Open-Vocabulary Scene Understanding with 3D Gaussian Splatting

Shaohui Dai[1*](23020241154342), **Hengyan Ma**[1*](31520241154519), **Dian Chen**[2*](36920241153197), **Shuyan Ke**[2*](36920241153224), **Yijun Wang**[2*](36920241153252)

[1]Deep Learning Class from *School of Informatics*
[2]Deep Learning Class from *Institute of Artificial Intelligence*

Abstract

Open-vocabulary 3D scene understanding is a challenging task with significant applications in areas such as embodied agents and augmented reality. While traditional methods, including those based on Neural Radiance Fields (NeRF), have demonstrated potential, they are hindered by slow training and rendering times due to their reliance on implicit scene representations. In this paper, we present OV Gaussian Splatting (OVGS), a novel open-vocabulary 3D scene understanding framework built on 3D Gaussian Splatting. Our approach leverages semantic features distilled from pre-trained image encoders into 3D Gaussian representations and introduces a compression mechanism to enable efficient scene understanding without extensive training. Experiments show that OVGS outperforms previous methods, achieving a 40% IoU improvement on the Mip-NeRF 360 dataset.

Introduction

Recent advancements in computer vision have significantly enhanced AI systems' ability to comprehend and interact with three-dimensional environments. A critical area of progress is open-vocabulary 3D scene understanding, which enables flexible, natural language-driven interactions with 3D spaces. This capability is increasingly vital for applications such as augmented reality (AR) and embodied AI, where seamless and human-like interaction with digital environments is essential.

Despite its promise, the development of efficient 3D scene understanding models has been hindered by the absence of large-scale 3D scene datasets with detailed language annotations. A common approach to addressing this challenge involves leveraging 2D semantic knowledge derived from images and extending it into 3D space. By learning from these 2D representations, semantic information can be projected into 3D, enabling accurate segmentation from novel viewpoints. This method also supports tasks like retrieving specific 3D objects from scenes, which is crucial for object recognition, manipulation, and broader scene understanding. Semantic segmentation, in particular, remains a cornerstone of 3D scene understanding.

Previous approaches have tackled these challenges by distilling knowledge from vision-language models, such as CLIP (Radford et al. 2021), into Neural Radiance Fields (NeRF) (Mildenhall et al. 2021). While NeRF excels at generating novel scene views, its implicit scene representation leads to slow training and rendering times, limiting its scalability. Moreover, NeRF-based methods (Zhi et al. 2021; Kerr et al. 2023) that project 2D semantic features into 3D often suffer from blurred semantic boundaries, which hinders their performance for open-vocabulary queries.

Recently, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has emerged as a compelling alternative for scene reconstruction, offering faster and more scalable 3D representations. Building on this, Feature 3DGS (Zhou et al. 2023) combines efficient 3DGS with feature distillation, delivering notable improvements over NeRF-based approaches. However, these methods continue to face challenges such as high memory consumption and prolonged training times.

In this paper, we introduce Open-Vocabulary Gaussian Splatting (OVGS), a novel framework that distills vision-language models (VLMs) to reconstruct a 3D semantic field. We leverage the APE model, a sota VLM, to extract more accurate and fine-grained semantic features, significantly enhancing the precision of object recognition and scene understanding. Additionally, we introduce an efficient compression mechanism that condenses abundant, noisy scene semantics into a compact Codebook, optimizing both storage and computational performance while maintaining high-quality semantic reconstruction. This approach enables precise and efficient scene understanding, which is crucial for open-vocabulary queries and real-time applications.

In summary, the main contributions of our work include:

- We introduce OVGS, a novel framework based on 3D Gaussian Splatting designed for accurate and scalable open-vocabulary 3D semantic perception.
- We propose an efficient compression mechanism that reduces noisy, high-dimensional semantic features into compact, low-dimensional representations, facilitating faster rendering and scene understanding.
- We demonstrate the superiority of OVGS in segmentation accuracy and efficiency, achieving significant improvements over existing methods and enabling a variety of downstream tasks.

[*]Equal Contribution.

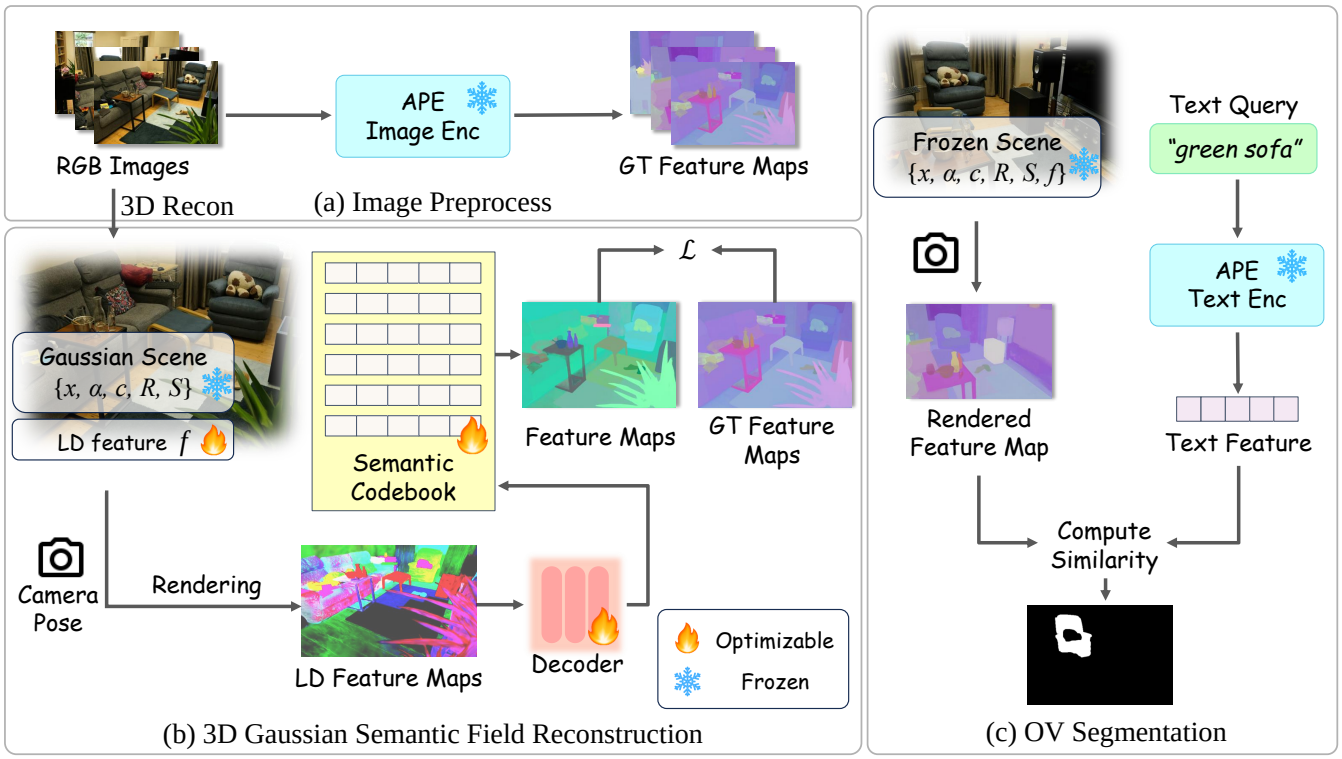

Figure 1: Overall framework of the proposed method. Top left: Image preprocessing for encoding multi-view images. Bottom left: 3D semantic field reconstruction. A low-dimensional (LD) feature map is rendered and then transformed into the predicted feature map using the Semantic Codebook. The loss between the predicted and ground truth feature maps is computed for optimization. Right: Pipeline illustrating the open-vocabulary querying process.

Related Works

Neural Rendering

Recent advancements in neural network-based 3D scene representation have significantly pushed the boundaries of what is possible. Among these, Neural Radiance Fields (NeRF) (Mildenhall et al. 2021; Barron et al. 2021, 2022, 2023) have stood out for their exceptional performance in novel view synthesis. However, NeRF's reliance on a fully implicit neural representation results in time-consuming training and rendering processes. To address these limitations, many subsequent methods have focused on optimizing NeRF's performance (Chen et al. 2022; Müller et al. 2022; Reiser et al. 2023).

A significant leap in the field was introduced with 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023), which substantially boosts rendering speed and achieves high-quality scene reconstruction. 3DGS represents a scene as a collection of Gaussian primitives. Rendering is made efficient by rasterizing these primitives into images. This method enables fast learning and facilitates the manipulation of specific scene elements without impacting other components.

3D Scene Understanding

Recent advancements in 3D scene understanding have made significant strides by incorporating semantic information into NeRF-based frameworks. For example, Semantic-NeRF (Zhi et al. 2021), an influential early work, introduced the integration of manually labeled semantic information into the output of NeRF, enabling the generation of high-quality semantic maps alongside appearance and geometry reconstructions. DFF (Kobayashi, Matsumoto, and Sitzmann 2022) and LERF (Kerr et al. 2023) adopt a different approach by extracting semantic features from Vision-Language Models (VLMs) such as LSeg and CLIP, allowing for open-vocabulary queries based on natural language descriptions. These methods pave the way for more flexible, language-driven 3D scene understanding.

More recently, Feature 3DGS has combined the efficiency of 3DGS with VLMs, embedding semantic features directly into Gaussian-based 3D scenes. This approach benefits from fast rendering speeds and scalable scene reconstruction while maintaining the flexibility to handle complex queries in open-vocabulary settings. However, the high dimensionality of the semantic features in Feature 3DGS leads to challenges such as long training times and excessive memory usage, limiting its practical scalability for large or complex scenes.

Methodology

Given a set of posed images $\mathcal{I} = \{I_1, I_2, \ldots, I_K\}$, we employ the Gaussian Splatting technique (Kerbl et al. 2023)

to reconstruct a photorealistic 3D scene, denoted as S. Our method extends S with open-vocabulary semantics, allowing us to accurately identify objects within the scene based on arbitrary natural language descriptions.

Figure 1 illustrates the overall pipeline of our method. Initially, we employ a frozen image encoder from a pre-trained Vision-Language Model (VLM) to process each image I_k, generating 2D semantic feature maps $\mathcal{V} = \{V_1, V_2, \ldots, V_K\}$. To integrate these high-dimensional 2D feature maps into the 3D Gaussian Splatting framework, we propose an efficient compression mechanism that ensures minimal storage overhead and optimal computational performance. This allows us to extend 3DGS and reconstruct a 3D semantic field. Following the reconstruction, we enable object querying through natural language descriptions, which results in 2D segmentation masks for the desired objects.

Preliminary: 3D Gaussian Splatting

3D Gaussian Splatting employs a set of 3D Gaussian primitives to model the appearance and geometry of a scene, closely resembling traditional point cloud representations. The discrete nature of these Gaussian primitives enables fast rendering, as they can be efficiently rasterized into images given the camera poses.

Each Gaussian primitive is parameterized by its center $\mu \in \mathbb{R}^3$, a 3D anisotropic covariance matrix Σ in world coordinates, an opacity value α, and a color c. During the rendering process, these Gaussian primitives are projected onto the 2D image plane, where the splatting process transforms the 3D Gaussian primitives into 2D ellipses.

Following splatting, a volumetric rendering process is applied in the rasterization step to compute the color c_{2D} of each pixel. This process is consistent with traditional NeRF methods:

$$c_{\text{2D}} = \sum_{i \in \mathcal{G}} c_i \alpha_i T_i, \tag{1}$$

where $\mathcal{G}$ denotes a set of Gaussian primitives sorted by depth, and T_i represents the transmittance. The transmittance is defined as the cumulative product of the opacity values of all Gaussian primitives superimposing on the same pixel, computed as $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$.

Pixel-aligned Feature Extraction

Prior research has widely employed CLIP for feature lifting in 3D radiance fields, due to its superior capability in handling open-vocabulary queries. Works such as (Kobayashi, Matsumoto, and Sitzmann 2022) and (Zhou et al. 2023) use LSeg (Li et al. 2022) to extract pixel-aligned CLIP features. However, LSeg struggles to capture the appearance characteristics of objects and is particularly ineffective at recognizing long-tail objects.

To address the limitation of CLIP, which provides only image-level features, methods like LERF (Kerr et al. 2023) adopt a feature pyramid approach, using cropped image encodings to represent local features. While these methods extract pixel-level features from the CLIP model, the resulting feature maps lack geometric boundaries and do not correspond directly to the objects in the scene.

To overcome these challenges, we turn to the Aligning and Prompting Everything All at Once (APE) model (Shen et al. 2023), which efficiently aligns vision and language features to produce pixel-level image features. This makes APE a robust solution for feature lifting. We make minor modifications to the APE model to efficiently extract pixel-aligned features with fine boundaries (approximately 2 seconds per image). We treat these encoded pixel-aligned feature maps as pseudo ground truth features, hereafter referred to as *GT*.

Using the APE-generated feature maps $\mathcal{V}$ for all training images in $\mathcal{I}$, we embed the semantic features into each Gaussian primitive, enabling the reconstruction of a 3D semantic field.

Condensed Semantic Feature Embedding

Since VLMs like APE and CLIP are trained on large datasets with extensive text-image pairs, the feature dimensionality they produce is quite high (often exceeding 256). Directly embedding these high-dimensional semantic features into each Gaussian primitive could lead to excessive storage and computational overhead. Given that 3DGS supports real-time rendering, we expect our approach to also achieve real-time performance for scene understanding and semantic rendering under optimal conditions.

However, we observe that the semantics of a single scene occupy only a small portion of the VLM feature space. This allows us to exploit scene priors to compress the semantic features embedded within the scene, reducing both storage and computational costs. Additionally, due to the inherent multi-view inconsistency of the encoded 2D semantic feature maps, Gaussians tend to overfit to individual training viewpoints, inheriting these inconsistencies and leading to discrepancies between the 3D and 2D representations of an object.

Therefore, we introduce the *Semantic Codebook*, which leverages scene priors to compress the semantic space of a scene into an N-length codebook, allowing for a more efficient and consistent representation across views.

3D Semantic Field Reconstruction

We introduce a low-dimensional semantic feature, denoted as f, into each Gaussian primitive, exploiting the redundancy of high-dimensional semantics across the scene and its dimensions to facilitate efficient rendering. Similar to previous works, to create a 2D semantic representation, we employ a volumetric rendering process analogous to color rendering, applied to the low-dimensional semantic feature.

$$f_{\text{2D}} = \sum_{i \in \mathcal{G}} f_i \alpha_i T_i. \tag{2}$$

Here, f_{2D} represents the pixel-wise low-dimensional feature. We then use a Multi-Layer Perceptron (MLP) as a feature decoder to obtain logits e, which are activated using the Softmax function to determine the index of the corresponding entry in the Semantic Codebook.

Thus, the low-dimensional feature f_{2D} can be recovered to semantic feature v through the MLP decoder $\mathcal{D}$ and the Semantic Codebook $\mathcal{T}$.

$$v = \mathcal{T}\left[\underset{j=1,2,\ldots,N}{\operatorname{argmax}}\left(e_j\right)\right], \tag{3}$$

where $e = \mathcal{D}(f_{2D}) \in \mathbb{R}^N$ and $\mathcal{T}[i]$ represents the i-th item in array. Thus, embedded features can be restrained to a compact and finite semantic space.

In the initial phase of semantic field optimization, our focus is on learning the Codebook from GT features. We observe a similarity between the learning process of the Codebook and the contrastive pre-training used in CLIP: features in the Codebook are aligned with the GT features, and each GT feature, denoted as v_{gt}, is assigned to the Codebook entry with the highest similarity. However, the assignment of a pixel feature to a particular entry is not predetermined, rather it pivots on similarity. Therefore, we devise a self-supervised loss function aimed at reducing the self-entropy of the clustering process, thereby improving the alignment between the Semantic Codebook entries and the GT features.

$$\mathcal{L}_{\text{sc}} = -\sum\nolimits_{j=1}^{N} p_j \log(p_j), \tag{4}$$

where $p_j = \text{Softmax}\left(\cos\left\langle v_{gt}, \mathcal{T}[j]\right\rangle \cdot t\right)$ and t is an annealing temperature.

Subsequently, we undertake a joint optimization of the low-dimensional features $\hat{f}$ and the MLP decoder $\mathcal{D}$. Ideally, the feature recovered from the low-dimensional feature should closely correlate with the GT feature v_{gt}. As a result, we impose a cross-entropy constraint geared towards aligning the entries' logits of the low-dimensional features with the assigned GT entry d,

$$\mathcal{L}_{\text{joint}} = -\sum\nolimits_{j=1}^{N} \text{onehot}(d)_j \log(e_j), \tag{5}$$

Finally, to bolster the robustness of this procedure, we introduce an end-to-end regularization, directly optimizing the cosine similarity of 2D semantic feature and corresponding ground truth,

$$\mathcal{L}_{\text{e2e}} = 1 - \cos\left\langle v_{gt}, v\right\rangle. \tag{6}$$

The comprehensive loss function designated for our semantic field reconstruction process is represented as $\mathcal{L}$,

$$\mathcal{L} = \lambda_{\text{sc}}\mathcal{L}_{\text{sc}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{e2e}}\mathcal{L}_{\text{e2e}}. \tag{7}$$

Experiments

Evaluation Setup

To assess the effectiveness of our approach, we conduct experiments on the Mip-NeRF 360 dataset (Barron et al. 2022). Mip-NeRF 360 is a high-quality real-world dataset containing a variety of objects with rich details in the images. It is extensively used in 3D reconstruction and novel view synthesis. We selected four scenes (Room, Bonsai, Garden, and Kitchen), representing both indoor and outdoor environments, for our evaluations. The test set of Mip-NeRF 360 includes semantic annotations from partial viewpoints, along with text prompts and 2D masks. During testing, we use the given text prompt as queries to predict the 2D mask of the object. We then evaluate the quality of the predicted mask.

To assess the accuracy of open-vocabulary querying results, we employ mean intersection over union (mIoU), mean accuracy (mAcc), and mean precision (mP) as evaluation metrics. Additionally, to further evaluate model performance, we measure both the training duration and rendering frame rate (Frame Per Second).

Comparision

We conduct a comparative evaluation of our approach in contrast with Feature 3DGS (Zhou et al. 2023), and LERF (Kerr et al. 2023).

Quantitative Results. Table 1 presents a comparative analysis of our method's performance against other state-of-the-art approaches. As shown, our segmentation results notably surpass those of both LERF and Feature 3DGS, achieving a significant 40% improvement in mean Intersection over Union (mIoU) on the Mip-NeRF 360 dataset. This substantial enhancement demonstrates the superior accuracy and efficiency of our approach in open-vocabulary 3D scene understanding.

Additionally, Table 2 highlights the time efficiency of our method. We report the image encoding time, scene training duration, total processing time, and rendering frame rates for each approach. By deriving a highly efficient visual encoder from APE, we reduced the encoding time for a single image to 2 seconds. Moreover, with the help of our effective compression mechanism and training regularization, we are able to reconstruct a semantic field in as little as 40 minutes, significantly improving processing speed compared to previous methods.

Method	mIoU ↑	mAcc ↑	mP ↑
LERF	0.2698	0.7796	0.3293
Feature 3DGS	0.4155	0.9189	0.5152
Ours	**0.8646**	**0.9569**	**0.9362**

Table 1: Evaluation metrics for comparing our method with others on Mip-NeRF 360 dataset.

Qualitative Results. We present qualitative results generated by our method, compared with other state-of-the-art approaches. Figure 2 showcases the performance of our open-vocabulary query system on the Mip-NeRF 360 test data, with a particular emphasis on the ability to handle queries that describe the appearance and texture of the objects in the scene.

LERF (Kerr et al. 2023) produces imprecise and ambiguous 3D features, which impede the clear delineation of boundaries between the target region and surrounding objects. This lack of precision leads to blurry segmentations and makes it difficult to distinguish between different regions within the same scene. Feature 3DGS (Zhou et al. 2023) utilizes a 2D semantic segmentation model, LSeg (Li et al. 2022), as its feature extractor. However, like LSeg, it struggles with open-vocabulary queries. Specifically, it tends

Figure 2: Visualization of open-vocabulary querying results. From top to bottom: Ground Truth images, querying results from OVGS, Feature 3DGS, and LERF. The corresponding textual descriptions for each row are provided at the bottom.

to retrieve all objects related to the query prompt, making it ineffective in handling nuanced distinctions. For instance, it has difficulty differentiating between a sofa and a toy placed on it, leading to less precise object segmentation and an overall lower quality of scene understanding.

Method	Preprocess	Training	Total	FPS
LERF	**3min**	50min	54min	0.04
Feature3DGS	25min	623min	648min	12
Ours	9min	**40min**	**49min**	**75**

Table 2: Time evaluation for training and rendering on Mip-NeRF 360 dataset.

Downstream Application

Our method can be applied to various downstream tasks, with a direct application in 3D scene editing. As shown in Figure 3, we use the text query “flowerpot” to locate the relevant 3D Gaussian primitives. Our approach allows for the highlighting of target areas, localized deletion, and movement. Additionally, by integrating with Stable-Diffusion (Rombach et al. 2021), we utilize the Score Distillation Sampling (SDS) (Poole et al. 2022) loss function to perform high-quality 3D generation and inpainting in specific regions. “A beautiful vase” is used as the prompt for the 3D inpainting process after locating the object.

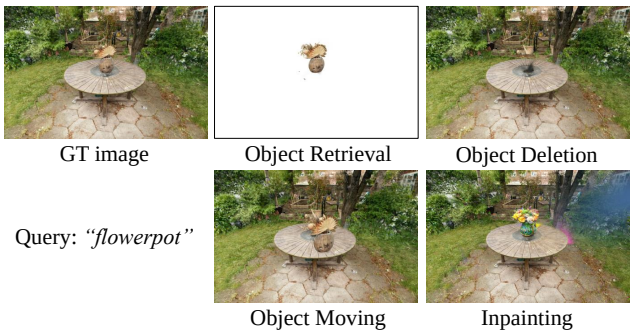

Figure 3: Visualization of scene manipulation results with OVGS. The query text is used to query the 3D object.

Conclusion

In this paper, we introduce OVGS, a novel method for reconstructing 3D semantic fields that enables precise open-vocabulary querying in 3D. By leveraging the APE encoder and a Semantic Codebook, OVGS efficiently compresses high-dimensional semantic features and integrates them into 3DGS with dense, low-dimensional representations. Compared to previous methods, OVGS significantly reduces memory and rendering costs while maintaining clear semantic distinctions between objects. Extensive experiments demonstrate that OVGS outperforms existing approaches, and it is highly applicable to downstream tasks such as localized scene editing.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 5460–5469. IEEE.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2023. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19697–19705.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, 333–350. Springer.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4): 1–14.

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. LERF: Language Embedded Radiance Fields. In *International Conference on Computer Vision (ICCV)*.

Kobayashi, S.; Matsumoto, E.; and Sitzmann, V. 2022. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35: 23311–23330.

Li, B.; Weinberger, K. Q.; Belongie, S. J.; Koltun, V.; and Ranftl, R. 2022. Language-driven Semantic Segmentation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1): 99–106.

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

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv preprint*, abs/2209.14988.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Reiser, C.; Szeliski, R.; Verbin, D.; Srinivasan, P.; Mildenhall, B.; Geiger, A.; Barron, J.; and Hedman, P. 2023. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4): 1–12.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Shen, Y.; Fu, C.; Chen, P.; Zhang, M.; Li, K.; Sun, X.; Wu, Y.; Lin, S.; and Ji, R. 2023. Aligning and Prompting Everything All at Once for Universal Visual Perception. *ArXiv preprint*, abs/2312.02153.

Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. J. 2021. 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*, 15818–15827. IEEE.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2023. Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields. *ArXiv preprint*, abs/2312.03203.