SpatialSplat: Efficient Semantic 3D from Sparse Unposed Images

Yu Sheng[1], Jiajun Deng[2†], Xinran Zhang[1], Yu Zhang[1], Bei Hua[1], Yanyong Zhang[1], Jianmin Ji[1†]
[1]University of Science and Technology of China, [2]The University of Adelaide
{shengyu724,zxrr}@mail.ustc.edu.cn, jiajun.deng@adelaide.edu.au
{zhangyu,bhua,yanyongz,jianmin}@ustc.edu.cn †: Corresponding Author

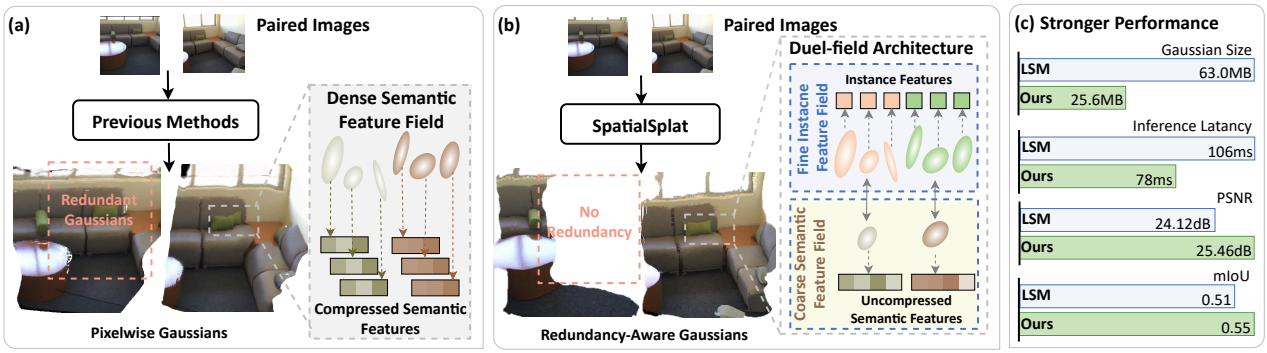

Figure 1. **Comparison between previous methods and our SpatialSplat.** (a): Previous methods predict pixel-wise Gaussians, associating each primitive with compressed semantic feature. (b): Our SpatialSplat avoids generating redundant Gaussian primitives for overlapping pixels, and it represents semantics with a dual-field architecture to better preserve the information. (c): SpatialSplat outperforms the state-of-the-art method LSM [15] for both novel-view rendering quality and semantic segmentation while being more efficient.

Abstract

*A major breakthrough in 3D reconstruction is the feedforward paradigm to generate pixel-wise 3D points or Gaussian primitives from sparse, unposed images. To further incorporate semantics while avoiding the significant memory and storage costs of high-dimensional semantic features, existing methods extend this paradigm by associating each primitive with a compressed semantic feature vector. However, these methods have two major limitations: (a) the naively compressed feature compromises expressiveness, affecting the model's ability to capture fine-grained semantics, and (b) the pixel-wise primitive prediction introduces redundancy in overlapping areas, causing unnecessary memory overhead. To this end, we introduce **SpatialSplat**, a feedforward framework that produces redundancy-aware Gaussians and capitalizes on a dual-field semantic representation. Particularly, with the insight that primitives within the same instance exhibit high semantic consistency, we decompose the semantic representation into a coarse feature field that encodes uncompressed semantics with minimal primitives, and a fine-grained yet low-dimensional feature field that captures detailed instance relationships. Moreover, we propose a selective Gaussian mechanism, which retains only essential Gaussians in the scene, effectively eliminating redundant primitives. Our proposed Spatialsplat learns accurate semantic information and detailed instances prior with more compact 3D Gaussians, making semantic 3D reconstruction more applicable. We conduct extensive experiments to evaluate our method, demonstrating a remarkable 60% reduction in scene representation parameters while achieving superior performance over state-of-the-art methods. The code is available at **SpatialSplat**.*

1. Introduction

Reconstructing and understanding 3D scenes from 2D images [5, 9, 32, 36, 53] is a fundamental topic in computer vision, aiming to obtain semantic-aware 3D structure from low-cost devices, *i.e.*, RGB cameras. This technique is significant for various applications, such as robotics, autonomous driving, and VR/AR. Recently, with the emergence of the new 3D representation, *i.e.*, Neural Radiance Fields (NeRF) [29] and 3D Gaussian Splatting (3DGS) [19], a popular paradigm of performing semantic-aware 3D re-

construction is to distill the feature field of NeRF or 3DGS with powerful 2D vision-language foundation models [20, 31, 33, 52]. However, these methods typically rely on per-scene optimization and complex multi-step preprocessing, limiting their ability to generalize across multiple scenes within a single model.

Recent breakthroughs [3, 7, 49] in 3D reconstruction have incorporated feed-forward networks to improve the generalization of 3DGS models and accelerate reconstruction. These methods follow a paradigm that generates pixel-wise Gaussian primitives from sparse posed images. Building on this, further efforts [38, 46] have extended the paradigm to reconstruct scenes from unposed images. Moreover, methods [15, 43] have advanced this approach by integrating semantic understanding, enabling semantic 3D reconstruction from sparse unposed images.

Despite significant progress, these methods have two major limitations. First, as shown in Fig.1 (a), pixel-wise Gaussian prediction introduces substantial redundancy in overlapping regions while providing minimal accuracy gains. Method [44] attempts to mitigate this issue by projecting primitives onto a plane and identifying overlaps, it depends on precise camera extrinsics, which are often difficult to obtain in real-world scenarios—especially when only a few sparse, textureless images are available. Second, methods [15, 31, 33, 43, 52] enforce per-primitive semantic learning to form a dense semantic feature field, as shown in Fig.1 (a). This primitive-level approach incurs significant memory and storage costs due to high-dimensional (512 or more) language features. To mitigate this, the methods often compress features into lower-dimensional representations (e.g., 64-128), but this compression leads to irreversible information loss [52], leading to suboptimal scene understanding performance.

In this paper, we take a slightly different viewpoint. First, inspired by MASt3R [25], which successfully extracts 3D geometry and pixel correspondences from images, we observe that redundant primitives often share similar geometry and appearance, allowing them to be identified directly from images without relying on complex geometric priors. Second, we find that per-primitive semantics are not essential for high-performance scene understanding. Instead, a coarse semantic representation, when combined with fine-grained instance information, is sufficient for strong performance, as Gaussian primitives within the same instance exhibit high semantic consistency.

By consolidating this idea, we introduce SpatialSplat, a feed-forward network generating compact yet expressive semantic 3D representations from unposed images. As depicted in Fig. 1 (b), SpatialSplat introduces a dual-field architecture that decomposes the dense semantic feature field into two components: a coarse feature field encoding uncompressed instance-level semantics through minimal Gaussians, and a low-dimensional fine-grained feature field capturing inter-instance relationships. Guided by multiple 2D foundation models, our method learns both accurate semantic and instance priors, achieving superior performance while reducing representation parameters. Additionally, we introduce a Selective Gaussian Mechanism (SGM) to eliminate redundancy in overlapping areas caused by pixelwise representations, along with a novel loss function that jointly optimizes redundancy-aware Gaussians and scene fidelity. Extensive experiments demonstrate that our method achieves state-of-the-art performance on multiple downstream 3D tasks while using only 40% of the representation parameters required by the baseline, all without any 3D supervision. Our contributions are threefold:

- A novel feed-forward 3DGS framework that, to the best of our knowledge, is the first to simultaneously learn semantic and instance priors with guidance from foundation models while maintaining a high-fidelity radiance field.
- A dual-field semantic representation that significantly reduces storage consumption while enhancing open-vocabulary 3D segmentation performance.
- A mechanism that identifies redundant primitives from images without relying on any geometric priors.

2. Related Work

2.1. Feed-forward 3D Reconstruction

The emergence of NeRF [29] and 3DGS [19] has significantly transformed the paradigm of 3D reconstruction from images. Despite achieving remarkable results in 3D reconstruction and novel view synthesis, these methods rely on time-consuming per-scene optimization, with even improved techniques [4, 16, 30] still requiring several minutes to hours. To overcome these limitations, recent approaches [3, 7, 11, 18, 41, 45, 48, 49] have explored using a single feed-forward network to learn reconstruction priors from large datasets, enabling generalizable 3D reconstruction. DUSt3R serials [25, 42] further simplifies the reconstruction pipeline, making 3D reconstruction possible from images without relying on camera extrinsics. This advancement has inspired a new wave of methods [13, 38, 46] capable of achieving dense reconstruction and view synthesis without camera extrinsics.

2.2. Feature Field Distillation

Early approaches of feature field distillation were largely based on NeRF [29]. For example, methods [37, 51] successfully embedded 2D label data into 3D space, achieving sharp and precise 3D segmentation. Later approaches [12, 20, 22, 36, 40] incorporated more complex pixel-aligned feature vectors from technologies like LSeg [26] or DINO [2] into NeRF frameworks, enabling open-vocabulary 3D segmentation. With the rise of

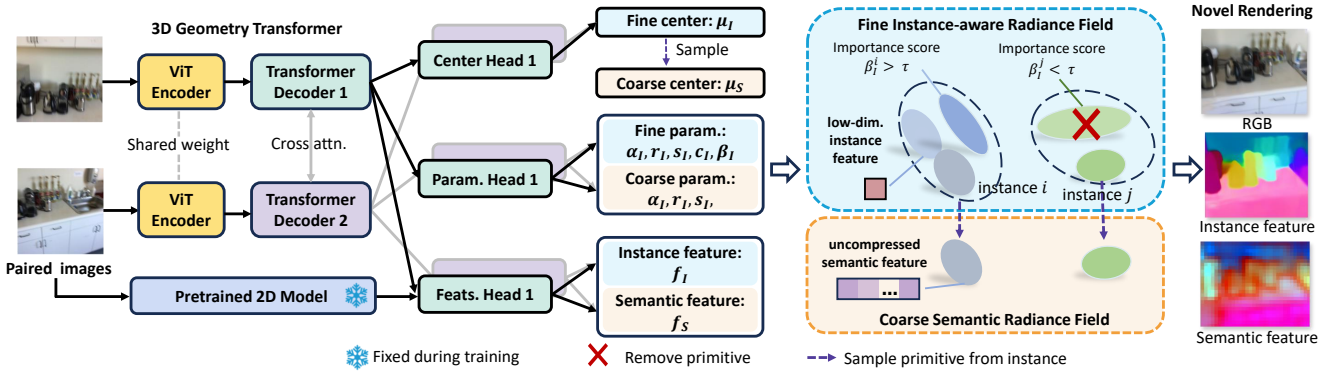

Figure 2. **Pipeline of SpatialSplat.** The SpatialSplat processes unposed images along with their intrinsics through a 3D geometry transformer. The extracted features from the geometry transformer and the pretrained 2D model are then fed into separate Dense Prediction Transformer (DPT) heads to predict different Gaussian attributes, resulting in a fine-grained instance-aware radiance field $\boldsymbol{\mathcal{F}_I}$ and a coarse semantic feature field $\boldsymbol{\mathcal{F}_S}$. This enables the synthesis of RGB and feature maps from novel viewpoints.

3DGS [19], methods such as [17, 31, 33, 52] have extended this concept by embedding language features into Gaussian attributes to enhance semantic understanding. Large Spatial Model (LSM) [15] was the first to achieve a generalizable semantic radiance field based on feed-forward 3DGS. However, due to the explicit nature of 3DGS representations, attaching feature vectors to each primitive incurs high memory costs. To mitigate this, existing methods often compress semantic features into lower-dimensional representations, inevitably resulting in information loss.

In contrast, our approach decomposes semantics into coarse instance-level semantics and fine-grained instance information, achieving superior performance with significantly lower storage requirements.

2.3. Compact 3D Gaussian

Our work is also related to compact 3D Gaussian representation. Recent works have attempted to compress 3DGS by exploring the spatial relationships among Gaussians. Scaffold-GS [27] introduces anchor-centered features to reduce parameter counts, while HAC [6] builds on this by incorporating a hash grid to better organize scene primitives. Methods [14, 24] utilize techniques like distillation and Gaussian pruning to minimize storage requirements.

Notably, our focus is on addressing the additional representational redundancy introduced by the feed-forward pipeline (i.e., overleaping Gaussians) rather than the inherent redundancy of 3DGS itself. This issue is prevalent across existing generalizable 3DGS methods, and the aforementioned compression techniques are not suitable for these frameworks.

3. Method

Overview. As illustrated in Fig.2, SpatialSplat leverages a standard Vision Transformer (ViT) [10] backbone with a few Dense Prediction Transformer (DPT) [35] heads. Given V unposed images at resolution $H \times W$ with their intrinsics $\{\boldsymbol{I}^v, \boldsymbol{K}^v\}_{v=1}^V$, SpatialSplat learn a mapping f_θ parameterd by θ that transforms the input images to an fine-grained instance-aware radiance field $\boldsymbol{\mathcal{F}_I}$ and a coarse semantic feature field $\boldsymbol{\mathcal{F}_S}$, which can be formulated as:

$$
\begin{aligned}
&f_\theta : \{\boldsymbol{I}^v, \boldsymbol{K}^v\}_{v=1}^V \rightarrow \boldsymbol{\mathcal{F}_I} + \boldsymbol{\mathcal{F}_S}, \\
&\boldsymbol{\mathcal{F}_I} = \{\cup(\boldsymbol{f}_I^i, \boldsymbol{\beta}^i, \boldsymbol{\mu}^i, \boldsymbol{\alpha}^i, \boldsymbol{r}^i, \boldsymbol{s}^i, \boldsymbol{c}^i)\}_{i=1,\dots,V\times H\times W}, \quad (1) \\
&\boldsymbol{\mathcal{F}_S} = \{\cup(\boldsymbol{f}_S^j, \boldsymbol{\mu}^j, \boldsymbol{\alpha}^j, \boldsymbol{r}^j, \boldsymbol{s}^j, \boldsymbol{c}^j)\}_{j=1,\dots,V\times \frac{H}{S}\times \frac{W}{S}}.
\end{aligned}
$$

$\boldsymbol{\beta} \in \mathbb{R}$ is the importance score, which will be explained in Section 3.2. $\boldsymbol{f}_I \in \mathbb{R}^N$ is the learnable instance feature and $\boldsymbol{f}_S \in \mathbb{R}^M$ is the learnable semantic feature, as detailed in Section 3.3. The remains represent the vanilla Gaussian parameters [19], including the center position $\boldsymbol{\mu} \in \mathbb{R}^3$, opacity $\boldsymbol{\alpha} \in \mathbb{R}$, rotation factor in quaternion $\boldsymbol{r} \in \mathbb{R}^4$, scale $\boldsymbol{s} \in \mathbb{R}^3$, and spherical harmonics (SH) $\boldsymbol{c} \in \mathbb{R}^k$ with k degrees of freedom. In the following sections, we provide a detailed explanation of each component of our method.

3.1. 3D Geometry Prediction

Both the encoder and decoder in our geometric prediction module are built on pure ViT structures, requiring no geometric priors as in previous methods [3, 7]. The input image is patchified and flattened into image sequences, which along with the camera intrinsics processed by a linear layer, are fed into the encoder. The encoder weights are shared across different input views. The features from encoder are then passed to a ViT-based decoder, where cross-attention is applied to better capture spatial relationships and aggregate information across views. Finally, the features from different views are separately fed into a series of DPT heads to predict point maps and other Gaussian parameters. The point maps serve as Gaussians centers. Since the input im-

ages are unposed, the Gaussians centers suffer from scale ambiguity. Some methods [15, 38] rely on depth supervision to obtain per-scene scale, but this introduces additional errors due to imperfect depth. To overcome this limitation, we follow NoPoSplat's [46] approach by injecting camera intrinsics to resolve scale ambiguity. Experiments show that SpatialSplat effectively learns 3D priors from sparse unposed images without depth supervision, even while jointly learning multiple parameters and features.

3.2. Selective Gaussian Mechanism

Previous methods [3, 7, 15, 38, 46] directly use the pixel-wise pointmap as Gaussian centers which introduce significant redundancy. To address this, we propose a selective Gaussian mechanism that assigns each primitive an importance score to quantify its necessity for the scene representation. Primitives with importance scores below a certain threshold are considered redundant and discarded. To effectively supervise the learning of importance scores, we analyze Gaussian properties and the rendering pipeline, designing a simple yet effective strategy. Specifically, a primitive's influence can be reduced by decreasing its opacity or size [19]. However, we observe that even primitives with near-zero size still contribute to the rendering by occupying a pixel. Therefore, we multiply the importance score to the opacity of each primitive and modify the alpha blending formulation as follows:

$$\boldsymbol{c} = \sum_{i=1}^{n} \boldsymbol{c}_i \boldsymbol{\alpha}_i \boldsymbol{\beta}_i \prod_{j=1}^{i-1} (1 - \boldsymbol{\alpha}_j \boldsymbol{\beta}_j), \tag{2}$$

where $\boldsymbol{c}$ is the final intensities calculated by blending n ordered Gaussians overlapping the pixels, and $\boldsymbol{\beta}_i$ is the importance score processed through a sigmoid activation function. This ensures that primitives with low importance have minimal impact on the final intensities. Removing important primitives leads to a significant increase in photometric loss, whereas discarding redundant ones has little effect. Therefore, we optimize $\boldsymbol{\beta}_i$ through photometric loss minimization. Following methods [15, 46], we define the photometric loss as a weighted sum of MSE and LPIPS [50]:

$$\mathcal{L}_P(\boldsymbol{C}, \hat{\boldsymbol{C}}) = \mathcal{L}_{MSE}(\boldsymbol{C}, \hat{\boldsymbol{C}}) + \lambda \mathcal{L}_{LPIPS}(\boldsymbol{C}, \hat{\boldsymbol{C}}), \tag{3}$$

where $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$ are rasterized and GT pixel intensities, and $\lambda = 0.05$. Since redundant primitives are directly discarded during inference, we encourage their importance scores to approach zero to minimize the discrepancy between training and inference. Simultaneously, we aim to reduce the number of primitives used, striving to eliminate redundant ones as much as possible. To achieve these objectives, we first design the importance score with inspiration from Leaky ReLU [28], formulated as follows:

$$\boldsymbol{\beta_i} = \begin{cases} \boldsymbol{\beta}_i & \text{if } \boldsymbol{\beta}_i > \tau \\ \boldsymbol{\beta}_i \times 10^{-3} & \text{if } \boldsymbol{\beta}_i < \tau \end{cases}. \tag{4}$$

This ensures that the importance scores of primitives below the threshold τ are close to zero while preventing gradient vanishing. Additionally, we introduce a Binary Cross Entropy (BCE) loss with a regularization term that pushes $\boldsymbol{\beta}_i$ to either 0 or 1 while penalizing excessive use of primitives:

$$\mathcal{L}_I(\boldsymbol{S}) = \mathcal{L}_{BCE}(\boldsymbol{S}, \hat{\boldsymbol{S}}) + \frac{1}{\|\boldsymbol{S}\|} \sum_{\boldsymbol{\beta}_i \in \boldsymbol{S}} \boldsymbol{\beta}_i, \tag{5}$$

where

$$\boldsymbol{S} = \{\boldsymbol{\beta}_i\}_{i=1}^{H \times W}, \quad \hat{\boldsymbol{S}} = \begin{cases} 1 & \text{if } \boldsymbol{\beta}_i > \tau \\ 0 & \text{if } \boldsymbol{\beta}_i < \tau \end{cases}.$$

As a result, SpatialSplat achieves a compact yet high-fidelity scene representation by selectively retaining only the most essential primitives.

3.3. Dual-field Architecture

While distilling 2D semantic features into dense 3D representations has proven effective for scene understanding, conventional per-primitive compressed feature assignments [31, 36, 52] result in information loss and suboptimal performance. To mitigate this loss without increasing storage costs, we propose a dual-field architecture that decouples semantic representation into: 1) a fine-grained instance-aware radiance field, capturing scene geometry, textures, and instance correspondences, and 2) a coarse, uncompressed semantic feature field that reduces storage requirements while preserving semantic accuracy.

Fine-grained Instance-Aware Radiance Field. Our instance-aware radiance field builds upon Vanilla 3DGS, initializing the redundancy-aware Gaussians augmented with learnable instance features $\boldsymbol{f}_I$. We train the network using contrastive learning [1] to pull rendered instance features of the same instance closer while push those of different instances further apart. To achieve this, we lift multi-view 2D segmentation priors into 3D to help SpatialSplat learn inter-instance relationships. Given a target image $\boldsymbol{I}$, we first use Segment Anything Model (SAM) [21] to extract a set of masks $\mathcal{M} = \{\boldsymbol{M^k} \mid k = 1, 2, ..., m\}$. The instance feature map rendered at target image view with Eq. 2 is denoted as $\boldsymbol{F}_I$. The contrastive loss is then defined as:

$$\mathcal{L}_C(\boldsymbol{F}_I, \mathcal{M}) = -log \frac{1}{\|\boldsymbol{I}\|^2} \sum_{u \in \boldsymbol{I}} \sum_{u' \in \boldsymbol{I}^+} sim(\boldsymbol{F}_I(u), \boldsymbol{F}_I(u')) + \sum_{u' \notin \boldsymbol{I}^+} (1 - sim(\boldsymbol{F}_I(u), \boldsymbol{F}_I(u'))), \tag{6}$$

Methods	Semantic	Forward	Source View		Target View				
			mIoU ↑	Acc. ↑	mIoU ↑	Acc. ↑	PSNR ↑	SSIM ↑	LPIPS ↓
L-Seg [26]	✔	✗	0.5541	0.7824	0.5558	0.7856	N/A	N/A	N/A
NeRF-DFF [23]	✔	✗	0.5381	0.7724	0.5137	0.7582	22.49	0.7650	0.2829
Feature-3DGS [52]	✔	✗	0.4992	0.7362	0.3223	0.5674	17.96	0.5812	0.4894
NoPoSplat [46]	✗	✔	N/A	N/A	N/A	N/A	25.70	0.8159	0.1875
LSM [15]	✔	✔	0.5141	0.7719	0.5104	0.7662	24.12	0.7961	0.2526
SpatialSplat-Lite	✔	✔	0.5272	0.7801	0.5265	0.7693	25.45	0.8032	0.2039
SpatialSplat	✔	✔	0.5593	0.7814	0.5587	0.7924	25.46	0.8045	0.2046

Table 1. **Quantitative Comparison in 3D Tasks on Scannet dataset**. Our method outperforms both the latest SOTA semantic-aware feed-forward approach and per-scene optimization methods. "-Lite": replacing the pretrained 2D model LSeg with CLIP ViT-B/16.

where $\boldsymbol{I}^+$ is the set of pixels that belong to the same instance mask, $\boldsymbol{F}_I(u)$ is the feature vector of the $\boldsymbol{F}_I$ at coordinate u and $sim(.)$ is the cosine similarity. With this weak supervision, SpatialSplat effectively clusters radiance primitives by instance affiliation. Notably, the loss complexity in Eq. 6 scales quadratically with the number of pixels, making training infeasible. To address this, we employ a trick (detailed in the appendix) that enables efficient loss estimation with linear complexity.

Coarse Semantic Feature Field. Since primitives within the same instance share semantic consistency, SpatialSplat learns instance-level semantics using only a subset of primitives, avoiding the redundancy of per-primitive assignments. Manually selecting primitives for each instance, however, disrupts the network's cohesion and simplicity while increasing additional computational costs. To address this, we uniformly downsample the Gaussian centers from $\boldsymbol{\mathcal{F}_I}$ to form a coarse feature field, where each selected primitive is assigned a semantic feature $\boldsymbol{f}_S$. To enhance semantic learning, we inject image features from a pretrained 2D model into feature heads, as shown in Fig. 2. A 2D semantic feature map $\boldsymbol{F}_S$ is rendered using alpha blending. We minimize the loss between the rendered feature map at a novel view and the feature map $\hat{\boldsymbol{F}}_S$ of the ground truth image extracted from the pretrained 2D model:

$$\mathcal{L}_S(\boldsymbol{F}_S, \hat{\boldsymbol{F}}_S) = MSE(\boldsymbol{F}_S, \hat{\boldsymbol{F}}_S) \tag{7}$$

During open-vocabulary querying, we first cluster primitives in $\boldsymbol{\mathcal{F}_I}$ by instance feature similarity, assigning each instance the average feature of its clustered primitives as its label. For primitives in $\boldsymbol{\mathcal{F}_S}$, their labels correspond to the instance features before sampling, and their semantic features are assigned to the instance with the highest cosine similarity. If multiple primitives contribute semantic features to the same instance, we take their average to ensure consistency. This approach enables the network to acquire sharper and clearer semantics compared to the privious methods. An additional advantage of dual-field architecture is that, unlike dense semantic supervision from LSeg [26], it allows the use of a much lighter pretrained model (e.g., CLIP ViT-B/16 [34]), significantly improving inference speed.

3.4. Training Objective

Following previous methods [15], we perform end-to-end training to optimize our model with the following objective:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_P(\boldsymbol{C}, \hat{\boldsymbol{C}}) + \lambda_1 \mathcal{L}_I(\boldsymbol{S}) \\ &+ \lambda_2 \mathcal{L}_C(\boldsymbol{F}_I, \mathcal{M}) + \lambda_3 \mathcal{L}_S(\boldsymbol{F}_S, \hat{\boldsymbol{F}}_S).\end{aligned} \tag{8}$$

The parameters λ_1, λ_2, λ_3 are set to 0.01, 0.2, and 1.0, respectively. To save training time, the instance masks $\mathcal{M}$ are generated by SAM [21] prior to training, while the semantic feature map $\hat{\boldsymbol{F}}_S$ is predicted during training.

4. Experiment

4.1. Experimental Setup

Datasets. Following LSM[15], we primarily train our model on a combination of ScanNet++[47] and ScanNet [8]. We filter out bad scenes and those with incomplete extrinsic parameters, resulting in a training dataset of approximately 1,500 scenes. For evaluation, we follow LSM and select 40 unseen scenes from ScanNet to assess our model's performance. We test our method's cross-dataset generalization on the synthesized Replica Dataset [39].

Evaluation Metrics. We evaluate our method on two downstream tasks: novel view synthesis (NVS) and 3D open-vocabulary segmentation (OVS). For NVS, we adopt the standard metrics: PSNR, SSIM, and LPIPS [50]. For OVS, we evaluate performance using class-wise intersection over union (mIoU) and average pixel accuracy (mAcc).

Baselines. We primarily compare our method with LSM [15], the latest state-of-the-art (SOTA) approach for generalizable semantic 3D reconstruction. Additionally, we include the following baselines: 1): Feature-3DGS [52]

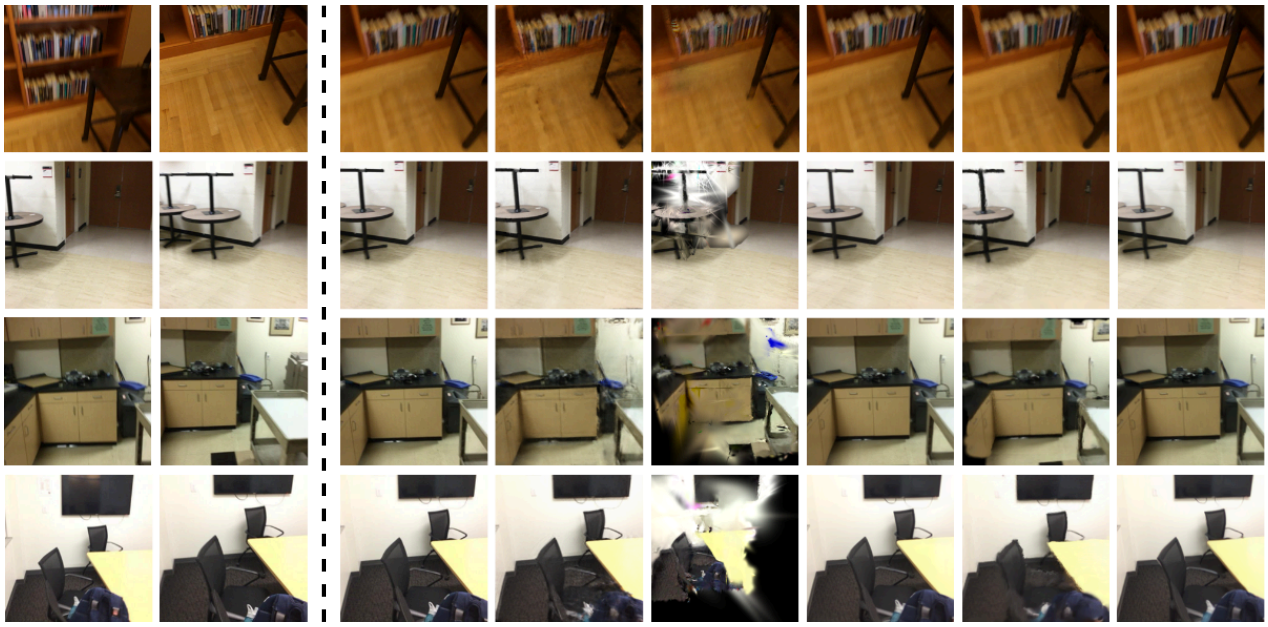

Figure 3. **Qualitative comparison in NVS.** SpatialSplat can synthesize realistic novel views. In challenging cases where LSM fails, such as the table legs in the first two rows and the corners in the last two rows, our method achieves significantly better results.

	LSM					SpatialSplat (Ours)				
Scene	mIoU ↑	Acc. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	Acc. ↑	PSNR ↑	SSIM ↑	LPIPS ↓
room0	0.5024	0.7423	17.61	0.5247	0.3228	**0.5131**	**0.7934**	**22.09**	**0.6760**	**0.1794**
room1	0.4661	0.7349	14.34	0.5913	0.3174	**0.4691**	**0.7628**	**20.62**	**0.6774**	**0.2265**
office3	0.5484	0.7627	21.28	0.7567	0.2084	**0.5582**	**0.7997**	**22.52**	**0.7913**	**0.1436**
office4	**0.5065**	**0.7723**	19.51	0.6962	0.2779	0.5016	0.7695	**21.13**	**0.7271**	**0.2072**

Table 2. **Out-of-distribution (OOD) comparison on Replica dataset.** SpatialSplat generalizes well on OOD data.

and NeRF-DFF [23], pre-scene optimization methods for semantic 3D reconstruction based on 3DGS [19] and NeRF [29], respectively. 2): NoPoSplat [46], a latest SOTA method for generalizable novel view synthesis. 3): LSeg [26], a 2D open-vocabulary segmenter used for feature distillation in all compared semantic-aware methods. Unless otherwise specified, our method also uses image features from LSeg for supervision.

Implementation details. For the 3D geometry prediction module, we use ViT-Large with a patch size of 16 as the encoder and ViT-Base as the decoder, both initialized with pre-trained weights from MASt3R [25], while the remaining modules are randomly initialized. We modify the 3DGS CUDA renderer to support instance and semantic feature map rendering. The importance score threshold τ is set to 0.5, the downsampling ratio to 8, the instance feature dimension N to 8, and the semantic feature dimension M to 512. For a fair comparison, we train the model at a resolution of 256×256, consistent with our baseline.

4.2. Results and Analysis

Results of Novel View Synthesis. Following previous methods [3, 46], we select two images from each test scene as input view for the generalizable 3DGS methods, and choose four additional images located between these two as target views to evaluate the performance of different methods in Novel View Synthesis. Notably, Feature-3DGS [52] and NeRF-DFF [23] relies on dense view inputs and performs poorly with only two input images. To account for this, we use five out of the six available images (the two input images and four target views) for training and evaluate on the remaining image. As shown in Tab. 1, SpatialSplat significantly outperforms latest SOTA method LSM. We observe that LSM struggles in certain scenes due to its reliance on accurate depth for aligning views during training as shown in Fig. 3. Although NeRF-DFF uses more images for training, our method still outperforms it and even demonstrates competitive performance with very re-

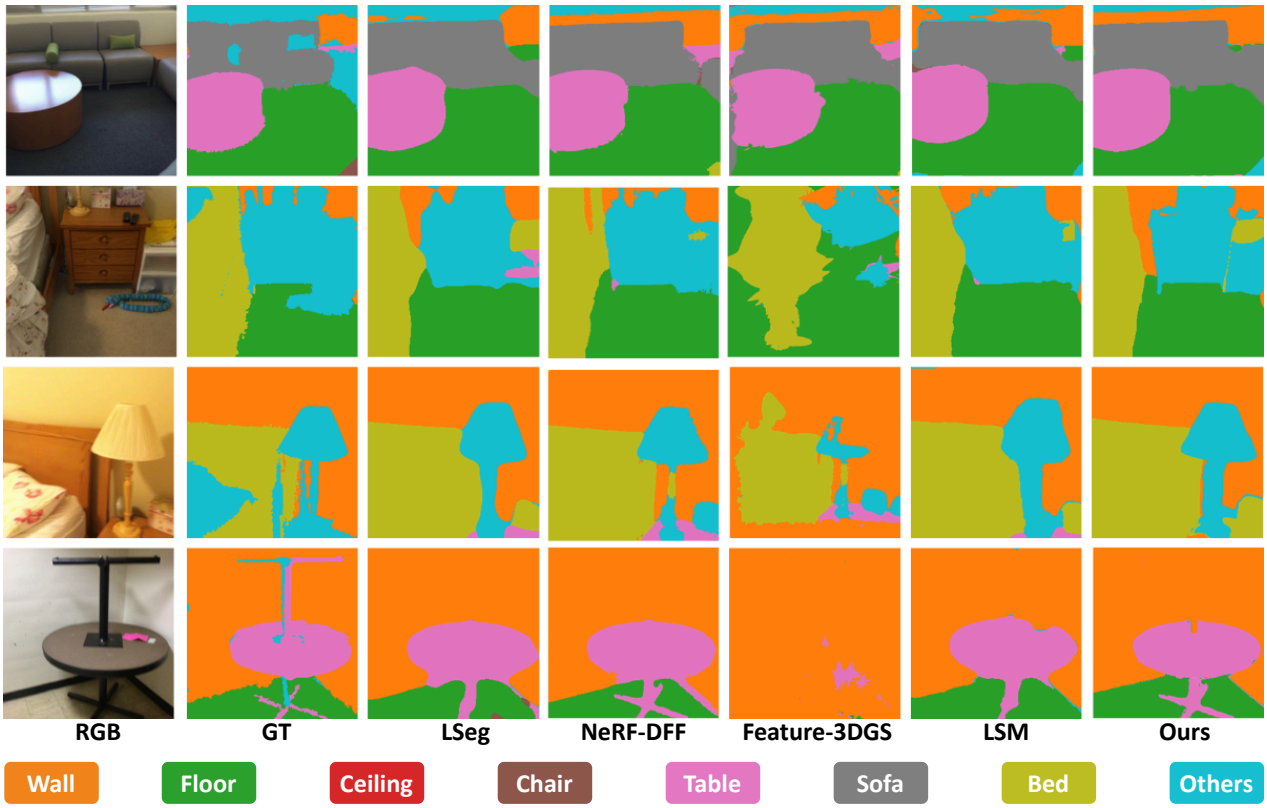

Figure 4. **Qualitative comparison in OVS.** SpatialSplat achieves sharper and more precise segmentation results compared to previous methods. Notably, our method excels in challenging details, such as distinguishing table legs from cabinet legs (the second the fourth rows), benefiting from our dual-field architecture that captures detailed instance information and uncompressed semantics.

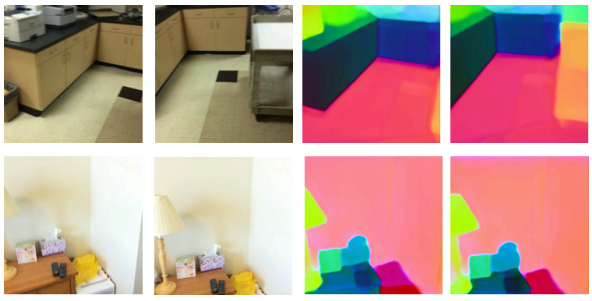

Figure 5. **The rendered instance features.** SpatialSplat predicts clear and consistent instance features across different views.

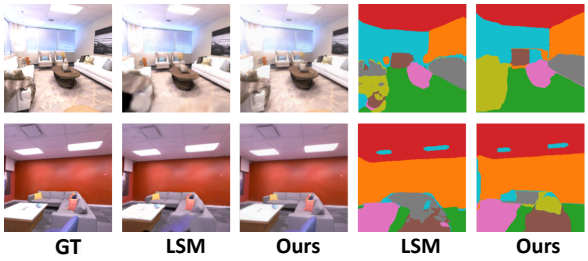

Figure 6. **Qualitative results of cross-dataset generalization.** Zoom out for clearer visualization.

cent SOTA methods designed specifically for novel view synthesis. This highlights that our method efficiently learns spatial priors from large-scale 2D data.

Results of Open-vocabulary 3D Segmentation. Following LSM's approach, we map category labels from the Scannet dataset into common categories: Wall, Floor, Ceiling, Chair, Table, Bed, Sofa, Others. As shown in Tab. 1, SpatialSplat outperforms all compared methods, even surpassing L-Seg, which provides semantic feature supervision for other compared methods. As illustrated in Fig. 4, SpatialSplat produces sharp and clear 3D semantic segmentation results. This is achieved by lifting both 2D instance and uncompressed semantic into 3D space, enabling more efficient semantic learning while mitigating the blurred contours and information loss associated with per-primitive compressed semantics. The visualizations in Fig. 5 show that SpatialSplat effectively learns a highly consistent instance prior with only weak supervision from 2D masks.

Results of Cross-Dataset Generalization. As shown in Tab. 2, our method significantly outperforms LSM on the

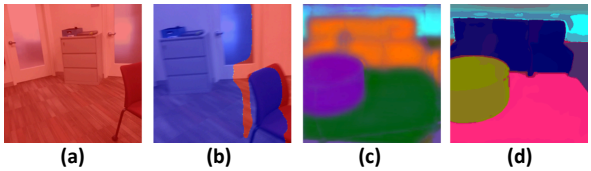

Figure 7. **Qualitative results of ablations.** (a) and (b) Qualitative results of importance score prediction, with red color indicating an importance score of 1 and blue indicating 0. (c) The rendered semantic feature without the dual-field. (d) The rendered semantic feature with the dual-field, which appears clearer and sharper.

Method	Latency↓	Gaussian Size ↓	Num. ↓
Feature-3DGS [52]	1069 s	229.65 MB	476K
NoPoSplat [46]	0.044 s	30.93 MB	131K
LSM [15]	0.104 s	63.00 MB	131K
SpatialSplat-Lite	**0.058 s**	25.58 MB	87.7K
SpatialSplat	0.071 s	**25.58 MB**	**87.7K**

Table 3. **Model efficiency comparison.** "-Lite": the model replaces LSeg with CLIP ViT-B/16.

unseen Replica dataset in both novel view synthesis and open-vocabulary segmentation. Notably, since Replica is a synthetic dataset with a different data modality from our training set, this underscores the strong generalization ability of our approach. The visualization in Fig. 6 further demonstrates that LSM struggles with coarse modeling at scene edges, leading to artifacts in synthesized views, whereas our method generates photorealistic renderings.

Analysis of Model Efficiency. We compare the model efficiency of different 3DGS-based methods. As shown in Tab. 3, our method is faster than LSM thanks to its streamlined architecture, requiring only 40% of the storage size and 65% of primitive number. Remarkably, our approach requires even fewer Gaussians parameters than NoPoSplat, which lacks semantic awareness. This demonstrates the effectiveness of our proposed dual-field architecture and selective Gaussian mechanism, which capture accurate semantics while eliminating redundant Gaussians. Furthermore, as our method does not rely on dense semantic supervision, we leverage a lightweight pretrained 2D model, significantly accelerating inference speed.

4.3. Ablations and Analysis.

We perform ablations to answer the following questions: (1) Are the primitives removed by our selective Gaussian mechanism truly redundant? (2) Does the improved scene understanding stem from the proposed dual-field architecture? (3) As the number of views increases, how does redundancy grow? Can our method effectively alleviate it?

Effect of selective Gaussian mechanism. To answer Question 1, we remove the SGM and retrain our model using the same configuration. As shown in Tab. 4, the SGM eliminates approximately 35% of Gaussian primitives while causing only a 0.08 drop in PSNR, indicating that the removed primitives have minimal impact on rendering quality. The visualization in Fig. 7 (a) and (b) further demonstrates that the SGM accurately localizes overlapping areas in the input images.

Model	mIoU↑	Acc.↑	PSNR↑	Num. ↓
No SGM	0.5569	0.7972	25.55	131k
No dual	0.5027	0.7786	25.46	87.6k
3-view	0.5125	0.7793	24.92	97.5k
SpatialSplat	0.5577	0.7920	25.47	87.7k

Table 4. **Ablation study on design choices.** The proposed Adaptive Gaussian mechanism and dual-field architecture jointly enable the construction of high-quality representations with significantly reduced dimensionality.

Effect of dual-field architecture. To address Question 2, we replace our dual-field architecture with single-level per-primitive semantic learning. Following the approach in LSM, we set the semantic feature dimension to 64 and use a convolutional layer to expand it to 512. As shown in Tab. 4, this results in a significant drop of 5% in mIoU for open-vocabulary segmentation performance. The primary issue is that per-primitive semantic learning struggles to maintain accurate semantics and fails to preserve clear instance boundaries, as illustrated in Fig. 7 (c) and (d).

Extend to more views. Our experiments primarily focus on two-view settings to ensure a fair comparison with our baseline. However, our method can be extended to more views for reconstructing larger scenes. As shown in Tab. 4, our approach remains effective even with three input views. Moreover, compared to pixel-wise methods, which require approximately 197K primitives to represent the scene, our method achieves a more compact representation by reducing about 50% primitives. More qualitative details can be found in the appendix.

5. Conclusion

We present an efficient feed-forward network that generates compact yet high-performance representations for 3D reconstruction and scene understanding from sparse, unposed images. Our proposed dual-field architecture achieves state-of-the-art performance while significantly reducing the number of representation parameters. Additionally, we introduce a Selective Gaussian Mechanism that directly identifies redundant Gaussians caused by pixel-wise predictions without requiring geometric priors. SpatialSplat achieves impressive results in 3D reconstruction and scene understanding without relying on any 3D data during training or inference, making semantic 3D reconstruction from sparse, unposed images more practical.

Acknowledgments

This work was partially supported by Hunan Province Major Scientific and Technological Project No. 2024QK2001, National Key R&D Program of China No. 2023YFB4704500.

References

[1] Yash Bhalgat, Iro Laina, João F. Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 4

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021. 2

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. Pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 19457–19467. IEEE, 2024. 2, 3, 4, 6

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022. 2

[5] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *CoRR*, abs/2401.03890, 2024. 1

[6] Yihang Chen, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. HAC: hash-grid assisted context for 3d gaussian splatting compression. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VII*, pages 422–438. Springer, 2024. 3

[7] 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 *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXI*, pages 370–386. Springer, 2024. 2, 3, 4

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5

[9] Jianmei Dai, Zhilong Zhang, Shiwen Mao, and Danpu Liu. A view synthesis-based 360° VR caching system over mec-enabled C-RAN. *IEEE Trans. Circuits Syst. Video Technol.*, 30(10):3843–3855, 2020. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[11] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4970–4980. IEEE, 2023. 2

[12] 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 *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[13] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *CoRR*, abs/2403.20309, 2024. 2

[14] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ FPS. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 3

[15] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, and Marco Pavone. Large spatial model: End-to-end unposed images to semantic 3d. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1, 2, 3, 4, 5, 8

[16] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12479–12488. IEEE, 2023. 2

[17] Qiao Gu, Zhaoyang Lv, Duncan P. Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIII*, pages 382–400. Springer, 2024. 3

[18] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18344–18347. IEEE, 2022. 2

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 1, 2, 3, 4, 6

[20] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19672–19682. IEEE, 2023. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 4, 5

[22] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[23] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 5, 6

[24] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21719–21728. IEEE, 2024. 3

[25] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2, 6

[26] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2, 5, 6

[27] Luca De Luigi, Damiano Bolognini, Federico Domeniconi, Daniele De Gregorio, Matteo Poggi, and Luigi Di Stefano. Scannerf: a scalable benchmark for neural radiance fields. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 816–825. IEEE, 2023. 3

[28] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, GA, 2013. 4

[29] 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. *Commun. ACM*, 65(1):99–106, 2022. 1, 2, 6

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[31] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 20051–20060. IEEE, 2024. 2, 3, 4

[32] Ri-Zhao Qiu, Yafei Hu, Ge Yang, Yuchen Song, Yang Fu, Jianglong Ye, Jiteng Mu, Ruihan Yang, Nikolay Atanasov, Sebastian A. Scherer, and Xiaolong Wang. Learning generalizable feature fields for mobile manipulation. *CoRR*, abs/2403.07563, 2024. 1

[33] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Language-driven physics-based scene synthesis and editing via feature splatting. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLI*, pages 368–383. Springer, 2024. 2, 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 5

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12159–12168. IEEE, 2021. 3

[36] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, pages 405–424. PMLR, 2023. 1, 2, 4

[37] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9043–9052. IEEE, 2023. 2

[38] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2, 4

[39] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[40] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *International Conference on 3D Vision, 3DV 2022, Prague, Czech Republic, September 12-16, 2022*, pages 443–453. IEEE, 2022. 2

[41] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4690–4699. Computer Vision Foundation / IEEE, 2021. 2

[42] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[43] Xingrui Wang, Cuiling Lan, Hanxin Zhu, Zhibo Chen, and Yan Lu. Gsemsplat: Generalizable semantic 3d gaussian splatting from uncalibrated image pairs. *CoRR*, abs/2412.16932, 2024. 2

[44] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2

[45] Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. Murf: Multi-baseline radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 20041–20050. IEEE, 2024. 2

[46] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 2, 4, 5, 6, 8

[47] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 5

[48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4578–4587. Computer Vision Foundation / IEEE, 2021. 2

[49] Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. *CoRR*, abs/2408.13770, 2024. 2

[50] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 4, 5

[51] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15818–15827. IEEE, 2021. 2

[52] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21676–21685. IEEE, 2024. 2, 3, 4, 5, 6, 8

[53] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 65, 2018. 1