\mu_i - \mu_j||_2)$ is the distance-based weighting coefficient. To prevent early-stage suboptimal convergence, λ is initialized at 0 and then increased to 1 after 4k iterations during training.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on two widely-used indoor scene datasets in terms of novel view synthesis: Replica [42] and ScanNet [8], using a sparse-input setting with only 12 input images for training. Following the standard sparse-input protocol [25, 60], every 10-th frame from each sequence is selected for testing, while 12 evenly sampled sparse views from the remaining images are used for training. These 12 views collectively encompass the entire scene. Notably, the *"inside-out"* viewing direction and limited overlap among views in sparse-input indoor scenes pose significant challenges for free-view synthesis.

Evaluation Metrics. We employ PSNR, SSIM [46], and LPIPS [57] for evaluating image quality. We report segmentation performance using the mean Intersection-over-Union (mIoU) and mean pixel accuracy (mAcc).

Implementation Details. We implement our approach using the official PyTorch 3D Gaussian Splatting [19] on an NVIDIA A100 GPU. The dimension of the learnable semantic parameter attached to 3D Gaussians is set to 32. We modify the CUDA kernel to enable parallel differentiable rasterization of both color and semantic attributes. We employ CLIP ViT-B/16 [37] for feature extraction and SAM2 Hiera-L [38] for region mask generation. More details are provided in the supplementary file.

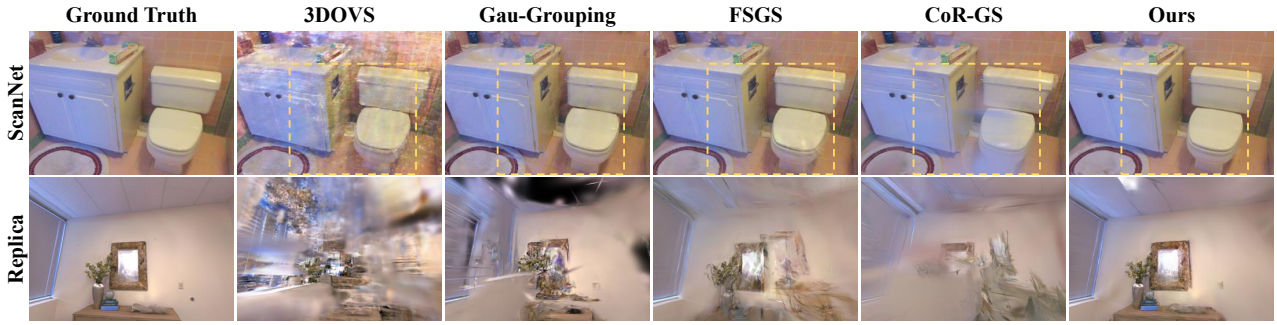

Figure 4. Visual reconstruction results on novel views. Our approach achieves superior global structure and photo-realistic details.

Method	Replica [42]					ScanNet [8]				
	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	mAcc ↑	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	mAcc ↑
3DOVS [31]	13.200	0.515	0.663	3.913	21.439	18.721	0.643	0.544	7.751	36.356
Feature 3DGS [59]	17.059	0.700	0.359	7.407	30.209	18.088	0.656	0.424	7.347	38.002
LangSplat [35]	17.715	0.701	0.420	3.292	18.621	18.780	0.663	0.476	3.339	24.491
Gau-Grouping [50]	17.297	0.691	0.376	15.028	50.422	19.006	0.675	0.421	15.172	57.355
DNGaussian [25]	18.513	0.712	0.402	16.657	54.400	18.392	0.691	0.484	13.912	54.814
FSGS [60]	18.694	0.725	0.328	16.101	53.474	20.574	0.696	0.406	14.031	56.850
CoR-GS [56]	18.955	0.758	0.312	15.933	52.476	20.815	0.721	0.393	14.462	55.812
Ours	**22.011**	**0.792**	**0.254**	**23.960**	**63.262**	**22.401**	**0.741**	**0.368**	**28.944**	**68.823**

Table 1. Quantitative results of reconstruction and segmentation on novel views on the Replica and ScanNet datasets, using the CLIP [37] for optimization with 12 training views. Our approach achieves **superior performances** across all metrics on various datasets.

4.2. Comparison with State-of-the-art Methods

We conduct novel view synthesis comparisons using sparse input data in indoor scenes, including two types of methods. *i) 3D open-vocabulary segmentation methods*, including 3DOVS [31], Feature 3DGS [59], LangSplat [35], and Gau-Grouping [50], are designed for 3D open world semantic understanding. *ii) Sparse-input free-view synthesis methods* contain DNGaussian [25], FSGS [60], and CoR-GS [56], which are designed to learn Gaussian radiance fields from sparse inputs, focusing primarily on appearance modeling. For concurrent 3D open-world segmentation, we extend these methods by adding semantic properties to Gaussian and incorporating CLIP-based semantic supervision, aligning with the semantic settings of our approach. We adopt consistent evaluations across all methods by employing the same sparse-input training protocol and test views.

Qualitative Results. *i) Reconstruction.* As shown in Fig. 4, the results from existing 3D open-vocabulary segmentation methods exhibit significant artifacts, since these methods lack sufficient view constraints. Although sparse-view reconstruction methods introduce geometric constraints for regularization, they struggle to recover photo-realistic details. In comparison, our approach incorporates scene-layout points prior and semantic consistency constraints, leading to a more detailed appearance.

ii) Open-world Segmentation. In Fig. 5, current methods struggle to obtain precise and complete results, since they easily encounter ambiguity arising from insufficient semantic view supervision. In contrast, our method delivers accurate results, benefiting from the SPC regularization that provides effective consistency constraints.

Quantitative Results. Table 1 shows that, when optimized using the CLIP model [37], our approach significantly outperforms existing methods across all metrics on both synthetic and real datasets. Competing methods struggle under sparse input conditions, due to either limited Gaussian representations or insufficient semantic consistency constraints. Notably, LangSplat faces considerable challenges, as it freezes Gaussian coordinates, scale, and rotation during semantic parameter optimization, which severely impedes semantic learning under sparse inputs. Besides, our method maintains superior results when optimized with CLIP-LSeg [24] (*see supplementary file*), validating its generalizability.

4.3. Ablation Studies

Key Components. In Table 2, applying the VGM-based Point Creation (PC, #1) and Scene-layout Point Generation (SPG, #2) notably improves reconstruction quality over setting #0, and consequently boosts segmentation results, validating SGI's effectiveness in enhancing Gaussian representation. Adding the full Semantic-Prompt Consistency regularization (SPC, #4) further yields significant segmentation gains, showing SPC's capability to address insufficient view constraints and enhance 3D scene understanding. Fig. 6 also shows that adding SGI can reduce artifacts, and applying SPC further enhances segmentation results.

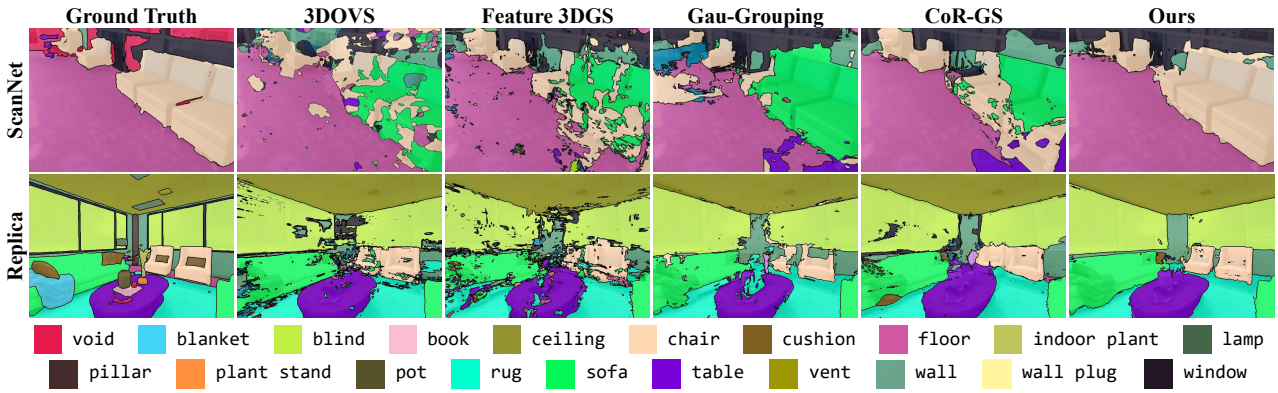

Figure 5. Visual open-world segmentation results on novel views. Our approach yields more accurate and complete results.

Case	SGI		SPC		Replica [42]				
	PC	SPG	2D	3D	PSNR	SSIM	LPIPS	mIoU	mAcc
#0					17.134	0.678	0.369	14.899	51.255
#1	✓				18.924	0.724	0.334	16.279	52.212
#2	✓	✓			21.176	0.783	0.260	19.401	58.281
#3	✓	✓	✓		21.814	0.791	0.257	22.849	61.362
#4	✓	✓	✓	✓	**22.011**	**0.792**	**0.254**	**23.960**	**63.262**

Table 2. Ablation study on key components.

Case	Configuration	PSNR	SSIM	LPIPS	mIoU	mAcc
#a	Ours (Full setting)	22.011	0.792	0.254	23.960	63.262
#b	w/o View-to-View Regular.	21.683	0.789	0.257	21.244	60.914
#c	w/o Contrastive Regular.	21.879	0.792	0.256	22.322	60.086
#d	w/o Boundary Erosion	21.950	0.792	0.254	22.964	61.706

Table 3. Ablation results of our SPC with different settings.

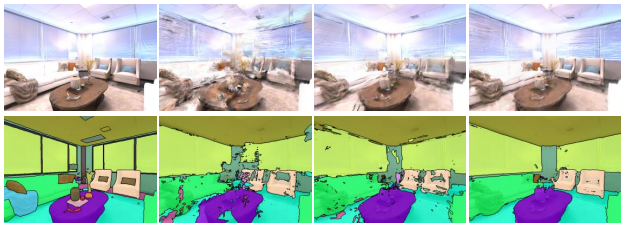

Figure 6. Visual ablation comparison of key components.

Components in SPC. In Table 3, removing either the 2D view-to-view or contrastive regularization (#b, #c), or omitting the region boundary erosion operation in the 3D constraint (#d), degrades performance. This validates their roles in mitigating insufficient supervision and filtering ambiguous Gaussian primitives spanning different objects.

Robustness. In Table 4, our strategies also achieve significant gains even with sparse SfM initialization from limited training images, validating their robustness.

Initialization. From top to bottom in Fig. 7, we can see that leveraging denser and more effective points for Gaussian initialization promotes 3DGS training that enhances scene representation, leading to improved rendered results.

Case	Configuration	PSNR	SSIM	LPIPS	mIoU	mAcc
#a	SfM Points Initialization	17.134	0.678	0.369	14.899	51.255
#b	#a + SGI	20.680	0.774	0.270	18.169	56.928
#c	#b + SPC	21.589	0.786	0.260	20.766	61.000

Table 4. Results of our strategies using SfM point initialization.

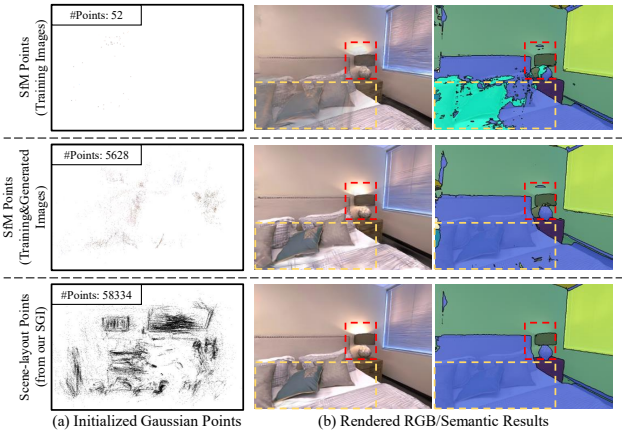

Figure 7. Visual comparison of different initializations and their rendering results after optimization with these initializations.

5. Conclusion

This paper presents SPC-GS, a novel approach to address the challenges of GS-based indoor open-world free-view synthesis with sparse views. In SPC-GS, the proposed Scene-layout-based Gaussian Initialization (SGI) effectively enhances Gaussian representation by providing an instructive point initialization, consequently promoting semantic learning. Additionally, the Semantic-Prompt Consistency (SPC) regularization mitigates limited supervision by introducing additional view-consistent semantic constraints, improving coherent segmentation performance. Experimental results show that our approach significantly outperforms SOTA methods in reconstruction and 3D open-world segmentation tasks across synthetic and real-world datasets with various CLIP models.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 3

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 2

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 6

[6] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *Proceedings of the European Conference on Computer Vision*, pages 370–386. Springer, 2024. 3

[7] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024. 3

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 6, 7, 1, 3, 4

[9] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics*, 36(4): 1, 2017. 2

[10] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. *ACM Siggraph Computer Graphics*, 22(4): 65–74, 1988. 2

[11] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[12] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[13] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 2

[14] Xu Hu, Yuxi Wang, Lue Fan, Junsong Fan, Junran Peng, Zhen Lei, Qing Li, and Zhaoxiang Zhang. Semantic anything in 3d gaussians. *arXiv preprint arXiv:2401.17857*, 2024. 3

[15] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[16] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 10608–10615, 2023. 1

[17] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[18] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua Tenenbaum, Celso de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. In *Proceedings of Robotics: Science and Systems*, 2023. 1

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3, 4, 6

[20] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2, 3

[21] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21530–21539, 2024. 3

[22] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 23311–23330, 2022. 3

[23] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. In *Computer Graphics Forum*, pages 29–43. Wiley Online Library, 2021. 3

[24] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *Proceedings of the International Conference on Learning Representations*, 2022. 3, 7, 6, 8, 10

[25] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. 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*, pages 20775–20785, 2024. 2, 3, 6, 7, 9, 10

[26] Guibiao Liao, Wei Gao, Qiuping Jiang, Ronggang Wang, and Ge Li. Mmnet: Multi-stage and multi-scale fusion network for rgb-d salient object detection. In *Proceedings of the ACM international conference on multimedia*, pages 2436–2444, 2020. 1

[27] Guibiao Liao, Jiankun Li, Zhenyu Bao, Xiaoqing Ye, Jingdong Wang, Qing Li, and Kanglin Liu. Clip-gs: Clip-informed gaussian splatting for real-time and view-consistent 3d semantic understanding. *arXiv preprint arXiv:2404.14249*, 2024. 2

[28] Guibiao Liao, Jiankun Li, and Xiaoqing Ye. Vlm2scene: Self-supervised image-text-lidar learning with foundation models for autonomous driving scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3351–3359, 2024. 1

[29] Guibiao Liao, Kaichen Zhou, Zhenyu Bao, Kanglin Liu, and Qing Li. Ov-nerf: Open-vocabulary neural radiance fields with vision and language foundation models for 3d semantic understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[30] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. *arXiv preprint arXiv:2403.03174*, 2024. 1

[31] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 53433–53456, 2023. 2, 3, 7, 9, 10

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2

[33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41 (4):1–15, 2022. 2

[34] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 3

[35] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 1, 3, 7, 2, 9

[36] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5328–5337, 2024. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6, 7, 8, 9

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 5, 6

[39] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. In *Proceedings of the European Conference on Computer Vision*, pages 456–472. Springer, 2024. 3

[40] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203, 2024. 3, 4

[41] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024. 3

[42] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 6, 7, 8, 1, 3, 4

[43] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *Proceedings of the International Conference on 3D Vision*, pages 443–453, 2022. 3

[44] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 3

[45] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 2

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[47] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *Proceedings of the ACM SIGGRAPH*, pages 1–11, 2024. 4, 1, 2

[48] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d

gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024. 2

[49] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 3

[50] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *Proceedings of the European Conference on Computer Vision*, pages 162–179. Springer, 2024. 1, 3, 7, 9, 10

[51] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20612–20622, 2024. 3

[52] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2

[53] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3, 4

[54] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 2

[55] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *Proceedings of the European Conference on Computer Vision*, pages 57–74. Springer, 2024. 5

[56] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. Cor-gs: sparse-view 3d gaussian splatting via co-regularization. In *Proceedings of the European Conference on Computer Vision*, pages 335–352. Springer, 2024. 3, 7, 9, 10

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[58] Xiaoyu Zhang, Guibiao Liao, Wei Gao, and Ge Li. Tdrnet: Transformer-based dual-branch restoration network for geometry based point cloud compression artifacts. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2022. 1

[59] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta 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*, pages 21676–21685, 2024. 1, 3, 7, 2, 9, 10

[60] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *Proceedings of the European Conference on Computer Vision*, pages 145–163. Springer, 2024. 2, 3, 5, 6, 7, 9, 10

[61] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *arXiv preprint arXiv:2401.01970*, 2024. 3