Reconstruction Using the Invisible: Intuition from NIR and Metadata for Enhanced 3D Gaussian Splatting

Gyusam Chang[1,2*], Tuan-Anh Vu[2], Vivek Alumootil[2], Harris Song[2], Deanna Pham[2], Sangpil Kim[1†], M. Khalid Jawed[2†]

[1]Korea University, Republic of Korea,
[2]University of California, Los Angeles, USA
{gsjang95, spk7}@korea.ac.kr
{tuananh.vu, vivekalumootil, songharris2006, deannapham2004, khalidjm}@ucla.edu

Abstract

While 3D Gaussian Splatting (3DGS) has rapidly advanced, its application in agriculture remains underexplored. Agricultural scenes present unique challenges for 3D reconstruction methods, particularly due to uneven illumination, occlusions, and a limited field of view. To address these limitations, we introduce NIRPlant, a novel multimodal dataset encompassing Near-Infrared (NIR) imagery, RGB imagery, textual metadata, Depth, and LiDAR data collected under varied indoor and outdoor lighting conditions. By integrating NIR data, our approach enhances robustness and provides crucial botanical insights that extend beyond the visible spectrum. Additionally, we leverage text-based metadata derived from vegetation indices, such as NDVI, NDWI, and the chlorophyll index, which significantly enriches the contextual understanding of complex agricultural environments. To fully exploit these modalities, we propose NIRSplat, an effective multimodal Gaussian splatting architecture employing a cross-attention mechanism combined with 3D point-based positional encoding, providing robust geometric priors. Comprehensive experiments demonstrate that NIRSplat outperforms existing landmark methods, including 3DGS, CoR-GS, and InstantSplat, highlighting its effectiveness in challenging agricultural scenarios.

Introduction

3D reconstruction has become increasingly crucial across various fields, including robotics, autonomous driving, augmented reality, and agricultural monitoring. Traditional methods for reconstructing three-dimensional structures from two-dimensional images often struggle to capture fine details, handle complex scenes, and maintain robustness under challenging environmental conditions. Recently, 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) has emerged as a significant advancement, enabling smoother, more detailed, and computationally efficient reconstructions. Unlike traditional approaches that rely heavily on discrete point representations (Sinha et al. 2017; Li et al. 2018; Lin, Kong, and Lucey 2018; Nguyen et al. 2019), 3DGS represents each 3D point as a Gaussian distribution, effectively capturing uncertainties and spatial continuity in complex environments.

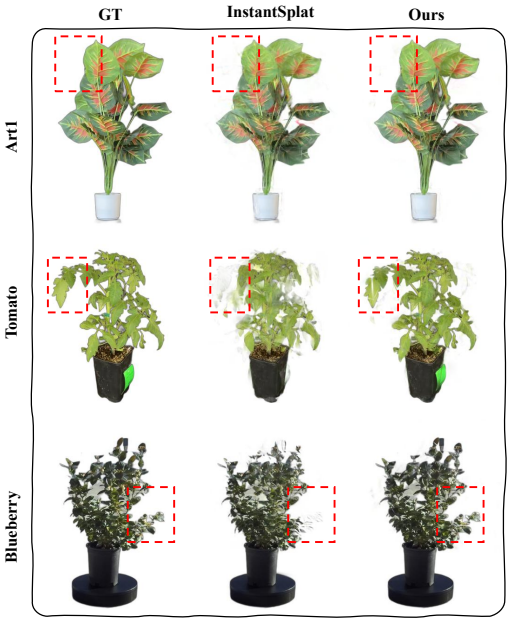


Figure 1: Qualitative comparisons in a 3-view setup. We highlight our improved semantic understanding, particularly in regions of interest within the image, where our method more accurately captures meaningful structures and distinctions.

Despite its demonstrated success in general scenarios, 3DGS faces significant challenges when applied to agriculture. As shown in Fig. 1, these environments pose unique challenges, including unpredictable lighting variations (*e.g.*, intense sunlight, low visibility, and sunset conditions), limited viewing angles, environmental instability due to weather fluctuations, and frequent occlusions by foliage. These factors can substantially degrade the performance of typical 3D reconstruction methods (Fan et al. 2024b; Zhang et al. 2024), resulting in incomplete or inaccurate plant modeling.

To overcome these limitations, we introduce the **NIRPlant** dataset, specifically designed to address the unique challenges of agricultural environments. **NIRPlant** incorporates comprehensive multimodal data, including RGB images, Near-Infrared (NIR) imagery, and rich textual metadata. The dataset (see Tab. 1) comprises diverse indoor and outdoor lighting scenarios captured from multiple perspectives, including artificial illumination, direct sunlight, and sunset con-

*Work done while the author was a visiting researcher at UCLA.
†Corresponding authors.

ditions. Integrating NIR imagery is particularly advantageous because NIR can capture plant-specific reflectance characteristics invisible to conventional RGB cameras, thus providing essential botanical information about plant health, water content, and structural integrity. For instance, high values of NDVI (Normalized Difference Vegetation Index) typically indicate robust vegetation health, NDWI (Normalized Difference Water Index) reflects water content, and chlorophyll index values directly correlate with photosynthetic efficiency and plant vigor. Such indices enrich our dataset and significantly enhance the model's ability to accurately interpret complex botanical scenarios, as shown in Fig. 1.

Moreover, textual metadata derived from both RGB and NIR images includes environmental conditions, precise lighting descriptions, and quantitative botanical indices, thus providing rich context for reconstructing detailed 3D models. Fusing this metadata with visual modalities enables our method to interpret plants more effectively and model them under diverse and challenging photometric conditions.

To leverage the full potential of our multimodal dataset, we propose **NIRSplat**, a novel Gaussian splatting framework optimized for multimodal data integration. **NIRSplat** employs a novel cross-attention mechanism (Zhu et al. 2020; Vaswani et al. 2017) that effectively combines NIR embeddings with RGB features. Our approach achieves superior scene understanding by exploiting the complementary strengths of RGB imagery and NIR-derived features. Moreover, inspired by the success of Vision-Language Models (VLMs) (Radford et al. 2021; Li et al. 2022a, 2023a; Liu et al. 2023), we integrate textual embeddings derived from metadata descriptions to further enhance semantic understanding. This multimodal interaction is further enhanced by employing a novel 3D point-based positional encoding method, which leverages spatial coherence from geometric priors to align and enrich 2D image features with 3D spatial information.

We conducted extensive evaluations to validate our proposed method, comparing **NIRSplat** with state-of-the-art approaches. Our results show that **NIRSplat** outperforms existing methods in terms of reconstruction accuracy, robustness to varying environmental conditions, and visual quality. We provide detailed analyses that highlight the contributions of each modality to the overall improvement in performance.

In summary, our key contributions include:

- The introduction of **NIRPlant**, a comprehensive multimodal agricultural dataset integrating RGB, NIR, and detailed textual metadata, enabling robust 3D reconstruction under varying lighting and environmental conditions.
- Development of **NIRSplat**, a multimodal Gaussian splatting framework employing cross-attention mechanisms and geometric priors, significantly improving scene reconstruction robustness.
- Extensive comparative analyses demonstrate our approach's effectiveness and advantages over leading methods, including 3DGS, CoR-GS, and InstantSplat, under diverse agricultural conditions (*e.g.*, intense sunlight, occlusion, low visibility).

Dataset	Modality	Lighting	# Scenes	# Views	Metadata
Barron et al. (2022)	R	✗	9	100-330	✗
Knapitsch et al. (2017)	R	✗	14	4-17	✗
Toschi et al. (2023)	R	✓	20	2000	✗
Voynov et al. (2023)	R,D,S	✗	107	100	✗
NIRPlant (Ours)	R,D,N,L,T	✓	34	360	✓

Table 1: Comparison with existing landmark dataset. R, D, N, S, L, and T denote RGB, Depth, NIR, Structured-Light Scanner (SLS), LiDAR, and Text, respectively. Lighting indicates whether there are various lighting conditions for supervision.

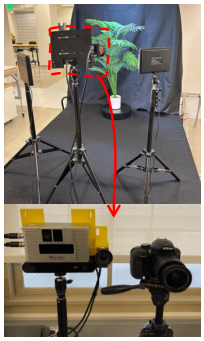
Related Works

Method for 3D Reconstruction

3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) uses a set of 3D Gaussian parameters and differentiable splatting to represent and render scenes more efficiently than traditional radiance fields (Martin-Brualla et al. 2021; Garbin et al. 2021; Barron et al. 2021). Mip-Splat (Yu et al. 2024) is another scene rendering algorithm that constrains the size of 3D Gaussian primitives and mitigates aliasing and dilation issues present in 3DGS. CF-3DGS (Fu et al. 2024) reduces the burden of pre-computation by leveraging the temporal continuity from video and the explicit point cloud representation. CoR-GS (Zhang et al. 2024) identifies and suppresses inaccurate reconstruction using Co-pruning, considers Gaussians, and Pseudo-view co-regularization. Furthermore, another InstantSplat (Fan et al. 2024b) is compatible with both the above methods and specializes in low-image-count representations through a neural network representation similar to that of NSR, utilizing a Gaussian Bundle Adjustment (GauBA). SplatFields (Mihajlovic et al. 2024) designs and regularizes splat features as the outputs of a corresponding implicit neural field. Recently, CATSplat (Roh et al. 2024) introduced a generalizable transformer-based framework, addressing the inherent constraints in monocular settings.

Dataset for 3D Reconstruction

Various 3D Reconstruction datasets have focused on advancements in lighting recognition and multimodal approaches. Mip-NeRF360 (Barron et al. 2022) synthesizes realistic object views in the real world, while MVimgNet enhances 3D capture through video-based 3D-aware signals (Yu et al. 2023). Adding on, ReLight and Tanks are designed to address lighting variation with different materials (Toschi et al. 2023; Voynov et al. 2023; Knapitsch et al. 2017). Previous adaptations of NeRF to real-world environments through LEGO bricks (Li et al. 2023b) and famous city sites (Martin-Brualla et al. 2021) improve lighting capture by training from photo datasets. OmniObject3D addresses surface reconstruction for dense and sparse-view surfaces (Wu et al. 2023). Additionally, GauU-Scene (Xiong, Li, and Li 2024) supports large-scale scene reconstruction using Gaussian Splatting for real-time scanning. NeRFBK (Yan et al. 2023) utilizes both real and synthetic data to capture objects of varying materials and lighting conditions, thereby comparing NeRF in outdoor views and for transparent objects. UniSDF (Wang et al. 2024) is another dataset that utilizes NeRF to capture 3D scenes with reflections, combining the traditional SDF with radiance fields to render scenes with and without reflections.



Sensor configuration:
- ✓ **Primary RGB & Depth**: ZED 2i,
- ✓ **NIR** : Alvium 1800 U-501 NIR,
- ✓ **Point cloud & calibration**: Neuvition Titan S2-70C
- ✓ **High-Resolution RGB** : Nikon D3400

Data Collection Procedure:
1. Installation: Set up the equipment at the scheduled time with established environmental conditions.
2. 360° Plant Rotation: Place the plant on the turntable and rotate it through a full 360-degree cycle.
3. Multi-View Image Capture: Capture images from the predetermined 36 viewpoints for full coverage.
4. Sensor Level Adjustment: Modify the sensor height and repeat steps 2 and 3.

(a) (b)

Figure 2: (a) Data Acquisition Platform: Sensor (Top) installation, and (Bottom) configuration. (b) Data collection procedure for both indoor and outdoor settings.

Our Multimodal NIRPlant Dataset

Collecting comprehensive multimodal datasets in diverse environmental conditions is inherently challenging, even within controlled laboratory settings. The manual process involved in collecting, observing, and annotating multiple data types, such as RGB and NIR imagery, and metadata, is labor-intensive, time-consuming, and costly. Moreover, achieving diversity in visual data and ensuring high-quality annotations significantly complicates the process. Please refer to the supplementary material for additional details.

Data Acquisition Platform

Our primary objective is to develop a comprehensive and versatile 3D plant reconstruction dataset under diverse real-world lighting conditions. However, dynamic environmental factors such as wind, shadows, and fluctuating sunlight pose significant challenges to data reliability. To mitigate these effects, we designed a controlled lighting configuration (see Fig. 2) that captures data during critical illumination periods, allowing us to observe lighting variance systematically. Precisely, objects were positioned to capture precise lighting conditions at defined times of the day (*i.e.*, noon and sunset). Furthermore, we utilized a multimodal sensor setup consisting of a ZED 2i and Nikon D3400 HD RGB cameras, an Alvium 1800 U-501 NIR sensor, and a Neuvition Titan S2-70C LiDAR sensor, ensuring accurate alignment and consistent distance between the objects and sensors. To enhance data quality under natural conditions, automatic adjustments for focus, exposure, and gain were employed to maintain consistency and adaptivity in capturing agricultural data, thus enabling accurate calibration and precise extraction of camera perspectives essential for reliable 3D reconstruction.

Data Construction and Processing

Collecting precise camera poses in agricultural environments is inherently challenging due to the dynamic nature of plants, whose structures change rapidly in response to environmental factors. To overcome this issue, extensive data were collected indoors and outdoors under strictly consistent conditions. Specifically, each object was captured from 360 multi-modal data samples per scene, and ground truth models were constructed using the landmark Structure-from-Motion (SfM) technique (Schönberger and Frahm 2016), ensuring high accuracy and robustness, particularly for texture-rich environments. Additionally, precise reconstruction was prioritized by meticulously removing background information from object images. In total, our dataset contains comprehensive multimodal data (as described in Tab. 1), covering four distinct lighting scenarios with 360 viewpoints per scene across up to 10 different plant categories, ensuring broad diversity and representativeness (see Fig. 3).

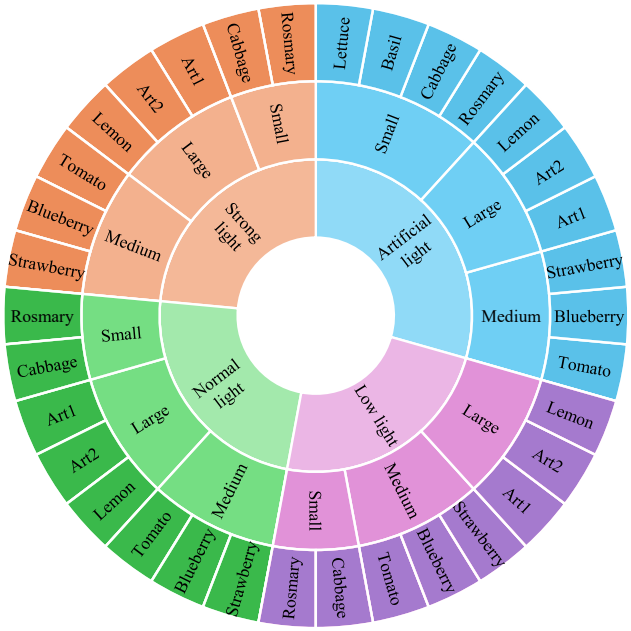


Figure 3: Hierarchical organization of the dataset taxonomy.

Dataset Specifications and Statistics

Data Organization. As illustrated in Fig. 3, our **NIRPlant** dataset categorizes plant data under four primary lighting conditions: artificial light, strong sunlight, low light, and normal daylight. It encompasses up to 10 diverse plant species, further classified into three distinct size categories (small, medium, and large). Each plant category and lighting condition was captured consistently from 360 viewpoints to ensure robust perspective coverage. This systematic data acquisition process was uniformly applied across RGB, NIR, Depth, and LiDAR. Note that botanic-aware prompts are generated for each scene, as illustrated in the supplementary material. Additionally, to effectively extract discriminative NIR signals, comparisons were conducted against artificial plants (Art 1 and Art 2). Leveraging this comprehensive multimodal dataset structure, we aim to enhance the understanding of plant-specific characteristics under various environments, thus substantially improving the performance and robustness of 3D reconstruction methods.

Dataset Split. Considering the practical agricultural environment, we adopted a sparse-view approach inspired by InstantSplat (Fan et al. 2024a). Specifically, from each set of 24 RGB, NIR, and textual metadata viewpoints, we randomly sampled 3, 6, and 12 views for training and testing

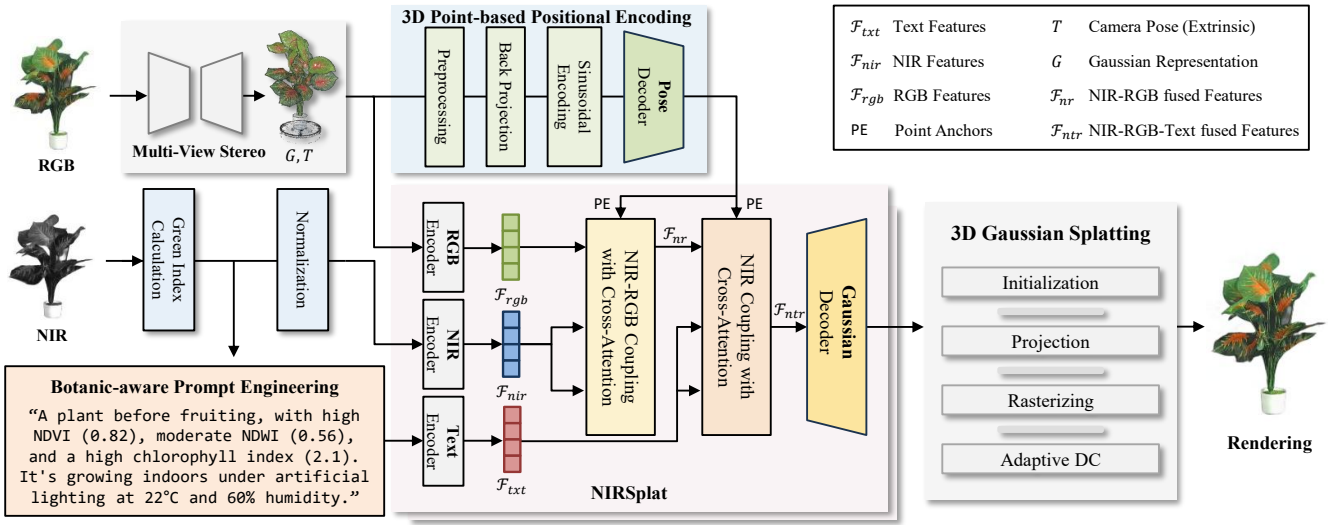


Figure 4: The overall architecture of NIRSplat framework. NIRSplat efficiently processes tri-modal inputs consisting of NIR, RGB, and Text, enabling joint reasoning. The details of the prompt engineering are included in the supplementary document.

purposes, respectively. This sampling strategy simulates real-world scenarios where limited views are common, ensuring the dataset's applicability and the generalization of reconstruction algorithms in realistic agricultural contexts.

Our Proposed Method

In this section, we describe our NIRSplat in detail, emphasizing the multimodal initialization through 3D positional encoding, the Transformer-based interactions for modality fusion, and the multimodal loss and regularization mechanisms. The technical background necessary for understanding our NIRSplat is provided in the supplementary material.

Gaussian-guided Positional Anchoring

Recently, the explicit way (Godard, Mac Aodha, and Brostow 2017; Godard et al. 2019; Yang et al. 2024) to interact with the 3D priors is to estimate depths from input RGB images. However, such an approach profoundly limits the advantage of 3DGS (*i.e.*, real-time NVS) by demanding additional deep-learning capacity. To mitigate this, we propose a lightweight and efficient alternative: **Gaussian-guided Positional Anchoring** inspired by (Shu et al. 2023; Liu et al. 2022), which provides strong geometric clues from initialized Gaussian positions without requiring external depth supervision.

Positional Anchoring leveraging Gaussian means. We leverage MASt3R (Leroy, Cabon, and Revaud 2024) to predict an initial dense 3D point map $\{\mathbf{p}_i\}_{i=1}^{N}$, which serves as the initialization for our Gaussian representation $\mathrm{G} = \{\mu_i, \Sigma_i, \alpha_i, c_i\}$, where $\mu_i = \mathbf{p}_i$ denotes the Gaussian center. Simultaneously, a coarse camera pose matrix $\mathrm{T} \in \mathrm{SE}(3)$ $(R \in SO(3), t \in \mathbb{R}^3)$ is obtained per view v, also from MASt3R. Given the current estimate of camera pose T and the intrinsic matrix $\mathrm{K} \in \mathbb{R}^{3\times3}$, we project each 3D point $\mathbf{p}_i \in \mathbb{R}^3$ onto the 2D image plane as below:

$$\tilde{\mathbf{u}}_i = \mathrm{K} \cdot (R \cdot \mathbf{p}_i + \mathbf{t}) \in \mathbb{R}^3, \tag{1}$$

$$\mathbf{u}_i = \left[\frac{\tilde{u}_i^x}{\tilde{u}_i^z}, \frac{\tilde{u}_i^y}{\tilde{u}_i^z}\right] \in \mathbb{R}^2. \tag{2}$$

Each projected 2D location $\mathbf{u}_i$ serves as a spatial anchor, from which we derive a positional embedding using either sinusoidal encoding or a lightweight MLP.

$$\mathtt{PE}_i = \Phi\left(\left[\sin(\lambda^\top \mathbf{u}_i) \oplus \cos(\lambda^\top \mathbf{u}_i)\right]\right), \tag{3}$$

where λ represents a learnable frequency scale and $\oplus$ denotes concatenation. $\Phi(\cdot)$ is a multilayer perceptron (MLP) that maps the encoded coordinates to a latent embedding space. This formulation enables efficient and geometry-aware interaction with 3D points directly on the image plane by providing a *unified positional reference*. Crucially, it preserves spatial correspondence and depth continuity without incurring the cost of full-scale depth estimation, thus maintaining the efficiency and lightweight design of 3DGS.

NIRSplat: A Multimodal Gaussian Splatting

Bridging the invisible: NIR-RGB Coupling. 3DGS (Kerbl et al. 2023) introduces a powerful Gaussian-based representation that enables real-time novel view synthesis, driving substantial progress across numerous 3D vision applications. However, in outdoor agricultural scenarios, where sensor configurations are often sparse and viewpoint coverage is inherently limited, we observe a considerable performance drop due to inconsistent lighting, occlusion, and textureless surfaces. To overcome these challenges, we incorporate **near-infrared (NIR)** sensing as a complementary modality to RGB. NIR images capture electromagnetic wavelengths beyond the visible spectrum, revealing latent structural information such as chlorophyll absorption, leaf water content, and surface reflectance properties that are often invisible in RGB. By leveraging this spectral prior, we aim to enhance feature

robustness under adverse imaging conditions. To this end, we design a Transformer-based NIR-RGB fusion module using a deformable cross-attention mechanism D_Attn (Vaswani et al. 2017; Zhu et al. 2020). Let $\mathbf{F}_{rgb} = \{f_{rgb}^i\}_{i=1}^N$ and $\mathbf{F}_{nir} = \{f_{nir}^i\}_{i=1}^N$ be the extracted features from RGB and NIR branches. Each modality is augmented with a shared positional encoding $\texttt{PE}_i = \texttt{PE}[:, u_i, v_i]$, and fused via:

$$\mathbf{F}_{nr}^{(i)} = D\_Attn(f_{rgb}^{(i)} \oplus \texttt{PE}_i,\ f_{nir}^{(i)} \oplus \texttt{PE}_i,\ f_{nir}^{(i)} \oplus \texttt{PE}_i) \quad (4)$$

Here, $D\_Attn(\cdot)$ applies a multi-head deformable attention operation. This formulation allows the RGB features to selectively attend to informative NIR signals guided by spatial anchors from the shared positional encoding. The fused representation is obtained by stacking L attention layers, yielding the final robust cross-modal feature set: $\mathbf{F}_{nr} = \left\{\mathbf{f}_{nr}^i\right\}_{i=1}^N$, where $\mathbf{F}_{nr}^{(i)} \in \mathbb{R}^C$. This transformer-driven NIR-RGB interaction enables effective exploitation of cross-spectral cues under limited views, leveraging both radiometric contrast from NIR and geometric alignment via positional encoding. As demonstrated in our experiments, this mechanism significantly enhances scene understanding and 3D reconstruction quality under real-world agricultural constraints.

Bridging the invisible: RGB-Text Coupling. Vision-Language Models (VLMs) (Radford et al. 2021; Alayrac et al. 2022; Li et al. 2022a,b; Zhang, Li, and Bing 2023; Li et al. 2023a; Liu et al. 2023) have recently achieved striking success across a wide range of tasks by tightly coupling visual inputs with rich textual descriptions. Despite their proven potential, these models remain largely unexplored in the domain of agricultural 3D reconstruction (*i.e.*, a field that urgently demands robust, high-level scene understanding to support smart farming systems). To address this gap, we propose a Transformer-based *RGB-Text interaction module* that semantically bridges RGB features with language-derived plant attributes, enabling better recognition of hard samples (*e.g.*, small objects, fine structures, and hard-to-perceive regions). Primarily, we observe that various factors (*e.g.*, descriptions, object attributes, environmental cues) in text prompts significantly contribute to view understanding (Oh et al. 2024; In Lee et al. 2024; Roh et al. 2024; Lee et al. 2024) by guiding the model's attention and perspectives. Inspired by this, we generate botanical-aware prompts $\mathcal{T} \in \mathbb{R}^{L \times C}$ that encapsulate detailed semantic information, such as vegetation index (*e.g.*, NDVI, NDWI), structural traits (*e.g.*, leaf shape, stem thickness), phenological stages (*e.g.*, sprouting, flowering), and context (*e.g.*, lighting, occlusion), as detailed in the supplementary document. These prompts are first encoded using a pre-trained VLM to obtain token-level text features: $\mathbf{F}_{txt} = \{f_{txt}^{(1)}, f_{txt}^{(2)}, ..., f_{txt}^{(L)}\}$, where each $f_{txt}^{(i)} \in \mathbb{R}^C$ represents a contextualized embedding. To align language and vision, we also leverage the *deformable attention* mechanism D_Attn (Vaswani et al. 2017; Zhu et al. 2020), injecting shared positional priors via $\texttt{PE}$ (see Eq. 3) in the same manner. We formulate the multimodal features from the NIR-RGB fusion $\mathbf{F}_{nr}^{(i)}$ and textual tokens $f_{txt}^{(i)}$ at pixel coordinate (u_i, v_i) as follows:

$$\mathbf{F}_{ntr}^{(i)} = D\_Attn(f_{nr}^{(i)} \oplus \texttt{PE}_i,\ f_{txt}^{(i)} \oplus \texttt{PE}_i,\ f_{txt}^{(i)} \oplus \texttt{PE}_i) \quad (5)$$

Here, $\oplus$ denotes vector concatenation, and the attention module facilitates fine-grained alignment between visual and linguistic features at both spatial and semantic levels. The resulting multimodal feature $\mathbf{F}_{\text{ntr}}$ integrates geometric anchors and contextual cues and is subsequently processed by a lightweight feed-forward decoder: $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{c}\} = \text{MLP}_{\text{gauss}}(\mathbf{F}_{ntr})$, where $\boldsymbol{\mu} \in \mathbb{R}^3$ denotes the 3D Gaussian mean, $\boldsymbol{\alpha}$ is the opacity, $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$ represents the anisotropic covariance (or its low-rank approximation), and $\boldsymbol{c}$ is the RGB appearance feature. These learned G are directly fed into a 3D Gaussian Splatting renderer equipped with an Adaptive Density Control (ADC) mechanism, allowing efficient, robust, and botanic-aware scene reconstruction under novel agricultural scenarios (*i.e.*, insufficient visual clues).

Cross-Modal Gaussian Field Reasoning

To successfully render a set of Gaussians G and corresponding pose T, we adopt a Gaussian rasterization as a differentiable operator following Gaussian Bundle Adjustment (Fan et al. 2024a) in a self-supervised manner. Specifically, after a highly informative cross-modal fusion phase, the refined pose $\mathbf{T}$ and Gaussian field $\mathbf{G}$ are jointly optimized by minimizing the photometric rendering loss:

$$\mathcal{L}_{\text{photo}} = \min_{G,T} \sum_v \sum_p \left\| \tilde{C}_{v,p} - C_{v,p}(G,T) \right\|_q, \quad q \in \{1,2\}. \quad (6)$$

where C and $\tilde{C}$ are the rasterization function and the observed 2D images, respectively. Consequently, it is worth noting that this formulation facilitates rapid optimization, seamlessly incorporating complementary multimodal knowledge into the underlying 3D Gaussian representation.

Experiments

Baselines. We selected recent state-of-the-art methods for comparison, including 3DGS (Kerbl et al. 2023), SplatFields (Mihajlovic et al. 2024), InstantSplat (Fan et al. 2024b) and CoR-GS (Zhang et al. 2024). These methods efficiently leverage a Gaussian parameter in real time, optimizing a position $\boldsymbol{\mu}$, an opacity $\boldsymbol{\alpha}$, a covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, and spherical harmonics (color) c with trivial computational overhead. Additionally, we adopt pose-free methods, Nope-NeRF (Bian et al. 2023) and CF-3DGS (Fu et al. 2024), which are supported by monocular depth maps and ground-truth camera intrinsics, following InstantSplat. Please refer to the supplementary document for detailed experimental setups.

Benchmarks and Discussions

Previous works often struggle under agricultural conditions due to uncertain camera perspectives, insufficient visual cues, and constraints on computational resources, making robust 3D reconstruction particularly challenging. Specifically, 3DGS suffers from limited 2D information (*i.e.*, due to sparse-view setups), leading to an inevitable performance drop (up to -31.9% SSIM, -5.8 PSNR, and +25.2% LPIPS gaps compared to our **NIRSplat**), as shown in Tab. 2. In the 3-view configuration, CoR-GS shows substantial structural degradation (lowest SSIM score) among recent sparse-view

Method	SSIM (↑)			PSNR (↑)			LPIPS (↓)		
	3-view	6-view	12-view	3-view	6-view	12-view	3-view	6-view	12-view
3DGS	0.5074	0.5590	0.6531	14.1552	15.5586	17.4352	0.4586	0.4469	0.4033
CoR-GS-1k	0.7179	0.7642	0.8081	16.3494	17.1991	19.5895	0.4124	0.3191	0.2289
CoR-GS-10k	0.7285	0.7776	0.8287	16.7118	18.8023	20.8714	0.4049	0.3094	0.2281
CoR-GS-30k	0.7143	0.7611	0.8131	15.8925	17.5348	20.2927	0.4120	0.3405	0.2489
SplatFields-1k	0.7429	0.7647	0.7886	11.5490	13.6087	13.4497	0.4301	0.3787	0.3164
SplatFields-10k	0.7624	0.7799	0.8070	12.1196	14.6183	14.4463	0.3965	0.4017	0.2898
SplatFields-30k	0.7664	0.7802	0.8163	12.8751	14.1314	15.6037	0.3764	0.3754	0.2721
InstantSplat-200	0.7559	0.7604	0.7720	17.6177	17.9250	18.5293	0.3048	0.2943	0.2784
InstantSplat-1k	0.7984	0.8126	0.8134	18.3849	18.9233	19.0333	0.2797	0.2689	0.2438
NIRSplat-200	0.7906	0.8099	0.8174	18.1747	18.7103	19.1921	0.2371	0.2267	0.2229
NIRSplat-1k	**0.8268**	**0.8311**	**0.8421**	**20.7182**	**21.0169**	**21.0814**	**0.2070**	**0.2071**	**0.2080**

Table 2: Main Performance with SOTA techniques (Fan et al. 2024b; Zhang et al. 2024; Mihajlovic et al. 2024) on NIRPlant dataset. We conduct experiments with 3, 6, and 12 view setups and calculate traditional three metrics: SSIM, PSNR, and LPIPS. 200, 1k, 10k and 30k denote iterations. Note that bold values indicate the best performance. Gray shading indicates Ours.

Method	Configuration	SSIM (↑)			PSNR (↑)			LPIPS (↓)		
		3-view	6-view	12-view	3-view	6-view	12-view	3-view	6-view	12-view
InstantSplat-S	I_{rgb} only	0.7984	0.8126	0.8134	18.3849	18.9233	19.0333	0.2797	0.2689	0.2438
	I_{rgb}, I_{nir}	0.7096	0.7383	0.7426	16.9913	17.6154	17.6345	0.2431	0.2265	0.2264
	$F_{rgb} \oplus F_{nir}$	0.8049	0.8079	0.8128	18.0318	18.7785	19.0927	0.3091	0.2923	0.2881
	$F_{rgb} \oplus F_{nir} \oplus F_{txt}$	0.7875	0.7883	0.7938	16.4174	17.1859	17.5160	0.2933	0.2742	0.2634
	$F_{rgb} + F_{nir}$	0.7605	0.7747	0.7789	16.1966	16.9488	17.1367	0.3623	0.3503	0.3505
	$F_{rgb} + F_{nir} + F_{txt}$	0.7514	0.7671	0.7735	14.6166	15.4207	16.5871	0.3405	0.3353	0.3246
NIRSplat-S	$attn(F_{rgb}, F_{txt})$	0.8053	0.8139	0.8160	18.8696	18.7132	19.1866	0.2765	0.2728	0.2457
	$attn(F_{rgb}, F_{nir})$	0.8205	0.8240	0.8314	20.0486	20.2083	20.9963	0.2244	0.2153	0.2130
	$attn(F_{rgb}, F_{nir}, F_{txt})$	**0.8268**	**0.8311**	**0.8421**	**20.7182**	**21.0169**	**21.0814**	**0.2070**	**0.2071**	**0.2080**

Table 3: Ablation study on various configurations.

approaches (Mihajlovic et al. 2024; Fan et al. 2024b). Meanwhile, SplatFields fails to preserve pixel-level fidelity, resulting in a notable drop in PSNR and suboptimal reconstruction quality. Furthermore, these models exhibit poor performance under extremely limited training budgets (1k iterations), suggesting a lack of inherent robustness. Although InstantSplat addresses these drawbacks, this paradigm is still limited in capturing visual details from challenging agricultural samples (*i.e.*, occlusion, uneven reflection), resulting in up to -2.6% SSIM, -2.5 PSNR, and +7.2% LPIPS loss, compared to Ours. To tackle these issues, we leverage a novel multimodal architecture, **NIRSplat**, which effectively generalizes agricultural environments. Notably, **NIRSplat** demonstrates its efficiency and validity by surpassing the performance of previous models that use 12 views despite using only 3 views.

Ablation Studies

Impact of Additional Modalities. This naturally raises a fundamental question: *Do additional modalities consistently yield performance improvements?* While additional modalities (NIR, Text) provide rich complementary cues, seamlessly integrating them remains a significant challenge. This difficulty largely stems from inherent modality gaps, spectral discrepancies, and disjoint embedding spaces across RGB, NIR, and textual inputs. To better understand this, we explore various fusion strategies in Tab. 3. Naïvely adding NIR signals significantly degrades performance, leading to up to a **9% drop in SSIM**. We attribute this to the *spectral discrepancy* and value distribution mismatch between visible and near-infrared modalities. Conventional fusion techniques (element-wise summation, feature concatenation) result in trivial improvements, failing to resolve the semantic and spatial misalignment. Importantly, this limitation potentially becomes more pronounced when incorporating textual metadata as shown in Tab. 3 and Tab. 4: without proper geometric references, semantic and geometric misalignment between visual and textual features causes suboptimal training (*i.e.*, **-1.52% SSIM**, **-2.7 PSNR**, **+6.74% LPIPS**). To address this, we introduce an effective cross-modal 3D reconstruction method, **NIRSplat** that leverages a geometry-guided 3D point-based positional encoding (PE) scheme anchoring features from all modalities to a shared 2D projection space. Consequently, NIRSplat facilitates robust alignment for uncertain cross-modal knowledge by allowing the model to leverage the most informative signals from each modality, thereby achieving superior 3D consistency and fidelity.

PE		6 views		
w/o	w/	SSIM (↑)	PSNR (↑)	LPIPS (↓)
✓		0.8159	18.3132	0.2745
	✓	**0.8311**	**21.0169**	**0.2071**

Table 4: Ablation study of PE (Eq. (3)).

Effect of Botanic-Aware Knowledge. To further enhance

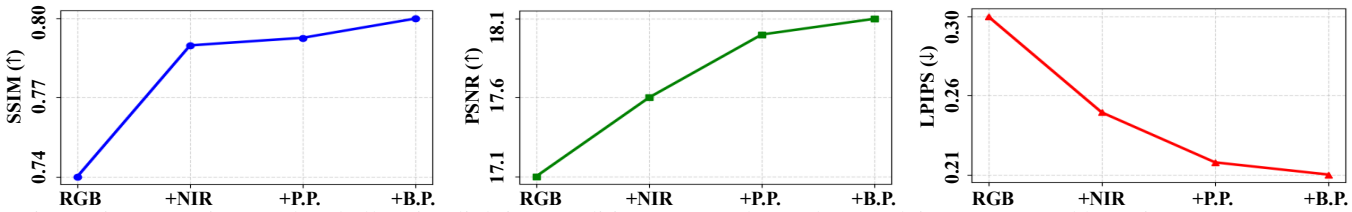


Figure 5: Comparison under challenging lighting conditions. P.P. and B.P. denote plain prompts and botanic-aware prompts.

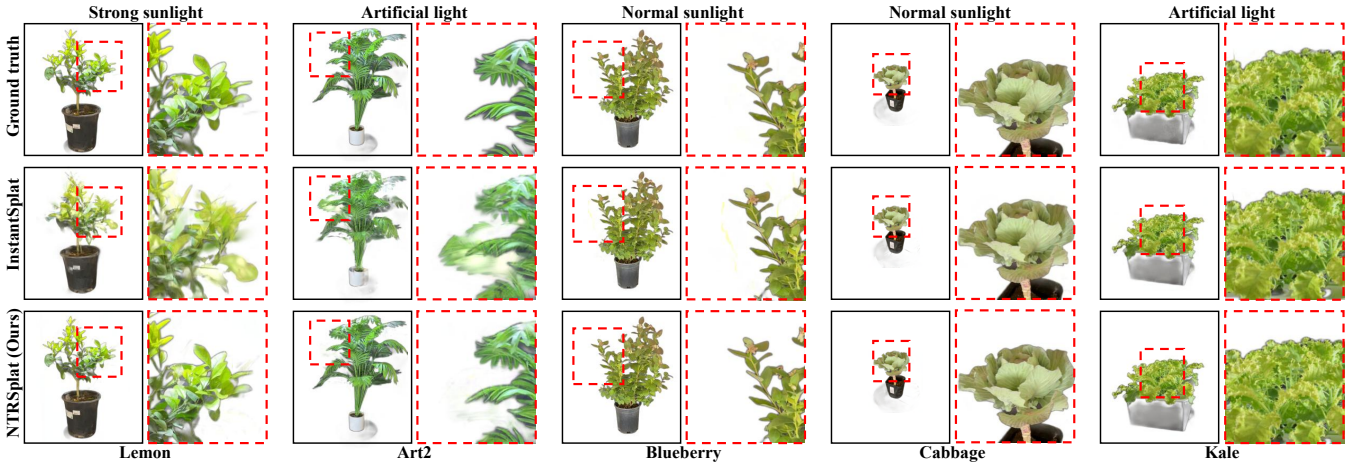


Figure 6: Qualitative visualization in a 3-view setup, demonstrating the results under diverse lighting conditions: Lemon (strong light), Kale (artificial light), Art2 (occlusion), and Cabbage (small object). The red boxes highlight the challenging semantic loss.

scene understanding in agricultural domains, we incorporate botanic-aware textual prompts that encode physiological and environmental context (*i.e.*, NDVI, NDWI, chlorophyll levels, growth stages, and lighting conditions), as detailed in the supplement. One might reasonably question the effectiveness of language-based guidance in dense 3D reconstruction, as textual descriptions are often semantically abstract and may lack spatial precision. To address this concern, we conduct an ablation study comparing plain prompts with botanic-aware prompts that explicitly embed spectral and biological indices. As shown in Fig. 5, botanic-aware prompts lead to consistent performance gains, with improvements compared to plain prompts under challenging scenarios. These results indicate that botanic-aware prompts, unlike plain prompts, act as high-level priors that reinforce correlations between NIR responses and botanical states, guiding cross-modal attention toward semantically and structurally relevant regions and improving reconstruction fidelity under ambiguity or occlusion.

Qualitative Analyses

We qualitatively evaluate our method under four challenging agricultural scenarios: (i) **Strong sunlight** (Lemon), (ii) **artificial lighting** (Kale, Art2), (iii) **occlusion** (Art2), and (iv) **small objects** (Cabbage), with Blueberry serving as a moderate-complexity reference (see Fig. 6). Conventional methods (*e.g.*, InstantSplat) often fail to preserve semantic and geometric fidelity, showing blurred textures and structural collapse under occlusion or extreme lighting. In contrast, **NIRSplat** achieves clear improvements by leveraging spectral cues (e.g., NIR) and botanic-aware priors, enabling better detail recovery and structural consistency. Notably, NIRSplat recovers saturated regions under strong illumination, resolves fine details in small-scale objects, and maintains coherence in occluded or low-texture areas—demonstrating its robustness across diverse agricultural conditions.

Conclusion

Summary. In this work, we introduced the **NIRPlant** dataset, which incorporates multimodal data from Near-Infrared (NIR), text, and RGB sensors in both indoor and outdoor agricultural environments. By leveraging the unique advantages of NIR and botanical-aware text, we addressed the challenges of 3D reconstruction in agriculture, including uneven lighting, occlusion, and novel perspectives. We also presented **NIRSplat**, an effective multimodal Gaussian Splatting framework that bridges these modalities through cross-attention and strong geometric priors from 3D point-based positional encoding. Importantly, **NIRSplat** significantly improves scene understanding, leveraging invisible NIR and contextual text knowledge. Through comprehensive experiments, we demonstrated that **NIRSplat** outperforms state-of-the-art methods, highlighting the potential of multimodal integration for robust agricultural 3D reconstruction.

Limitations and Future Work. We found that additional inputs lead to significant computational overheads, which limit the efficiency of real-time rendering. While our transformer-based approach effectively bridges the multimodality, it suffers from the cost of increased model complexity and capacity. In future work, we aim to address this issue by seamlessly aligning the three different modalities, ensuring more efficient integration and reducing computational overhead.

Acknowledgements

This research was supported by the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025), and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program, Korea University).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

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 (ICCV)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5470–5479.

Bian, W.; Wang, Z.; Li, K.; Bian, J.-W.; and Prisacariu, V. A. 2023. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4160–4169.

Fan, Z.; Cong, W.; Wen, K.; Wang, K.; Zhang, J.; Ding, X.; Xu, D.; Ivanovic, B.; Pavone, M.; Pavlakos, G.; Wang, Z.; and Wang, Y. 2024a. InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds. arXiv:2403.20309.

Fan, Z.; Wen, K.; Cong, W.; Wang, K.; Zhang, J.; Ding, X.; Xu, D.; Ivanovic, B.; Pavone, M.; Pavlakos, G.; et al. 2024b. InstantSplat: Sparse-view SfM-free Gaussian Splatting in Seconds. *arXiv preprint arXiv:2403.20309*.

Fu, Y.; Liu, S.; Kulkarni, A.; Kautz, J.; Efros, A. A.; and Wang, X. 2024. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20796–20805.

Garbin, S. J.; Kowalski, M.; Johnson, M.; Shotton, J.; and Valentin, J. 2021. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14346–14355.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 270–279.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3828–3838.

In Lee, D.; Park, H.; Seo, J.; Park, E.; Park, H.; Dam Baek, H.; Sangheon, S.; Kim, S.; et al. 2024. EditSplat: Multi-View Fusion and Attention-Guided Optimization for View-Consistent 3D Scene Editing with 3D Gaussian Splatting. *arXiv e-prints*, arXiv–2412.

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).

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics*, 36(4).

Lee, S. H.; Li, Y.; Ke, J.; Yoo, I.; Zhang, H.; Yu, J.; Wang, Q.; Deng, F.; Entis, G.; He, J.; et al. 2024. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. In *European Conference on Computer Vision*, 462–478. Springer.

Leroy, V.; Cabon, Y.; and Revaud, J. 2024. Grounding Image Matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 71–91.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, K.; Bian, J.-W.; Castle, R.; Torr, P. H.; and Prisacariu, V. A. 2023b. Mobilebrick: Building lego for 3d reconstruction on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4892–4901.

Li, K.; Pham, T.; Zhan, H.; and Reid, I. 2018. Efficient dense point cloud object reconstruction using deformation vector fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 497–513.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Lin, C.-H.; Kong, C.; and Lucey, S. 2018. Learning efficient point cloud generation for dense 3d object reconstruction. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. PETR: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision (ECCV)*, 531–548. Springer.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S. M.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mihajlovic, M.; Prokudin, S.; Tang, S.; Maier, R.; Bogo, F.; Tung, T.; and Boyer, E. 2024. Splatfields: Neural gaussian splats for sparse 3d and 4d reconstruction. In *European Conference on Computer Vision*, 313–332. Springer.

Nguyen, A.-D.; Choi, S.; Kim, W.; and Lee, S. 2019. GraphX-convolution for point cloud deformation in 2D-to-3D conversion. In *Proceedings of the IEEE/CVF International conference on computer vision*, 8628–8637.

Oh, G.; Jeong, J.; Kim, S.; Byeon, W.; Kim, J.; Kim, S.; and Kim, S. 2024. Mevg: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*, 401–418. Springer.

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

Roh, W.; Jung, H.; Kim, J. W.; Lee, S.; Yoo, I.; Lugmayr, A.; Chi, S.; Ramani, K.; and Kim, S. 2024. CATSplat: Context-Aware Transformer with Spatial Guidance for Generalizable 3D Gaussian Splatting from A Single-View Image. *arXiv preprint arXiv:2412.12906*.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shu, C.; Deng, J.; Yu, F.; and Liu, Y. 2023. 3DPPE: 3D Point Positional Encoding for Transformer-based Multi-Camera 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3580–3589.

Sinha, A.; Unmesh, A.; Huang, Q.; and Ramani, K. 2017. Surfnet: Generating 3d shape surfaces using deep residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6040–6049.

Toschi, M.; De Matteo, R.; Spezialetti, R.; De Gregorio, D.; Di Stefano, L.; and Salti, S. 2023. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20762–20772.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Voynov, O.; Bobrovskikh, G.; Karpyshev, P.; Galochkin, S.; Ardelean, A.-T.; Bozhenko, A.; Karmanova, E.; Kopanev, P.; Labutin-Rymsho, Y.; Rakhimov, R.; et al. 2023. Multi-sensor large-scale dataset for multi-view 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21392–21403.

Wang, F.; Rakotosaona, M.-J.; Niemeyer, M.; Szeliski, R.; Pollefeys, M.; and Tombari, F. 2024. UniSDF: Unifying Neural Representations for High-Fidelity 3D Reconstruction of Complex Scenes with Reflections. In *Advances in Neural Information Processing Systems*, volume 37, 3157–3184. Curran Associates, Inc.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; Lin, D.; and Liu, Z. 2023. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 803–814.

Xiong, B.; Li, Z.; and Li, Z. 2024. GauU-Scene: A Scene Reconstruction Benchmark on Large Scale 3D Reconstruction Dataset Using Gaussian Splatting. arXiv:2401.14032.

Yan, Z.; Mazzacca, G.; Rigon, S.; Farella, E. M.; Trybala, P.; Remondino, F.; et al. 2023. NeRFBK: a holistic dataset for benchmarking NeRF-based 3D reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48(1): 219–226.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *CVPR*.

Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Zhu, C.; Xiong, Z.; Liang, T.; et al. 2023. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9150–9161.

Yu, Z.; Chen, A.; Huang, B.; Sattler, T.; and Geiger, A. 2024. Mip-Splatting: Alias-free 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19447–19456.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, J.; Li, J.; Yu, X.; Huang, L.; Gu, L.; Zheng, J.; and Bai, X. 2024. CoR-GS: sparse-view 3D Gaussian splatting via co-regularization. In *European Conference on Computer Vision*, 335–352. Springer.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.