

CATSplat: Context-Aware Transformer with Spatial Guidance for Generalizable 3D Gaussian Splatting from A Single-View Image

Wonseok Roh^{1*} Hwanhee Jung^{1*} Jong Wook Kim¹ Seungwan Lee¹
 Innfarn Yoo³ Andreas Lugmayr² Seunggeun Chi⁴ Karthik Ramani⁴ Sangpil Kim^{1†}
¹Korea University ²Google ³CNAPS.AI Inc. ⁴Purdue University

Abstract

Recently, generalizable feed-forward methods based on 3D Gaussian Splatting have gained significant attention for their potential to reconstruct 3D scenes using finite resources. These approaches create a 3D radiance field, parameterized by per-pixel 3D Gaussian primitives, from just a few images in a single forward pass. Unlike multi-view methods that benefit from cross-view correspondences, 3D scene reconstruction with a single-view image remains an underexplored area. In this work, we introduce **CATSplat**, a novel generalizable transformer-based framework designed to break through the inherent constraints in monocular settings. First, we propose leveraging textual guidance from a visual-language model to complement insufficient information from single-view image features. By incorporating scene-specific contextual details from text embeddings through cross-attention, we pave the way for context-aware 3D scene reconstruction beyond relying solely on visual cues. Moreover, we advocate utilizing spatial guidance from 3D point features toward comprehensive geometric understanding under monocular settings. With 3D priors, image features can capture rich structural insights for predicting 3D Gaussians without multi-view techniques. Extensive experiments on large-scale datasets demonstrate the state-of-the-art performance of CATSplat in single-view 3D scene reconstruction with high-quality novel view synthesis.

1. Introduction

3D scene reconstruction and novel view synthesis are fundamental tasks in modern computer vision and graphics, driving advancements across diverse domains [2, 12, 27, 32], such as virtual reality and autonomous navigation. Together, they create 3D scene representations using 2D source images and produce realistic images from unseen perspectives. Early approaches [5, 8, 33, 37] (e.g., NeRF) have made impressive progress through differentiable vol-

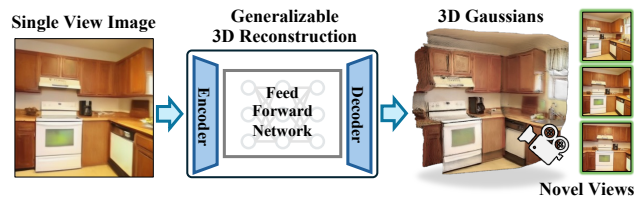


Figure 1. Overview of the generalizable 3D scene reconstruction pipeline. The feed-forward network creates a 3D radiance field using 3D Gaussians, all within an end-to-end differentiable system.

ume rendering. However, they are still far from real-time scenarios due to the heavy computational demands. Unlike previous methods, 3D Gaussian Splatting (3DGS) based approaches [20, 56, 59] have emerged as leading frontrunners, achieving high performance with real-time rendering capabilities. They employ 3D Gaussians for explicit scene representations via efficient rasterization-based rendering.

Recently, generalizable feed-forward methods [7, 9, 44, 50, 60] based on 3DGS [20] have attracted growing interest for their ability to reconstruct 3D scenes, even with limited resources like sparse view images. They create a 3D radiance field parameterized by per-pixel Gaussian primitives from just a few input images (typically one or two) in a single forward pass without scene-specific optimization, as outlined in Fig. 1. For example, pixelSplat [7] samples Gaussian centers from probabilistic depth distributions using a multi-view epipolar geometry, while MVSpLat [9] constructs cost volumes from source images to extract geometric cues. Both approaches benefit from the cross-view correspondences between a pair of images to capture useful cues for the Gaussian parameters prediction. In contrast to multi-view settings, monocular 3D scene reconstruction still remains challenging due to relatively constrained information, relying solely on an image. Although Flash3D [44] has pioneered a 3DGS-based generalizable single-view 3D scene reconstruction method with a pre-trained depth estimation network [38], this area has yet to be fully explored.

To tackle the challenges in monocular scenarios, we introduce **CATSplat**, a carefully designed transformer that leverages two intelligent guidance to supplement the insufficient information from single-view image features. Based

*Equal contribution.

†Corresponding author

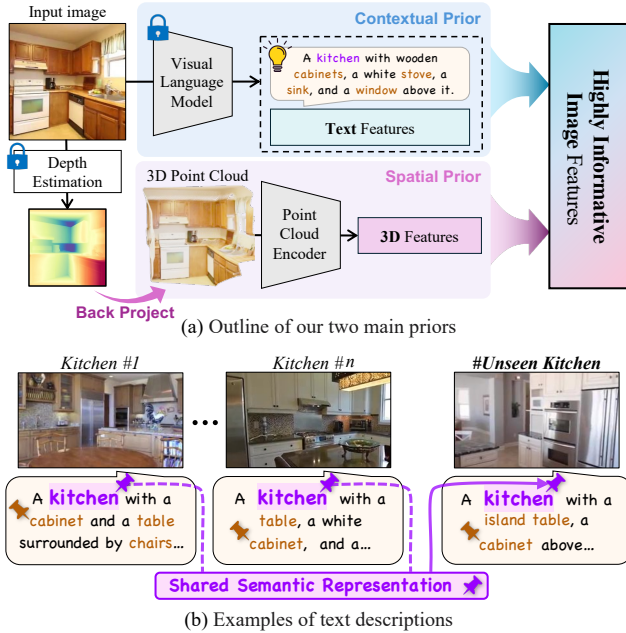


Figure 2. We introduce **CATSplat**, a Context-Aware Transformer with Spatial Guidance for Generalizable 3D Gaussian Splatting from a single image. (a) Our two main priors, and (b) Examples of text descriptions (from the VLM) representing an input image.

on the traditional generalizable 3DGS paradigm, which predicts Gaussian primitives from image features, we focus on enhancing these features with additional insights. First, we propose using textual guidance as contextual priors. One of the most promising ways to employ text guidance is through visual-language models (VLM) [1, 25, 28, 65]. They have showcased their potential in various vision tasks [18, 21, 22, 66] to provide visual-linguistic knowledge learned from large-scale multimodal data. Motivated by the success of VLMs, we utilize their text embeddings representing the input image to guide the network towards context-aware scene reconstruction, as in Fig. 2 (a). Here, text features usually capture spatial context (e.g., *kitchen*), object categories with their relationships (e.g., *a table surrounded by chairs*), as well as overall compositions, as in Fig. 2 (b). These text-embedded scene semantics not only enrich contextual understanding but also improve generalizability via shared semantic representations across similar yet distinct scenes. For instance, when evaluating on an unseen *kitchen* scene, the text embeddings serve as extra anchors, helping the network identify the scene as likely a *kitchen*. With explicit textual guidance, our network can better apply the knowledge learned from various *kitchen* scenes ($\#1 \sim n$), boosting generalizability beyond solely relying on visual features.

In addition to the contextual guidance, we explore additional avenues to enhance the image features. In generalizable tasks with sparse images, gaining insights into 3D geometrics is crucial to reconstruct robust scenes in 3D space. Typically, multi-view methods [7, 9] operate physical techniques like triangulation to capture rich 3D cues from cross-

view perspectives. However, such techniques are unavailable in monocular settings, leading to constrained geometric details. In this context, we advocate for integrating 3D guidance into 2D features to enrich their spatial understanding. Instead of merely using a 2D depth map from an off-the-shelf depth estimation model, as used in previous work [44], we further leverage its 3D representation as a backprojected point cloud. Unlike depth maps, which are confined to a 2D grid, point clouds encode continuous x , y , and z -axis information in 3D space, capturing richer geometric structures. Thus, we utilize point clouds as a source of structural guidance for reliable 3D perception. Specifically, as in Fig. 2 (a), we extract 3D features from these points and strengthen image features through attention mechanisms. Ultimately, our image features, influenced by two constructive priors, are highly informative for 3D Gaussian-based reconstruction.

Given landmark datasets, RealEstate10K (RE10K) [64], ACID [26], KITTI [15], and NYUv2 [42], we validate the generalizability and effectiveness of our novel framework. To summarize, our main contributions are listed as follows:

- We introduce **CATSplat**, a novel generalizable framework for monocular 3D scene reconstruction. We leverage the rich contextual cues of text embeddings from the VLM as insightful guidance toward context awareness, complementing limited information from a single image.
- We propose 3D spatial guidance from 3D point features to enrich geometric details in single-view settings. With 3D priors, image features can capture valuable cues for predicting 3D Gaussians without multi-view techniques.
- We analyze the effectiveness of our method on challenging datasets. Extensive quantitative and qualitative experiments demonstrate that ours achieves new state-of-the-art performance on single-view 3D scene reconstruction.

2. Related Work

Sparse-view 3D Reconstruction. Recent progress in neural fields [32, 43, 54] and volume rendering [29, 46] has advanced 3D reconstruction and novel view synthesis, even with sparse-view images. FreeNeRF [55] regularizes frequency to address few-shot neural rendering, while pixelNeRF [57] predicts a neural radiance field in the camera coordinate using a feed-forward approach from few-view images. More recently, 3D Gaussian Splatting (3DGS) [20] has revolutionized the field of 3D reconstruction, achieving real-time rendering. Inspired by the success of 3DGS, pixelSplat [7] has pioneered the feed-forward network, which reconstructs a 3D radiance field parameterized by 3D Gaussian primitives from a pair of images. Then, diverse multi-view generalizable 3DGS approaches [9, 50, 60] have since developed with a similar structure. MVSpLat [9] constructs cost volumes to capture cross-view similarities for accurate Gaussians, and latentSplat [50] proposes variational Gaus-

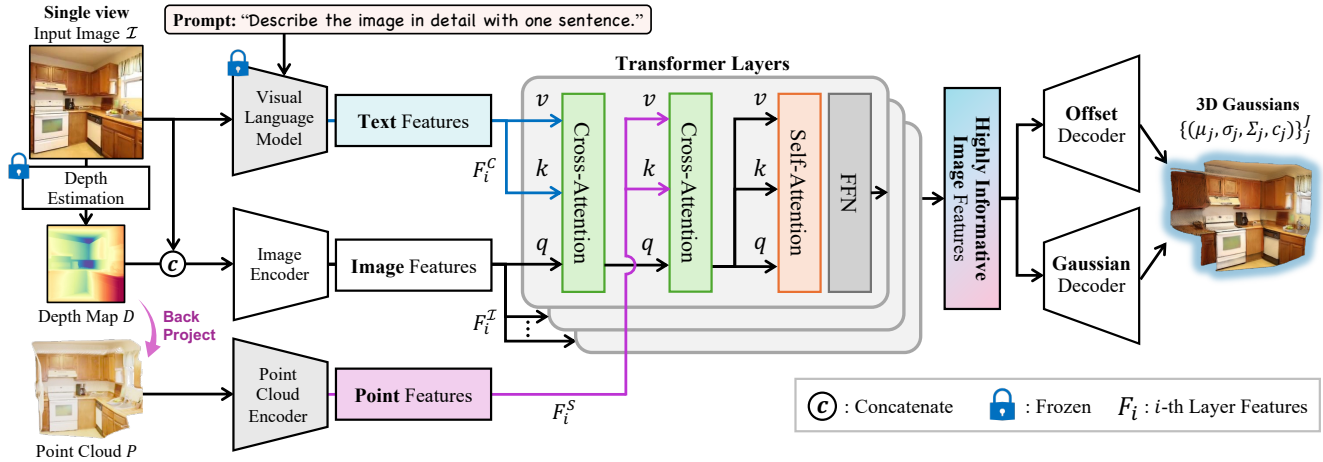


Figure 3. Overview of **CATSplat** framework. CATSplat takes an image \mathcal{I} and predicts 3D Gaussian primitives $\{(\mu_j, \alpha_j, \Sigma_j, c_j)\}_j^J$ to construct a scene-representative 3D radiance field in a single forward pass. Our primary goal is to go beyond the finite knowledge inherent in single-view image features leveraging our two innovative priors. Through cross-attention layers, we enhance image features F_i^I to be highly informative by incorporating valuable insights: contextual cues from text features F_i^C , and spatial cues from 3D point features F_i^S .

sians to encode uncertainty in a latent space. Unlike these methods, monocular tasks remain more challenging due to limited details, without benefits of cross-view properties.

Single-view 3D Reconstruction. Early approaches [48, 51] have proposed practical strategies to overcome the constraints of single-view settings. SynSin [51] introduces a differentiable point renderer that projects a 3D point cloud from a single image into target views. [48] predicts multi-plane images (MPI) [64] directly from a single image without correlations between multiple views. In line with recent trends, single-view 3D reconstruction quality has significantly improved, thanks to innovations in NeRF [32] and 3DGS [20]. Built upon NeRF, MINE [23] extends MPI to a continuous 3D representation, and BTS [52] predicts less complex continuous density fields from an image. Recently, Splatter Image [45] involves 3DGS based on an image-to-image neural network for monocular object reconstruction. Also, Flash3D [44] predicts pixel-wise Gaussian parameters for scene reconstruction in a single forward pass, using depth cues from a monocular depth estimation model [38]. Based on the core paradigm of generalizable 3DGS frameworks, we propose two beneficial priors to complement insufficient cues from an image, enhancing generalizability.

Vision-Language Models for Vision Tasks. Visual Language Models (VLMs) have emerged as powerful tools for bridging the gap between visual and textual modalities [14, 30], achieving outstanding performance in diverse vision tasks, such as image captioning [3, 24, 25, 35, 58], image-text retrieval [19, 31, 41, 63], and visual question answering (VQA) [17, 34, 40]. These models use large-scale image-text pair datasets to learn joint representations, encouraging seamless understanding and integration across modalities. Early approaches like CLIP [41] and ALIGN [19] leverage contrastive learning to relate image and text data within a shared embedding space, enabling effective zero-shot gen-

eralization between them. Recently, the success of Large Language Models (LLMs) [4, 6, 10, 47] has driven significant advancements in visual-language processing. For example, BLIP-2 [25] and LLaVA [28] achieve strong performance in image captioning with context-rich visual descriptions based on LLMs [10, 11, 62]. They aim to connect image features from visual encoders into the language space of pre-trained LLMs. Motivated by the significance of VLMs, we employ contextual clues of their text embeddings to provide additional perspectives and enhance generalizability.

3. Method

In this section, we introduce CATSplat, a novel generalizable framework for monocular 3D scene reconstruction with 3D Gaussian Splatting. We first provide an overview of the whole pipeline (Sec. 3.1 and Fig. 3) and then elaborate on technical details: Context-Aware 3D Reconstruction (Sec. 3.2) and Spatial Guidance for 3D Insights (Sec. 3.3).

3.1. Overview

Recent generalizable feed-forward frameworks [7, 9, 44, 50, 60] commonly follow a similar paradigm; they construct a 3D radiance field from N sparse-view images $\mathcal{I}^N \in \mathbb{R}^{N \times H \times W \times 3}$ in a single forward pass with pixel-aligned J Gaussian primitives $\{(\mu_j, \alpha_j, \Sigma_j, c_j)\}_j^J$, including position μ_j , opacity α_j , covariance Σ_j , and spherical harmonics coefficients c_j . In this paradigm, it is challenging to reconstruct the vivid scene from a single image due to limited resources, comparing with multi-view configurations. To overcome this constraint, we propose a carefully designed transformer that leverages two extra guidance for enhancing knowledge of single-view image features: (1) Textual Guidance, which provides explicit contextual clues for the scene, and (2) Spatial Guidance, which enriches three-dimensional structural information of 2D features, as illustrated in Fig. 3.

Feed-Forward Network with Transformer. From a single input image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, we first predict a depth map $D \in \mathbb{R}_+^{H \times W \times 1}$ as potential centers for Gaussians, employing a pre-trained monocular depth estimation model [38]. Given \mathcal{I} and its estimated depth map D , we channel-wise concatenate them as $\mathcal{I}' \in \mathbb{R}^{H \times W \times 4}$, then feed \mathcal{I}' into a ResNet-based image encoder [16] to produce hierarchical depth-conditioned image features $F_i^{\mathcal{I}} \in \mathbb{R}^{H_i \times W_i \times D_i^{\mathcal{I}}}$. Then, we utilize a multi-resolution transformer that encourages image features $F_i^{\mathcal{I}}$ to effectively represent both global structures and fine details across various resolutions, improving the overall understanding of the scene. We specifically use three layers with three resolution features. Based on transformer architecture, we extend the cross-attention mechanism to interact with our two novel priors, as described in Sec. 3.2 and Sec. 3.3, further enriching the feature representation. Through iterative layers, our transformer yields highly informative image features $\tilde{F}_i^{\mathcal{I}} \in \mathbb{R}^{H_i \times W_i \times D_i^{\mathcal{I}}}$ well-suited for effective scene reconstruction in 3D space. Finally, we estimate per-pixel 3D Gaussian parameters from $\tilde{F}_i^{\mathcal{I}}$ using ResNet-based decoders, as detailed in Sec 3.4.

3.2. Context-Aware 3D Reconstruction

In real-world scenarios, diverse objects are usually placed in inconsistent patterns without following conventional rules. These variabilities make monocular 3D scene reconstruction more challenging. To extend the representational capability of image features, we propose textual information as a rich source of hidden context, enhancing generalizability.

Incorporation of Textual Cues. Recent advancements in large-scale visual language models [1, 25, 28, 65] (VLM) have highlighted the benefits of their general embedded knowledge, which mirrors the diversity of real-world contexts. In this work, we take advantage of language-driven scene interpretations embedded in the text representations produced by these models. With a single-view source image \mathcal{I} , we prompt the pre-trained VLM [28] to generate a detailed, one-sentence description of the scene. During this procedure, we utilize text embeddings $F^C \in \mathbb{R}^{N_c \times D^C}$ from a well-aligned multimodal space before they are processed into linguistic descriptions. Our main focus is on the contextual details from F^C , such as object identities, spatial relationships, and scene semantics, which can potentially serve as influential biases for enhancing generalizability. To softly incorporate supplemental cues from F^C into image features $F^{\mathcal{I}}$, we employ iterative cross-attention layers. For each transformer layer designed to use multi-scale features, we convert F^C into $F_i^C \in \mathbb{R}^{N_c \times D_i^C}$ to align the dimension with its corresponding $F_i^{\mathcal{I}}$ using a linear layer, as illustrated in Fig. 4. Given $F_i^{\mathcal{I}}$ and F_i^C , queries \mathbf{Q}_i are projected from $F_i^{\mathcal{I}}$, and keys \mathbf{K}_i and values \mathbf{V}_i are from F_i^C , as follows:

$$\mathbf{Q}_i = W_q \cdot F_i^{\mathcal{I}}, \quad \mathbf{K}_i = W_k \cdot F_i^C, \quad \mathbf{V}_i = W_v \cdot F_i^C \quad (1)$$

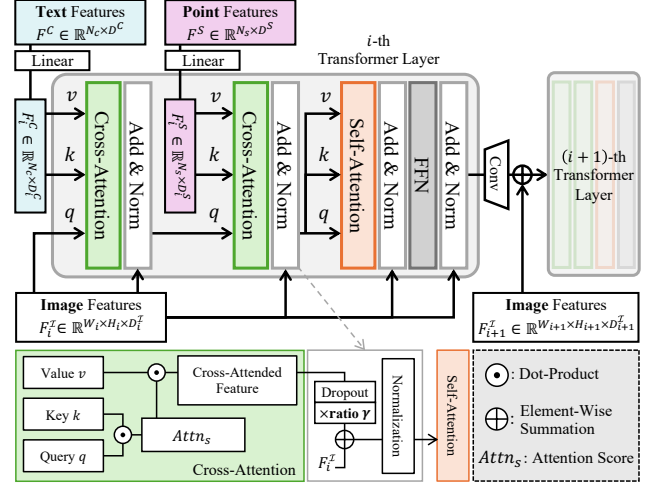


Figure 4. Detailed transformer pipeline. In the i -th layer, we first operate cross-attention between $F_i^{\mathcal{I}}$ and F_i^C , then proceed cross-attention with F_i^S . We also use a ratio γ to preserve visual information from $F_i^{\mathcal{I}}$ while incorporating extra cues from F_i^C and F_i^S .

where W denotes the learnable parameters of each projection layer. Then, we associate them through cross-attention:

$$F_i^{\mathcal{I}C} = \text{Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{Softmax}\left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_i^T}{\sqrt{D_i}}\right) \mathbf{V}_i \quad (2)$$

where $F_i^{\mathcal{I}C}$ represents output features containing not only visual clues from $F_i^{\mathcal{I}}$ but also textual clues from F_i^C . Finally, our iterative layers continuously establish valuable connections between source image features and additional contextual priors, facilitating more generalizable 3D reconstruction of real-world scenes under limited resources.

3.3. Spatial Guidance for 3D Insights

In multi-view configurations, each perspective contributes unique spatial information, boosting the reconstruction of complex three-dimensional structures. Yet, single-view often falls short of 3D cues for comprehensive geometric understanding. To bridge this gap, we introduce efficient spatial guidance based on the 3D representation of a 2D depth map, which provides a broader geometric context for reliable 3D perception independent of stereo vision expertise.

Incorporation of Spatial Cues. Solid geometric awareness is essential for accurately depicting a scene within 3D space. To capture 3D cues from a single image, traditional approaches [23, 44, 45] often rely on depth information in a two-dimensional format. Beyond its conventional use, we extend the estimated per-pixel 2D depth $d \in D$ into a full 3D representation for more direct spatial knowledge. Given camera parameters $K = \text{diag}(f_x, f_y, 1) \in \mathbb{R}^{3 \times 3}$, where f denotes the focal length, we unproject D into 3D space as point cloud $P \in \mathbb{R}^{H \times W \times 3}$, with each point $\mathbf{p} \in P$:

$$\mathbf{p} = K^{-1} \cdot \mathbf{u} \cdot d = (u_x d / f_x, u_y d / f_y, d) \quad (3)$$

where $\mathbf{u} = (u_x, u_y, 1) \in \mathcal{I}$ is one of the image pixels. From this set of points P , we extract point features $F^S \in \mathbb{R}^{N_s \times D^S}$

using a PointNet-based encoder [39] for better spatial reasoning. These point embeddings usually encode important geometric details, from depth relationships to surface orientations, going beyond static depth information. In order to integrate such valuable clues into image features while overcoming the domain gap between 2D and 3D representations, we leverage cross-attention layers. Similar to the approach for textual cues, we project F^S into $F_i^S \in \mathbb{R}^{N_s \times D_i^S}$ and further enrich context-guided image features F_i^{IC} (Eq. 2) from the previous cross-attention layer with F_i^S as follows:

$$F_i^{ICS} = \text{Attn}(\mathbf{Q}'_i, \mathbf{K}'_i, \mathbf{V}'_i) = \text{Softmax}\left(\frac{\mathbf{Q}'_i \cdot \mathbf{K}'_i{}^T}{\sqrt{D_i}}\right) \mathbf{V}'_i \quad (4)$$

where \mathbf{Q}'_i are projected from F_i^{IC} , and \mathbf{K}'_i and \mathbf{V}'_i are from F_i^S . During the add and normalization process after cross-attention, as shown in Fig. 4 below, we use the ratio γ to preserve core visual information from the source image while incorporating practical cues from our two novel priors as:

$$\tilde{F}_i^{ICS} = \text{Norm}(F_i^{IC} + \gamma \text{Dropout}(F_i^{ICS})) \quad (5)$$

Then, we refine \tilde{F}_i^{ICS} to \tilde{F}_i^I with the self-attention layer, ensuring seamless knowledge enhancement across the feature space. Ultimately, the final transformer output features \tilde{F}_i^I are highly informative for robust 3D scene reconstruction, even with an image. Note that we provide more technical details and the workflow algorithm in Supp (Sec.2).

3.4. Gaussian Parameters Prediction

With insightful features \tilde{F}_i^I , we predict parameters for J pixel-aligned 3D Gaussians $\{(\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j, \boldsymbol{\Sigma}_j, \mathbf{c}_j)\}_j^J$ through ResNet-based decoders [16] to represent the 3D scene.

Gaussian center $\boldsymbol{\mu}$. For precise scene reconstruction, we predict depth offsets $\delta \in \mathbb{R}_+^{H \times W \times 1}$ to refine per-pixel depth $d \in D$ and 3D offsets $\Delta_j \in \mathbb{R}^3$ for center-wise alignment following [44, 45]. Then, we unproject the 2D refined depth $\tilde{d} = d + \delta$ into 3D points using the provided camera parameters K to produce potential centers. Given Δ_j and projected points, the j^{th} Gaussian center $\boldsymbol{\mu}_j$ is set as follows:

$$\boldsymbol{\mu}_j = K^{-1} \cdot \mathbf{u} \cdot \tilde{d} + \Delta_j \quad (6)$$

$$= (u_x \tilde{d}/f_x + \Delta_x, u_y \tilde{d}/f_y + \Delta_y, \tilde{d} + \Delta_z) \quad (7)$$

where $\mathbf{u} = (u_x, u_y, 1) \in \mathcal{I}$ is one of the image pixels.

Opacity α , Covariance $\boldsymbol{\Sigma}$, and Color \mathbf{c} . In line with previous generalizable feed-forward methods [7, 9] using 3DGS, we operate convolutional layers to predict each parameter. We use the sigmoid activation function for the opacity α to ensure that values are bounded between 0 and 1. Additionally, we estimate a rotation matrix R and a scaling matrix S to construct the covariance matrix $\boldsymbol{\Sigma} = RSS^T R^T$. Also, for the color, we decode spherical harmonics coefficients \mathbf{c} .

Loss Function. Finally, we render images $\hat{\mathcal{I}}_t$ from novel viewpoints based on the reconstructed 3D scene using rasterization operation. For training, we calculate the following loss \mathcal{L}_{total} as the sum of the three losses to optimize the

quality of the rendered images $\hat{\mathcal{I}}_t$ with GT target images \mathcal{I}_t :

$$\mathcal{L}_{total} = \lambda_{\ell_1} \mathcal{L}_{\ell_1} + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}} + \lambda_{\text{lpiips}} \mathcal{L}_{\text{lpiips}} \quad (8)$$

where $\mathcal{L}_{\text{ssim}}$ and $\mathcal{L}_{\text{lpiips}}$ represent Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [61] losses, respectively, and each λ is a hyperparameter to handle the strength of the respective loss term.

4. Experiments

4.1. Experimental Setup

Datasets. In this study, we train and evaluate the overall performance using a large-scale dataset, RealEstate10K (RE10K) [64], containing home walkthrough videos. We also use three additional datasets, NYUv2 (indoor) [42], ACID (nature) [26], and KITTI (driving) [15], for cross-dataset experiments. Detailed descriptions of datasets and implementation details are provided in the supplementary.

Evaluation Metrics. We quantitatively evaluate the 3D reconstruction performance using three traditional metrics for novel view synthesis: PSNR, SSIM [49], and LPIPS [61]. For comparisons with single-view 3D reconstruction methods, we evaluate three metrics on unseen target frames located 5 and 10 frames away from the input source image as well as a randomly sampled frame within a ± 30 frame range, following the standard evaluation protocol of previous methods [23, 44]. Also, to further evaluate our method, we adopt conventional interpolation and extrapolation protocols from pixelSplat [7] and latentSplat [50], respectively, following Flash3D [44]. For extrapolation, we sample target views up to 45 frames before or after the source frame.

4.2. Performance Comparisons with SOTA Models

Comparisons with Single-view Methods. In this section, we quantitatively compare our proposed framework CATSplat with existing state-of-the-art single-view 3D reconstruction methods [23, 44, 45, 48, 51, 52]. Despite significant advancements through robust radiance field rendering techniques [20, 32], monocular 3D scene reconstruction has yet to be fully explored and still faces challenges under resource constraints. To address this challenging task, we introduce a carefully designed transformer-based architecture with two novel priors, enriching image features to predict precise 3D Gaussians for scene representation. As reported in Tab. 1, we evaluate novel view synthesis performance on the RealEstate10K [64] dataset. CATSplat consistently outperforms previous methods with new state-of-the-art scores in terms of PSNR, SSIM, and LPIPS across three target frame at distinct locations. Specifically, CATSplat achieves high-quality rendering not only for nearby frames, such as those 5 or 10 frames apart, but also for frames randomly located at far distances (within a ± 30 frame range). These results demonstrate that our novel priors effectively complement limited cues from ResNet-based single image features.

Method	$n = 5$ (frames)			$n = 10$ (frames)			$n = \text{Random}$ (frames)		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
MPI [48]	27.10	0.870	–	24.40	0.812	–	23.52	0.785	–
BTS [52]	–	–	–	–	–	–	24.00	0.755	0.194
Splatter Image [45]	28.15	0.894	0.110	25.34	0.842	0.144	24.15	0.810	0.177
MINE [23]	28.45	0.897	0.111	25.89	0.850	0.150	24.75	0.820	0.179
Flash3D [44]	28.46	0.899	0.100	25.94	0.857	0.133	24.93	0.833	0.160
CATSplat (Ours)	29.09	0.907	0.094	26.44	0.866	0.125	25.45	0.841	0.151

Table 1. Comparisons of Novel View Synthesis (NVS) performance with state-of-the-art **single-view** 3D reconstruction approaches on the RealEstate10K [64] dataset. Following the standard protocol from [23, 44], we evaluate NVS metrics on unseen target frames located n frames away from the input source frame. Also, we randomly sample an extra target frame within 30 frames apart from the source frame.

Input	Method	Framework	RE10K Interpolation			RE10K Extrapolation		
			PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Two-View	pixelNeRF [57]	NeRF	20.51	0.592	0.550	20.05	0.575	0.567
	Du <i>et al.</i> [13]	NeRF	24.78	0.820	0.213	21.83	0.790	0.242
	pixelSplat [7]	3DGS	26.09	0.864	0.136	21.84	0.777	0.216
	latentSplat [50]	3DGS	23.93	0.812	0.164	22.62	0.777	0.196
	MVSplat [9]	3DGS	26.39	0.869	0.128	23.04	0.813	0.185
Single-View	Flash3D [44]	3DGS	23.87	0.811	0.185	24.10	0.815	0.185
	CATSplat (Ours)	3DGS	25.23	0.835	0.159	25.35	0.837	0.159

Table 2. Comparisons of NVS performance with state-of-the-art **few-view** 3D reconstruction approaches on the RealEstate10K [64]. Although we mainly focus on comparing with the leading single-view method, Flash3D [44], we also provide scores of two-view methods for additional references. Following Flash3D, we use interpolation and extrapolation protocols from previous works, [7] and [50], respectively.

Interpolation and Extrapolation. In multi-view setups, novel view synthesis is typically evaluated on target frames within the range of multiple input images (interpolation) and outside their range (extrapolation). In Tab. 2, to further validate our method, we evaluate CATSplat across both conventional settings, as established in Flash3D [44], a prominent single-view 3D scene reconstruction model. While our primary focus is on comparing with Flash3D, we also provide scores of multi-view methods [7, 9, 13, 50, 57] for additional references. First, CATSplat significantly surpasses Flash3D in the interpolation setup. Although our results are somewhat lower than recent two-view methods, which are robust for intermediate views using cross-view correspondence, ours achieves competitive scores. Moreover, for the extrapolation setup, CATSplat outperforms Flash3D by large margins. Notably, these impressive scores even exceed previous two-view methods despite using only a single image. In such extrapolation setups, target frames are usually over 45 frames away from the source image, representing nearly unseen views. These findings highlight the effectiveness of our novel priors, providing valuable insights for handling distant target views. Specifically, textual anchors based on the shared semantic representation of the scenes, along with deep spatial understanding from point features encoding continuous x, y, and z-axis knowledge, enhance generalizability in relatively restrictive monocular settings.

Cross-dataset Generalization. In Tab. 3, we demonstrate the strong generalizability of CATSplat across three differ-

Cross Dataset	Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
RE10K	Flash3D [44]	25.09	0.775	0.182
→ NYUv2	CATSplat (Ours)	25.57	0.781	0.157
RE10K	Flash3D [44]	24.28	0.730	0.263
→ ACID	CATSplat (Ours)	24.73	0.739	0.250
RE10K	Flash3D [44]	21.96	0.826	0.132
→ KITTI	CATSplat (Ours)	22.43	0.833	0.122

Table 3. Comparisons of cross-dataset generalization with the state-of-the-art single-view 3DGS method, Flash3D [44], on various real-world datasets: NYUv2 [42], ACID [26], and KITTI [15].

ent cross-dataset settings. In each case, we train our model on RE10K [64] and directly test it on the target datasets in a zero-shot manner. We first evaluate the generalization on the NYUv2 [42], which contains indoor scenes similar to the RE10K. CATSplat adeptly synthesizes images for previously unseen indoor environments. Then, we focus on outdoor scenarios with more significant domain gaps; specifically, the ACID [26] includes nature landscapes captured by aerial drones, and KITTI [15] comprises driving scenes tailored for autonomous driving. Within these challenging conditions, where filming techniques or object types (*e.g.*, *cars*, *buildings*) are dissimilar, CATSplat showcases superior generalizability than the latest method, Flash3D [44]. Through a series of experiments, we confirm the power of our intelligent priors in improving the informativeness of single-view image features for dynamic real-world scene reconstruction. We thoroughly analyze this in Supp (Sec.4.1).

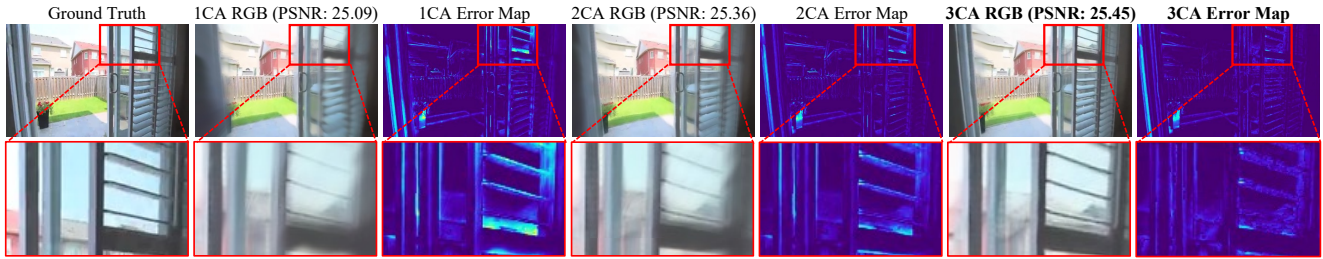


Figure 5. Ablation study to see the effect of iteratively incorporating our novel priors on the RE10K [64] ($n=Random$). For clear ablations, we keep the number of entire transformer layers consistent across the experiments and adjust only the number of cross-attentions (CA).

Method	$n = 10$ (frames)			$n = Random$ (frames)		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Baseline	26.04	0.857	0.132	25.02	0.834	0.159
w/ Contextual	26.40	0.864	0.127	25.40	0.838	0.153
w/ Spatial	26.38	0.864	0.127	25.42	0.837	0.153
CATSplat	26.44	0.866	0.125	25.45	0.841	0.151

Table 4. Ablation study to explore the effect of our two intelligent priors (Contextual and Spatial) across three different settings, as in Tab. 1, on the RE10K [64] dataset. Here, the “Baseline” indicates our basic transformer architecture without any proposed priors.

4.3. Ablation Studies

Effect of Contextual and Spatial Priors. In Tab. 4, we evaluate variants of our method with/ and w/o Contextual and Spatial priors. Here, the Baseline refers to our basic multi-resolution transformer architecture, excluding cross-attention with any of our proposed priors. The addition of each prior consistently enhances the visual quality of the rendered images from target novel perspectives. With contextual priors, the improvements across all metrics underscore the significance of incorporating extra context details for effective scene reconstruction. Also, spatial priors contribute impressive gains within all target settings, providing a more extensive geometric context for 3D understanding. Ultimately, combining both valuable priors together leads to further advancements, achieving the best scores. These results highlight that each prior plays a meaningful role in complementing limited details from single-view features.

Iteratively Incorporating Priors. Based on transformer, our feed-forward network seamlessly integrates two novel priors through iterative cross-attention layers. In Fig. 5, we explore the effect of varying the number of cross-attention iterations using rendered images with corresponding error maps. Specifically, we keep the total layers of the transformer consistent at three and apply cross-attention either in the first layer only, across two layers, or throughout all three layers. Across experiments, increasing cross-attention iterations leads to more precise, less blurry image synthesis with fewer errors. These improvements in visual quality through iterative incorporation underline the potential of our priors.

Vs. Pre-trained Visual Priors. Our primary goal in incorporating our two priors is to utilize comprehensive scene details that a simple image encoder may fail to capture from an image. In Tab. 5, we validate the practical benefits of our two priors by replacing them in the cross-attention layers

Method	$n = 10$ (frames)			$n = Random$ (frames)		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
w/ ConvNeXt-B	26.15	0.856	0.132	25.09	0.832	0.158
w/ ConvNeXt-L	26.17	0.857	0.132	25.12	0.833	0.157
w/ DINOv2-B	26.17	0.858	0.131	25.11	0.833	0.157
w/ DINOv2-g	26.19	0.859	0.131	25.17	0.834	0.156
w/ Contextual	26.40	0.864	0.127	25.40	0.838	0.153
w/ Spatial	26.38	0.864	0.127	25.42	0.837	0.153

Table 5. Ablation study to see the effect of our novel priors versus visual priors from large-scale pre-trained image encoders, such as DINOv2 [36] and ConvNeXt V2 [53], on the RE10K [64] dataset.

Method	$n = 10$ (frames)			$n = Random$ (frames)		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Baseline	26.04	0.857	0.132	25.02	0.834	0.159
w/ Scene Type	26.14	0.859	0.130	25.13	0.835	0.158
w/ Object List	26.23	0.862	0.128	25.25	0.836	0.155
w/ Extended	26.31	0.862	0.128	25.29	0.837	0.154
w/ Single Sent.	26.40	0.864	0.127	25.40	0.838	0.153

Table 6. Ablation study to see the impact of different text description formats on generalizable tasks. The “Baseline” is as in Tab. 4.

with visual priors from pre-trained image encoders, such as DINOv2 [36] and ConvNeXt V2 [53], on large-scale image datasets. The advanced image processing intelligence of these models, developed from extensive real-world images, contributes to moderate gains. However, since they mainly focus on visible aspects of the scene and our model is inherently designed to handle image data, injecting additional visual cues may introduce redundancy rather than meaningful benefits. In contrast, VLM extends beyond pure image processing by integrating language-driven insights, providing extra markers for our network to address unseen scenes. Also, with point features, our network captures more global geometric properties than the localized features from image encoders, enabling more accurate 3D Gaussian predictions. By leveraging both priors, we overcome the straightforward dependence on visual patterns, achieving notable increases.

Analysis of Context Details. In Tab. 6, we explore how different context details embedded in text features from pre-trained VLM [28] influence generalizability. We conduct experiments with four different prompt styles: identifying the scene type (e.g., *bedroom*), listing objects (e.g., *lamp*, *bed*), describing the scene with a detailed single sentence, and two or more sentences. While scene type or object list provides certain clues, their impact on performance is relatively modest. In contrast, sentence-level text embeddings

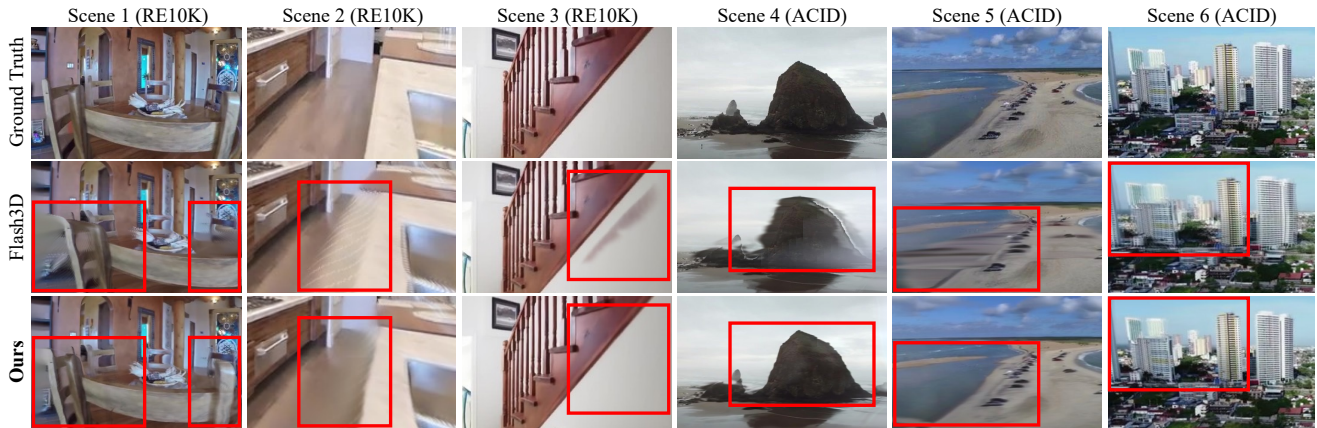


Figure 6. Qualitative comparisons of NVS performance between Flash3D [44] and ours with Ground Truth on the novel view frames from RealEstate10K [64] and ACID [26] (cross-dataset). We provide more visual results and details of user study in the supplementary material.

Method	$n = 10$ (frames)			$n = \text{Random}$ (frames)		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Baseline	26.04	0.857	0.132	25.02	0.834	0.159
w/o Depth Conc.	25.91	0.855	0.134	24.82	0.827	0.165
w/ Point Conc.	26.06	0.857	0.132	25.04	0.834	0.158
w/ Depth Feat.	26.18	0.859	0.130	25.16	0.835	0.157
w/ Point Feat.	26.38	0.864	0.127	25.42	0.837	0.153

Table 7. Ablation study to explore strategies for enriching geometric knowledge from a single image. The ‘‘Baseline’’ is as in Tab. 4.

contain more practical context details, such as texture, object relations, and overall composition, for enhancing generalizability. But extended versions may introduce overstatements, potentially confusing the network. Thus, we use single sentence embeddings that offer proper yet unexaggerated details. We further discuss this in Supp (Sec.4.2, 4.3).

Analysis of Geometric Cues. To capture geometric cues under limited resources, it is crucial to guide the network with practical spatial information. In Tab. 7, we examine strategies to enhance geometrical knowledge from a single image. Our base transformer network, called Baseline, concatenates depths with an image to extract depth-conditioned features. We first evaluate using only the image, excluding depth concatenation, and observe drops in overall scores. This highlights the meaningful role of the geometric condition. Then, we replace the depth concatenation in the Baseline with unprojected 3D point concatenation. While using 3D points yields slight gains, there is no significant benefit over depth. Beyond simple concatenation, we employ attention strategies to integrate geometric cues seamlessly. We finally observe that cross-attention with point features greatly contributes to comprehensive 3D understanding. These validate the efficacy of our 3D spatial guidance incorporation.

4.4. Visual Comparisons

Qualitative Analysis. In Fig. 6, we qualitatively compare rendered images from ours and Flash3D [44], along with ground truth for solid comparisons. In Scene 1 (*chair*) and 2 (*sink*), ours achieves more precise object placement with less blurriness compared to Flash3D. Also, in Scene 3 (*stair*), CATSplat clearly represents a low-texture area,

whereas Flash3D struggles with blotchy artifacts. Moreover, ours outperforms Flash3D in cross-dataset scenarios. In Scene 4 and 5, ours captures well-defined edges; in Scene 6, ours renders a more detailed image from an aerial view of the complex cityscape. In addition to comparing rendered RGBs, we assess the quality of 3D Gaussians with corresponding depth maps (Supp Sec.4.6). These findings confirm the significance of our priors for novel view synthesis.

User Study. In Tab. 8, we validate our method through human evaluation. We randomly selected 60 and 20 scenes from the RE10K [64] and ACID [26] datasets, and recruited 100 participants via Amazon Mechanical Turk. We present two types of questions with rendered images: (i) preferring between ours and Flash3D [44] based on performance, and (ii) rating the visual quality on a 7-point Likert scale. For all evaluations, ours strongly outperforms Flash3D by a significant margin across both datasets. Also, the narrow confidence interval highlights the consistency of these results.

Method	RE10K [64]		ACID [26]	
	Preference (%)	Likert \uparrow	Preference (%)	Likert \uparrow
Flash3D [44]	11.58 \pm 1.09	4.56 \pm 0.30	8.59 \pm 0.63	4.14 \pm 0.21
CATSplat (Ours)	88.42\pm1.09	6.04\pm0.22	91.41\pm0.63	5.27\pm0.18

Table 8. User study comparisons. We report mean preference percentage and a 7-point Likert scale with a 95% confidence interval.

5. Conclusion

We introduce CATSplat, a novel generalizable 3DGS framework using a single-view image. Our core objective is to transcend the constraints of relying on a single image. To this end, we propose two priors: (i) contextual priors from VLM text embeddings towards context-aware 3D scene reconstruction, and (ii) spatial priors from 3D point features for comprehensive geometric understanding. Extensive experiments demonstrate the superiority of CATSplat. While our method excels in monocular 3D scene reconstruction, ours might be less effective in occluded or truncated areas. Besides, our current training relies on the RealEstate10K dataset; however, with diverse large-scale datasets, CATSplat would be more suitable for real-world applications.

Acknowledgement

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 38%; Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 25%; Development of sketch-based semantic 3D modeling technology for creating user-centric Metaverse content spaces for indoor spaces, RS-2023-00227409, 10%), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00521602, 25%), Institute of Information & communications Technology Planning & Evaluation (IITP) & ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), 1%; IITP-2025-RS-2024-00436857, 1%), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4
- [2] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2): 4606–4613, 2022. 1
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3
- [4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 3
- [5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 1
- [6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3
- [7] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 1, 2, 3, 5, 6
- [8] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 1
- [9] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 1, 2, 3, 5, 6
- [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 3
- [11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 3
- [12] Anurag Dalal, Daniel Hagen, Kjell G Robbersmyr, and Kristian Muri Knausgård. Gaussian splatting: 3d reconstruction and novel view synthesis, a review. *IEEE Access*, 2024. 1
- [13] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4970–4980, 2023. 6
- [14] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022. 3
- [15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5, 6
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5
- [17] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975, 2023. 3
- [18] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 2

- [19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3
- [20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3, 5
- [21] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [22] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2
- [23] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 3, 4, 5, 6
- [24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3, 4
- [26] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snaveley, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 2, 5, 6, 8
- [27] Fangfu Liu, Hanyang Wang, Weiliang Chen, Haowen Sun, and Yueqi Duan. Make-your-3d: Fast and consistent subject-driven 3d content generation. *arXiv preprint arXiv:2403.09625*, 2024. 1
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3, 4, 7
- [29] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2
- [30] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. *arXiv preprint arXiv:2204.07356*, 2022. 3
- [31] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15692–15701, 2022. 3
- [32] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 1, 2, 3, 5
- [33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [34] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 3
- [35] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7
- [37] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1
- [38] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 1, 3, 4
- [39] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 5
- [40] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 3
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2, 5, 6

- [43] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2
- [44] 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. *arXiv preprint arXiv:2406.04343*, 2024. 1, 2, 3, 4, 5, 6, 8
- [45] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 3, 4, 5, 6
- [46] Andrea Tagliasacchi and Ben Mildenhall. Volume rendering digest (for nerf). *arXiv preprint arXiv:2209.02417*, 2022. 2
- [47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [48] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 3, 5, 6
- [49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [50] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. 1, 2, 3, 5, 6
- [51] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 3, 5
- [52] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9076–9086, 2023. 3, 5, 6
- [53] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 7
- [54] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 2
- [55] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023. 2
- [56] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 1
- [57] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2, 6
- [58] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3
- [59] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 1
- [60] Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. *arXiv preprint arXiv:2408.13770*, 2024. 1, 2, 3
- [61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5
- [62] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3
- [63] Yan Zhang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. User: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*, 2024. 3
- [64] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2, 3, 5, 6, 7, 8
- [65] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 4
- [66] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8806–8817, 2024. 2