Enhanced monocular 3D Object Reconstruction

Adam Deryło Małgorzata Gwiazda Emin Sadikhov
Technical University of Munich
{adam.derylo, malgorzata.gwiazda, emin.sadikhov}@tum.de

Abstract

Splatter Image, a recent approach for monocular 3D object reconstruction, achieves high efficiency using Gaussian splatting while maintaining state-of-the-art performance. In this work, we propose enhancements to this project through three contributions: (1) incorporating semantic embeddings from pre-trained vision-language models to provide richer contextual understanding, (2) integrating monocular depth estimation to improve geometric accuracy, and (3) enhancing loss calculations by using Total Variation and Edge losses to refine reconstruction details. Our experiments show that semantic conditioning, particularly using DINO embeddings, significantly improves view consistency and generalization. Depth information further enhances reconstruction quality by constraining the solution space, but loss modifications do not bring substantially improvements. Code is available at https://github.com/splatter-works/splatter-image.

1. Introduction

Single-view 3D reconstruction remains a challenging problem in computer vision, particularly due to its inherently ill-posed nature [8, 25]. While recent advances in neural rendering and 3D representation have shown promising results [30], achieving both high-quality reconstruction and real-time performance has remained elusive. The recently proposed Splatter-Image method [22] made significant strides in this direction by introducing an ultra-fast approach to single-view 3D reconstruction using Gaussian splatting [11]. However, despite its impressive speed and state-of-the-art performance, the method shares certain limitations in reconstruction quality with its NeRF-based predecessors, particularly in achieving high-fidelity object reconstructions [29].

Building on observations from Tatarchenko et al. [25], we hypothesize these limitations stem from an underconstrained optimization problem. The underlying network is inherently underparameterized relative to the degrees of freedom possible in plausible reconstructions, primarily due to the inherent ambiguity of the task [8]. This results in a form of representational "superposition" [7], where the network attempts to encode more features than the available dimensions in its hidden layers permit, leading to interference between these representations. To resolve this interference, the network learns non-linear transformations that effectively serve as shortcuts.

Rather than truly learning reconstruction, these networks often rely on classification and retrieval, as shown by Tatarchenko et al. [25]. Monocular object reconstruction methods like Splatter Image tend to prioritize high PSNR and SSIM over true fidelity, similar to NeRF-based approaches like PixelNeRF [30], as reported by Watson et al. [29]. This often results in "blobby" reconstructions that average training shapes rather than capturing fine-scale details. The need for depth conditioning in Flash3D [23] further suggests that substantial modifications are required for scene reconstruction.

To improve Splatter Image while maintaining its state-of-the-art performance, we explore three complementary enhancements:

1. **Semantic Enhancement:** We leverage pre-trained foundation models [16, 18] with strong semantic understanding capabilities to provide additional contextual information to the reconstruction process. We hypothesize that by conditioning the network on semantic embeddings from pre-trained models, we free the rest of the encoder to focus on local features rather than redundantly capturing semantic information, thus boosting overall performance metrics.
2. **Depth Integration:** We investigate the integration of monocular depth estimation [20, 34] as an auxiliary signal, hypothesizing that explicit depth information can help constrain the solution space and improve geometric accuracy. This builds upon previous work demonstrating the synergistic effects of combining geometric and color information in 3D reconstruction tasks [5, 32]. Furthermore, the validity of this extension is reinforced by a similar approach in the follow-up scene reconstruction work, Flash3D [24].
3. **Loss Schedule Optimization:** We explore optimized

loss function schedules to guide training toward detailed and accurate reconstructions. By combining Total Variation (TV) loss and edge-preservation loss with the standard photometric supervision, we prevent convergence to geometrically inconsistent solutions while maintaining sharp features. Our approach adapts the curriculum learning benefits demonstrated by Chen et al. [2] within the end-to-end training paradigm of the original Splatter-Image framework.

Our experiments demonstrate that these enhancements show promising improvements in reconstruction quality while preserving the computational advantages of the original Splatter-Image framework. Despite constraints in computational resources, our preliminary results suggest that depth integration and semantic embeddings are promising research directions for addressing the inherent ambiguity in monocular 3D object reconstruction.

2. Related Work

2.1. Semantic Conditioning in 3D Reconstruction

Incorporating semantic cues has been shown to improve 3D reconstruction and scene understanding. Several methods use pre-trained vision-language models to enhance semantic consistency and generalization. For example, prior work employs CLIP-based losses to refine occluded surfaces and leverage 2D DINO features for automated object discovery in SfM point clouds [10, 28]. Other approaches transfer dense 2D CLIP features to 3D scene representations, enabling annotation-free segmentation, while vision-language embeddings have been integrated into Gaussian Splatting to improve multi-view semantic coherence [13, 33, 36]. Furthermore, aligning 3D features with large-scale language models has proven to be effective for large-vocabulary segmentation [21]. Inspired by these works, we integrate both CLIP and DINO features as semantic conditioning signals, to improve the reconstruction quality to achieve better geometry and appearance.

2.2. Monocular Depth Estimation

Monocular depth estimation has seen remarkable progress with deep learning approaches [6, 12], with recent methods leveraging transformer architectures and multi-scale feature fusion to achieve state-of-the-art performance [1, 20, 35]. These approaches provide valuable geometric cues for 3D reconstruction tasks. Several works have integrated monocular depth as an auxiliary signal for novel view synthesis [14, 32], with MonoSDF [32] incorporating depth priors to guide SDF optimization and Depth-supervised NeRF [5] showing significant improvements in novel view synthesis quality, especially in regions with limited visibility.

Flash3D [24] extends this approach by generating multiple layers of 3D Gaussians to model both visible surfaces and occluded regions. Ablation studies show that depth priors improve reconstruction quality by guiding models toward geometrically consistent solutions. This also enables better generalization, as it reduces reliance on limited 3D training data by leveraging pre-trained depth models. In our work, we treat Splatter Image as an extension of depth prediction networks and integrate explicit depth information as a conditioning signal to improve geometric accuracy.

2.3. Loss Functions in 3D Reconstruction

The choice of loss functions is crucial for 3D reconstruction quality. Traditional methods use L1/L2 losses on point clouds, voxels, or meshes [3, 9, 27]. However, recent works show that perceptual and adversarial losses on 2D renderings can produce more visually appealing results [4, 15]. SPSG [4] introduces a self-supervised framework that infers unobserved geometry and color in RGB-D scans. Instead of 3D losses, it applies adversarial and perceptual losses on 2D renderings, reducing artifacts from inconsistent camera poses. Similarly, PixelNeRF [30] and other NeRF-based methods optimize photometric quality, sometimes sacrificing geometric accuracy. Recent work explores curriculum learning in 3D reconstruction. Chen et al. [2] propose a gradual transition from 3D to 2D supervision, addressing shape-appearance ambiguity by ensuring the model first learns object shape before refining appearance details.

3. Method

3.1. Semantic Conditioning

The original method for monocular 3D reconstruction maps RGB image pixels directly to colored Gaussians, relying solely on a CNN-based architecture to infer occluded regions. This requires the model to learn view-invariant features, which is challenging without extensive training data and for few-view 3D reconstruction. To address this, we integrate semantic embeddings from large vision-language models such as CLIP [19] and DINO [17], which have been shown to learn robust domain-invariant representations, even in zero-shot settings.

To incorporate these features, we modify the U-Net backbone by injecting pre-trained CLIP and DINO embeddings into the bottleneck layer. This allows the network to leverage high-level semantic features alongside spatial and appearance cues during reconstruction. By adding these embeddings, the model focuses more on object shape and structure rather than just pixel details, leading to improved inference of unseen object parts and more consistent reconstructions across different viewpoints. As a result, the model relies less on large labeled datasets and captures view-invariant features, enhancing monocular 3D object reconstruction.

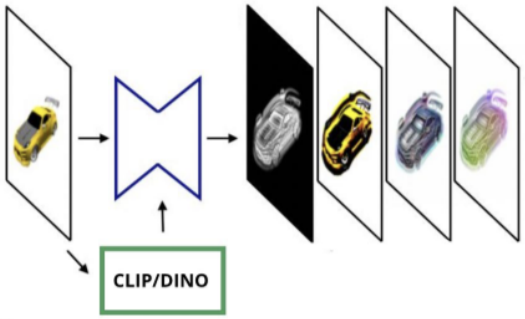

Figure 1. Semantic conditioning of image. CLIP and DINO embeddings are extracted from the input image and injected into the bottleneck layer of the U-Net.

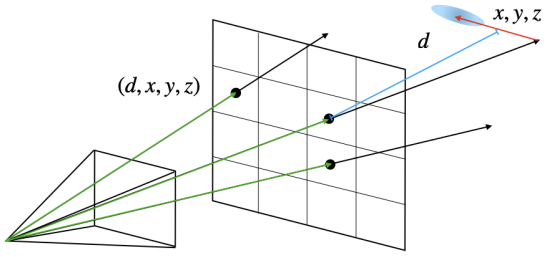

Figure 2. Illustration of the splatter image method as an extension of depth estimation networks. Each Gaussian's location is defined by a predicted depth d (blue) and a 3D offset (x, y, z) (red). The Gaussians are projected to the image plane along camera rays (green) and then displaced by the offset, capturing both observed and unobserved geometry.

3.2. Depth Conditioning

The standard Splatter-Image method operates solely on RGB images without depth information. Consequently, the network must implicitly estimate depth when predicting the placement of colored Gaussians, as illustrated in Figure 2. Due to the scarcity of 3D training data, this learned shape estimation is significantly less robust than dedicated depth estimation methods trained on abundant RGB-D datasets.

To enhance reconstruction robustness and guide the network to leverage domain-invariant features rather than defaulting to classification-based retrieval [25], we integrate depth information in two complementary approaches. First, we augment the input with an additional channel containing depth estimates from a pre-trained monocular depth estimation network. Second, we condition the bottleneck layer through cross-attention between image features F and depth features D obtained from the pre-trained monocular depth prediction network:

$$\text{Attention}(F, D) = \text{softmax}\left(\frac{FW_Q(DW_K)^T}{\sqrt{d_k}}\right)DW_V$$

where W_Q, W_K, and W_V are learnable convolutional projection layers that operate on the image features and depth map, respectively, and d_k is the dimension of the key vectors. This mechanism enables the network to selectively focus on depth map features during reconstruction and thus achieve better geometric consistency.

3.3. Loss Modification

In order to enhance the quality of images that the model outputs, we examined possible extensions in terms of loss functions. We proposed the addition of Total Variation loss to improve the smoothness of the image by penalizing high-frequency variations such as noise or abrupt intensity changes. It is commonly used in image denoising tasks [26]. Mathematically, it measures changes in pixel intensities along horizontal and vertical directions.

Secondly, we added Edge loss, which preserves edges by penalizing changes in gradient magnitude. It uses a Sobel operator to highlight intensity variations, aiming to improve sharpness and structural details. Loss terms were scaled, with 0.01 found to be optimal for both.

4. Results

4.1. Baseline Performance

We first evaluate the baseline performance of the Splatter Image model and compare it with PixelNeRF [31]. Since all modifications were trained with 30k iterations, we also trained the baseline Splatter Image model for 30k iterations to ensure consistency in our comparisons while considering time constraints. Table 1 presents the reconstruction quality metrics for these models at different training iterations.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
PixelNeRF	21.76	0.78	0.203
Splatter Image (800k)	**21.80**	**0.80**	**0.150**
Splatter Image (30k)	20.95	0.773	0.240

Table 1. Baseline novel-view performance comparison

The results show that the baseline Splatter Image model performs better than PixelNeRF, achieving state-of-the-art results for monocular 3D reconstruction. However, training for longer improves performance, as the model trained for 800k iterations achieves better results than the 30k iteration setup used in our experiments.

4.2. Effect of Semantic Embeddings

To evaluate the impact of adding semantic embeddings, we compare our modified models using CLIP-based and DINO-based features. Table 2 shows the performance on both conditioned and novel views.

Method	Conditioned Views			Novel Views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DINO-based	**32.21**	**0.961**	**0.072**	**21.26**	**0.786**	**0.221**
CLIP-based	31.72	0.959	0.077	20.91	0.775	0.230

Table 2. Comparison of CLIP-based and DINO-based embeddings on conditioned and novel views.

The results show that using DINO-based embeddings improves reconstruction quality, especially for novel views, while CLIP-based embeddings do not lead to significant gains. We believe this is due to how these models process visual features. CLIP, trained with contrastive learning on image-caption pairs, creates more global embeddings, whereas DINO, trained through self-distillation, captures more local, object-invariant features. These local features seem more useful for monocular 3D reconstruction as they help the model better predict unseen object parts. Additionally, DINO embeddings improve generalization to novel viewpoints, leading to better view-invariant feature learning. This reduces reliance on large labeled datasets and results in more consistent 3D reconstructions.

4.3. Effect of Depth Embeddings

Method	Depth Model	Conditioned Views			Novel Views		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Input	$\text{MiDaS}_{\text{SMALL}}$	30.17	0.938	0.111	20.76	0.776	0.230
	$\text{DPT}_{\text{HYBRID}}$	28.77	0.929	0.126	20.23	0.766	0.250
	$\text{DPT}_{\text{LARGE}}$	32.36	0.962	0.083	21.12	0.789	0.200
X-Attn	$\text{MiDaS}_{\text{SMALL}}$	**32.77**	**0.966**	0.060	21.11	0.787	**0.206**
	$\text{DPT}_{\text{LARGE}}$	32.71	**0.966**	**0.059**	**21.15**	**0.788**	**0.206**

Table 3. Comparison of depth integration methods using different depth models on conditioned and novel views.

We evaluate the impact of depth embeddings by comparing different ways of integrating monocular depth information into the reconstruction pipeline. Table 3 presents results for both direct depth input and cross-attention conditioning, using depth predictions from different MiDaS and DPT models.

The results show that the choice of depth integration method does not make a big difference, but the quality of the depth estimates plays a key role. Models using $\text{DPT}_{\text{LARGE}}$ consistently achieve better PSNR and SSIM while lowering LPIPS, especially for novel views. This suggests that higher-quality depth estimates help the model better understand object geometry.

Cross-attention conditioning leads to the best overall performance, with both $\text{DPT}_{\text{LARGE}}$ and $\text{MiDaS}_{\text{SMALL}}$ performing better than the baseline. The improvements are more noticeable in novel views, showing that strong depth cues help the model generalize beyond conditioned perspectives.

Our results show that depth information improves reconstruction quality, but its effectiveness depends on the accuracy of depth predictions.

4.4. Effect of Loss Modification

Table 4 compares results achieved by the two scenarios of baseline with Total Variation loss and Edge loss terms added to main loss.

Method	Conditioned Views			Novel Views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Var loss	32.31	**0.960**	**0.079**	**21.19**	**0.773**	**0.243**
Edge loss	13.61	0.743	0.297	12.52	0.683	0.342

Table 4. Addition of Total Variational loss and Edge Loss compared on conditioned and novel views.

Incorporation of Total Variation loss gave the best results, but only marginally surpassing original work. Usage of Edge loss resulted in significantly reduced performance.

5. Conclusion

We explored three modifications to the Splatter Image framework: incorporating semantic embeddings, integrating depth information, and adjusting the loss function. Our results show that using DINO embeddings improves generalization to novel views, while CLIP embeddings do not, likely because DINO captures more local, object-invariant features. Adding depth information from MiDaS enhances reconstruction quality, with better depth estimates leading to better results. For loss modifications, Total Variation loss provided a slight improvement, while Edge loss significantly reduced performance, suggesting that preserving edges is not beneficial for Gaussian Splatting.

Future work could explore pre-training strategies to improve geometric understanding, similar to how pre-training on mathematical tasks enhances entity recognition in NLP. Additionally, techniques from generative modeling, such as mixture-of-experts architectures and autoregressive methods seen in recent image generation models, could further improve view consistency and generalization in monocular 3D reconstruction. Our results show the benefits of incorporating semantic and depth information while suggesting promising directions for future research.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 2

[2] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. 2023. 2

[3] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision – ECCV 2016*, pages 628–644, Cham, 2016. Springer International Publishing. 2

[4] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans, 2021. 2

[5] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free, 2024. 1, 2

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 2

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. 2022. 1

[8] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. 2016. 1

[9] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image, 2016. 2

[10] Yubin Hu, Sheng Ye, Wang Zhao, Matthieu Lin, Yuze He, Yu-Hui Wen, Ying He, and Yong-Jin Liu. O^2-recon: Completing 3d reconstruction of occluded objects in the scene with a pre-trained 2d diffusion model, 2024. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics, volume 42(4), July 2023*, 2023. 1

[12] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks, 2016. 2

[13] 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, 2024. 2

[14] Fanqing Lin, Brian Price, and Tony Martinez. Generalizing interactive backpropagating refinement for dense prediction, 2021. 2

[15] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3504–3515, 2020. 2

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. 2023. 1

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2

[18] 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. 2021. 1

[19] 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, 2021. 2

[20] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. 1, 2

[21] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild, 2022. 2

[22] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. 2023. 1

[23] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F. Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. 2024. 1

[24] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F. Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image, 2024. 1, 2

[25] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? 2019. 1, 3

[26] Mengyuan Wang, Wei He, and Hongyan Zhang. A spatial–spectral transformer network with total variation loss for hyperspectral image denoising. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 3

[27] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images, 2018. 2

[28] Yuang Wang, Xingyi He, Sida Peng, Haotong Lin, Hujun

Bao, and Xiaowei Zhou. Autorecon: Automated 3d object discovery and reconstruction. In *CVPR*, 2023. 2

[29] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. 2022. 1

[30] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. 2020. 1, 2

[31] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2021. 3

[32] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction, 2022. 1, 2

[33] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip, 2023. 2

[34] Zhengming Zhou and Qiulei Dong. Self-distilled feature aggregation for self-supervised monocular depth estimation, 2022. 1

[35] Zhengming Zhou and Qiulei Dong. Two-in-one depth: Bridging the gap between monocular and binocular self-supervised depth estimation, 2023. 2

[36] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding, 2024. 2