C^3-GS: Learning Context-aware, Cross-dimension, Cross-scale Feature for Generalizable Gaussian Splatting

Yuxi Hu[1]
yuxi.hu@tugraz.at
Jun Zhang[1]
jun.zhang@tugraz.at
Kuangyi Chen[1]
kuangyi.chen@tugraz.at
Zhe Zhang[2]
doublez@stu.pku.edu.cn
Friedrich Fraundorfer[1]
friedrich.fraundorfer@tugraz.at

[1] Institute of Visual Computing, Graz University of Technology, Graz, Austria
[2] School of Electronic and Computer Engineering, Peking University, Beijing, China

Abstract

Generalizable Gaussian Splatting aims to synthesize novel views for unseen scenes without per-scene optimization. In particular, recent advancements utilize feed-forward networks to predict per-pixel Gaussian parameters, enabling high-quality synthesis from sparse input views. However, existing approaches fall short in encoding discriminative, multi-view consistent features for Gaussian predictions, which struggle to construct accurate geometry with sparse views. To address this, we propose C^3-GS, a framework that enhances feature learning by incorporating context-aware, cross-dimension, and cross-scale constraints. Our architecture integrates three lightweight modules into a unified rendering pipeline, improving feature fusion and enabling photorealistic synthesis without requiring additional supervision. Extensive experiments on benchmark datasets validate that C^3-GS achieves state-of-the-art rendering quality and generalization ability. Code is available at: https://github.com/YuhsiHu/C3-GS.

1 Introduction

Novel view synthesis is intended to generate realistic images from novel viewpoints given a series of posed images. It is widely used in applications such as AR/VR, digital content creation, and autonomous driving [2, 3, 6, 13]. Neural Radiance Fields (NeRF) [25] achieve impressive results with implicit MLPs but suffer from slow rendering due to dense point querying. 3D Gaussian Splatting (3D-GS) [16] replaces implicit fields with explicit Gaussians, enabling real-time synthesis via efficient splatting. Although standard 3D-GS achieves high-quality, photorealistic rendering, it requires dense input views and lengthy per-scene optimization.

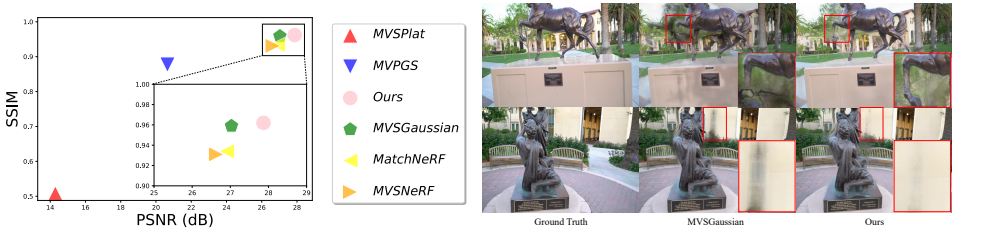

Figure 1: **Comparison with existing methods. Left:** Generalization results on DTU [15] with 3 input views, where our method achieves higher PSNR and SSIM. **Right:** Synthesized images on "Horse" and "Family" scenes from Tanks and Temples [18], highlighting improved visual fidelity.

Recent developments include generalizable methods [4, 7, 23, 29, 33, 40], which are inspired by generalizable NeRF [5, 8, 36], and aim to predict Gaussian parameters via a feed-forward network without further optimization. Early approaches [4, 7] relied on two-view feature matching, while later works [23, 33] incorporated multi-view stereo (MVS) cues to better exploit multi-view geometry. Despite demonstrable progress, challenges still remain in fully utilizing multi-scale feature information for consistent sparse-view geometry constraint, which is essential for predicting accurate Gaussian parameters for high-fidelity rendering. To this end, we propose **C**^{3}-GS, a generalizable GS framework that enhances feature learning across coordinate, spatial, and scale dimensions. Specifically, building upon the baseline MVSGaussian [23], we introduce three novel lightweight modules: (1) Coordinate-Guided Attention (CGA) that embeds coordinate-aware context into 2D features for robust cost matching; (2) Cross-Dimensional Attention (CDA) that constructs 3D spatially-aware descriptors by fusing 2D features with 3D volumes; and (3) Cross-Scale Fusion (CSF) that refines Gaussian opacity across scales to preserve both global structure and fine details.

We validate **C**^{3}-GS on four benchmarking datasets of varying scenarios, demonstrating better rendering performance and generalization capability over the state-of-the-art generalizable rendering methods, as showcased in Fig. 1. Our method excels in challenging cases by capturing fine structures (e.g., horse legs in Fig. 1 (right)) without artifacts. Additionally, our **C**^{3}-GS is capable of predicting more accurate depth maps against the state-of-the-art generalizable methods (Fig. 4, Table 4), thanks to our enhanced feature learning modules.

2 Related Works

Multi-View Stereo (MVS) aims to reconstruct 3D scenes from multiple overlapping images. Traditional methods [10, 11, 27, 28] rely on rule-based features and struggle in complex scenes. Learning-based approaches such as MVSNet [35] build cost volumes from deep features, while cascade frameworks [12, 34, 37, 39] refine depth progressively. Recent transformer-based methods [1, 9, 19, 20] further capture long-range spatial dependencies. In this paper, we build on a novel view synthesis framework [23] that combines a cascade MVS architecture with 3D Gaussian Splatting to generate features and depth information.

Neural Radiance Fields (NeRF) represent scenes as continuous color and density fields using MLPs [25], achieving high-quality view synthesis but requiring costly per-scene training. Generalizable NeRF methods [5, 8, 36] address this by learning feature encoders for 3D points, which are decoded into radiance and density. In particular, MatchNeRF [8] incorporates feature matching to achieve accurate scene representation. [5, 22] leverage cost volume

to better capture multi-view geometry of the scene. Despite their appreciable success, achieving a balance between rendering quality, speed, and generalization remains challenging.

3D Gaussian Splatting (GS) represents scenes with anisotropic Gaussians, enabling real-time rendering via differentiable rasterization, avoiding dense points querying as in NeRF, while still requiring dense input data for per-scene optimization. Inspired by generalizable NeRFs, recent works propose to predict Gaussian parameters in a feed-forward manner. For instance, PixelSplat [4] estimates Gaussians from two views using an epipolar Transformer, which is extended by MVSplat [2] to multi-view with cost volume representation. MVPGS [33] incorporates pre-trained depth estimation models to enhance supervision. Meanwhile, MVSGaussian [23] introduces a hybrid rasterization-volumetric rendering framework, achieving superior performance in rendering quality and efficiency. Despite the success of incorporating multi-view constraints for generalizable rendering, existing methods overlook the enhancement of encoded features in exploiting the intra- and inter-view constraints, resulting in limited generalizability when applied across different datasets. Besides that, the reliance on fixed-view settings and extensive supervision often restricts their flexibility and adaptation to new scenes with varying view counts and configurations. To address these limitations, we extend a generalizable 3D-GS framework [23] with three lightweight modules that enhance feature aggregation and multi-view reasoning, leading to better generalization without additional supervision.

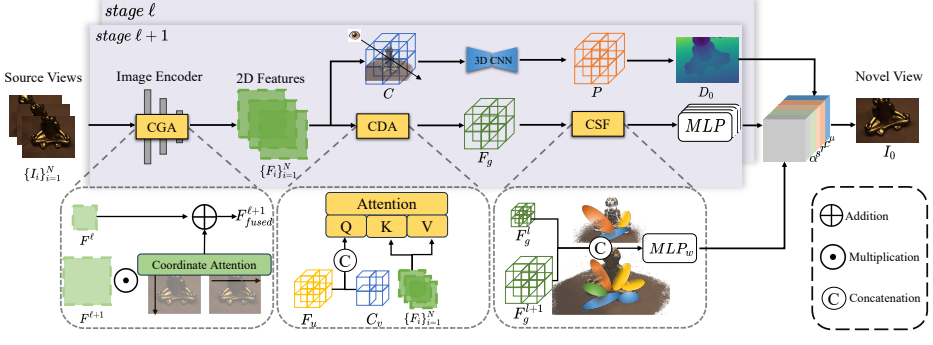

Figure 2: **Overall architecture.** The proposed $\mathbf{C}^3$-GS is a coarse-to-fine framework that estimates depths and Gaussian representations from low resolution (*stage* ℓ) to high resolution (*stage* $\ell+1$). It extracts features $\{\mathbf{F}_i\}_{i=1}^N$ from N source images $\{\mathbf{I}_i\}_{i=1}^N$ using a feature pyramid network and CGA. These features are warped into the target camera frustum planes to construct the 3D cost volume $\mathbf{C}$. Regularization via 3D CNN yields the probability volume $\mathbf{P}$ that is regressed to generate the depth map $\mathbf{D}_0$, which serves as the Gaussian centers after unprojection. CDA is designed to enhance the interaction between 2D features from input views and 3D scene information from the cost volume $\mathbf{C}$ and the compressed features $\mathbf{F}_u$ from 2D features using a U-Net. The enhanced features $\mathbf{F}_g$ from CDA, are decoded to obtain the other Gaussian parameters (i.e., opacity, color, scale and rotation) via lightweight MLPs. By fusing the enhanced features from coarse and fine stages, the CSF produces weights for modifying Gaussian opacity. The output Gaussians are used to render the novel view $\mathbf{I}_0$.

3 Method

In this section, we reveal the details of our proposed framework. Our method is built upon MVSGaussian [23], which estimates 3D Gaussian centers from multi-view stereo (MVS) and decodes Gaussian parameters via MLPs. We contribute three modules for image feature extraction, multi-view feature encoding, and multi-scale Gaussian fusion, enabling improved novel view synthesis without increasing complexity.

Specifically, given a set of posed source views $\{\mathbf{I}_i\}_{i=1}^N$, the goal of our framework is to render a target view I_0 from a novel viewpoint with predicted 3D Gaussians. As illustrated in Fig. 2, we first extract multi-scale 2D features $\{\mathbf{F}_i\}_{i=1}^N$ from source images using a Feature Pyramid Network (FPN) [21]. These features are enhanced by our Coordinate-Guided Attention (CGA) module and then warped to the target camera frustum to construct a cost volume $\mathbf{C}$ via differentiable homography. The cost volumes are passed through a 3D CNN to regularize the features and regress a depth map $\mathbf{D}_0$, which is unprojected to 3D space and used as the centers of pixel-aligned Gaussians. The enhanced features are further fed into our proposed Cross-Dimensional Attention (CDA) module to learn cross-view and context-aware constraints, the output features $\mathbf{F}_g$ are then decoded to obtain the other Gaussian parameters (i.e., opacity, color, scale, and rotation) via lightweight MLPs. To bridge Gaussian representations across different scales, we propose Cross-Scale Fusion (CSF) module to adaptively update Gaussian opacity α in finer stages, improving the rendering quality.

3.1 Image Feature Extraction with Coordinate-Guided Attention

MVSGaussian [23] employs standard FPN for feature extraction. Inspired by [14, 39], we further enhance the expressiveness via introducing the Coordinate-Guided Attention (CGA) module, which models the relative importance of spatial positions between features, by applying two attention maps that capture long-range, feature-channel dependencies.

Specifically, given N input images $\{\mathbf{I}_i\}_{i=1}^N$, we extract their multi-scale features $\{\mathbf{F}_i^\ell\}_{i=1}^N$ using a Feature Pyramid Network (FPN), where ℓ denotes the pyramid level. For clarity, since the same operations are performed independently for each image, we omit the image index i in the following and denote the features at level ℓ by $\mathbf{F}^\ell$. At level ℓ, the feature map $\mathbf{F}^\ell$ has lower resolution, while $\mathbf{F}^{\ell+1}$ has double the height and width. We first compute the weights as $\mathbf{T}_h \in \mathbb{R}^{C\times H\times 1}$ and $\mathbf{T}_w \in \mathbb{R}^{C\times 1\times W}$ according to Eq. 1. Here, H and W represent the height and width of the features, respectively. Given a feature value in the x-th channel, the pooling operations along the height and width at position (h,w) are defined as:

$$\mathbf{T}_h(x,h,1) = \frac{1}{W}\sum_{w=1}^{W}\mathbf{F}(x,h,w)\ ,\mathbf{T}_w(x,1,w) = \frac{1}{H}\sum_{h=1}^{H}\mathbf{F}(x,h,w)\ . \tag{1}$$

Subsequently, the weights $\mathbf{T}_h$ and $\mathbf{T}_w$ are first reshaped and concatenated along the spatial dimension ($cat(\cdot,\cdot)$) to form $\mathbf{T}_{\text{cat}} \in \mathbb{R}^{C\times(H+W)\times 1}$. A convolution layer, followed by a sigmoid activation $\sigma(\cdot)$, is applied to merge the height and width information into a unified representation while normalizing the attention scores:

$$\mathbf{T}_{\text{cat}} = cat(\mathbf{T}_h,\mathbf{T}_w) \in \mathbb{R}^{C\times(H+W)\times 1} \tag{2}$$

$$\mathbf{T}_{\text{attn}} = \sigma\left(\text{Conv1D}(\mathbf{T}_{\text{cat}})\right) \in \mathbb{R}^{C\times(H+W)\times 1}\ . \tag{3}$$

After fusion, the output $\mathbf{T}_{\text{attn}}$ is split to generate two distinct attention maps, $\mathbf{A}_h \in \mathbb{R}^{C\times H\times 1}$ and $\mathbf{A}_w \in \mathbb{R}^{C\times 1\times W}$. Each attention map captures long-range dependencies along one spatial axis while maintaining coordinate cues along the orthogonal direction, enabling more

efficient and interpretable representation learning. The attention maps $\mathbf{A}_h$ and $\mathbf{A}_w$ are broadcasted to align with the shape of the input feature map $\mathbf{F}^{\ell+1}$ and applied multiplicatively to modulate the feature responses along the height and width dimensions:

$$\mathbf{F}_A^{\ell+1} = \mathbf{A}_h \odot \mathbf{A}_w \odot \mathbf{F}^{\ell+1} \,. \tag{4}$$

Meanwhile, the coarser feature map $\mathbf{F}^{\ell}$ from the upper pyramid level is upsampled and combined with the enhanced feature $\mathbf{F}_A^{\ell+1}$ via element-wise addition:

$$\mathbf{F}_{fused}^{\ell+1} = upsamp(\mathbf{F}^{\ell}) \oplus \mathbf{F}_A^{\ell+1} \,, \tag{5}$$

where $upsamp(\cdot)$, $\odot$, and $\oplus$ stand for the upsampling operation, element-wise multiplication (Hadamard product), and element-wise addition, respectively. By constructing 2D features with coordinate-aware long-range encoding, Coordinate-Guided Attention significantly improves global feature discrimination compared to standard convolutional operations, which are limited to local receptive fields.

3.2 Gaussian Parameter Prediction

Gaussian Center Prediction from MVS After obtaining the coordinated-ware fused features $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ from source images, we follow a standard MVS pipeline [23, 30], to warp multi-scale features from source images into the target view by differentiable homographies, constructing a 3D cost volume $\mathbf{C}$ along fronto-parallel planes. The cost volume is processed by a 3D CNN to predict a depth probability volume $\mathbf{P}$, from which we regress the final depth map $\mathbf{D}_0 \in \mathbb{R}^{H\times W\times 1}$, which are back-projected to 3D space with the target camera pose and served as the centers of new generated Gaussians, same as in MVSGaussian [23].

Feature Descriptor Construction with Cross-Dimensional Attention To encode multi-view features for a complete Gaussian representation, [23] directly aggregates them through a simple pooling network, followed by a 2D U-Net for spatial enhancement. However, such encoding features mainly capture view-wise information and lack awareness of 3D context consistency, leading to discontinuities in feature representation. To address these limitations, we propose Cross-Dimensional Attention (CDA) that jointly integrates 2D feature details and 3D volumetric consistency.

In particular, we first extract per-pixel voxel features $\mathbf{C}_v \in \mathbb{R}^{HW\times 8}$ from the cost volume $\mathbf{C}$ through grid sampling, where HW denotes the number of pixel locations after flattening the spatial dimensions. Simultaneously, we extract per-pixel image features across multiple source views, augmented with RGB values and ray direction differences, resulting in a feature tensor $\mathbf{F}_{img} \in \mathbb{R}^{HW\times N\times(C+3+4)}$, where N denotes the number of source views, C denotes the number of feature channels, the additional leading 3 channels correspond to RGB colors, and the trailing 4 channels encode ray direction differences.

To aggregate multi-view information at each pixel independently, we follow [23] to apply a lightweight U-Net Ψ_u along the source view dimension, compressing $\mathbf{F}_{img}$ into a compact feature $\mathbf{F}_u \in \mathbb{R}^{HW\times 16}$:

$$\mathbf{F}_u = \Psi_u(\mathbf{F}_{img}) \,. \tag{6}$$

We then concatenate the voxel consistency feature $\mathbf{C}_v$ and the aggregated image feature $\mathbf{F}_u$ to form a combined representation:

$$\mathbf{F}_c = cat(\mathbf{F}_u, \mathbf{C}_v) \,, \tag{7}$$

where $\mathbf{F}_c \in \mathbb{R}^{HW\times 24}$. This fused feature integrates both fine-grained 2D appearance information and multi-view geometric consistency, enhancing the robustness of 3D representation.

Finally, we leverage $\mathbf{F}_c$ as the query input to an attention-based aggregation module Ψ_t, which retrieves and integrates 2D features across multiple views:

$$\mathbf{F}_g = \Psi_t(\mathbf{F}_c, \{\mathbf{F}_i\}_{i=1}^N), \tag{8}$$

where $\mathbf{F}_c$ serves as the query, and $\{\mathbf{F}_i\}_{i=1}^N$ provide the key and value sequences. This cross-view attention mechanism enables the encoded features to capture both spatially consistent 3D geometry and rich appearance details, leading to improved spatial-aware reconstruction.

Gaussian Attribute Prediction with Cross-Scale Fusion The encoded features $\mathbf{F}_g$ are used to predict the parameters of each Gaussian. Specifically, each Gaussian is defined by a set of attributes, $\{\mu, s, r, \alpha, c\}$, where μ denotes the center, s the scale, r the rotation, α the opacity, and c the color. As mentioned before, the center μ is determined by back-projecting pixel locations according to the regressed depth map $\mathbf{D}_0$. The remaining parameters are decoded from $\mathbf{F}_g$ as:

$$\begin{cases} s = Softplus(MLP_s(\mathbf{F}_g)) \\ r = Norm(MLP_r(\mathbf{F}_g)) \\ \alpha = Sigmoid(MLP_\alpha(\mathbf{F}_g)) \\ c = Sigmoid(MLP_c(\mathbf{F}_g)) \end{cases}, \tag{9}$$

where MLP_s, MLP_r, MLP_α, and MLP_c are dedicated heads for estimating scale, rotation, opacity, and color, respectively.

Previous works [23, 33] adopt cascaded multi-view stereo designs to predict depth maps from coarse to fine. However, they produce fixed-resolution images with single-scale features, without explicitly modeling the relationships across different scales of Gaussians. To address this, we propose the Cross-Scale Fusion (CSF) module, which modulates the opacity of Gaussians across scales to enhance consistency.

Specifically, given the Gaussian features from two adjacent stages, $\mathbf{F}_g^\ell$ and $\mathbf{F}_g^{\ell+1}$, we first upsample $\mathbf{F}_g^\ell$ to match the spatial resolution of $\mathbf{F}_g^{\ell+1}$, and then concatenate them along the feature dimension:

$$\mathbf{F}_g^{\ell+1} = cat(upsamp(\mathbf{F}_g^\ell), \mathbf{F}_g^{\ell+1}), \tag{10}$$

where $upsamp(\cdot)$ denotes spatial interpolation, and $cat(\cdot,\cdot)$ denotes feature concatenation. The concatenated feature is passed through a lightweight MLP, MLP_w, to produce a modulation weight $\mathbf{w}^{\ell+1}$:

$$\mathbf{w}^{\ell+1} = MLP_w(\mathbf{F}_g^{\ell+1}), \tag{11}$$

which is used to adjust the opacity by element-wise multiplication:

$$\alpha = \alpha \odot \mathbf{w}^{\ell+1}. \tag{12}$$

With the updated opacity α and the other Gaussian attributes $\{\mu, s, r, c\}$, novel views $\mathbf{I}_0$ can be rendered via rasterization. Although in principle all Gaussian parameters could be refined across scales, we empirically find that updating only the opacity achieves a good balance between performance and computational efficiency.

3.3 Loss Function

Following [23], our model is trained using images as the only source of supervision. We combine a standard pixel-wise error (MSE) $\mathcal{L}_{\text{pixel}}$ with additional perceptual [58] (VGG16) $\mathcal{L}_{\text{feature}}$ and structural similarity [51] (SSIM) $\mathcal{L}_{\text{structure}}$ losses to improve generalization. Our framework consists of two stages, corresponding to a coarse and a fine reconstruction level. The overall loss is computed as the weighted sum of the losses at each stage:

$$\mathcal{L}_{\text{total}} = \sum_{\ell} \gamma^{(\ell)} \left(\mathcal{L}_{\text{pixel}} + \beta_s \cdot \mathcal{L}_{\text{structure}} + \beta_p \cdot \mathcal{L}_{\text{feature}} \right), \tag{13}$$

where $\mathcal{L}_{\text{total}}$ represents the cumulative loss, and $\gamma^{(\ell)}$ denotes the importance weight of the ℓ-th stage loss. The loss at each stage consists of three components: a pixel-wise term $\mathcal{L}_{\text{pixel}}$, a structure-based term $\mathcal{L}_{\text{structure}}$ scaled by β_s, and a feature-based term $\mathcal{L}_{\text{feature}}$ scaled by β_p.

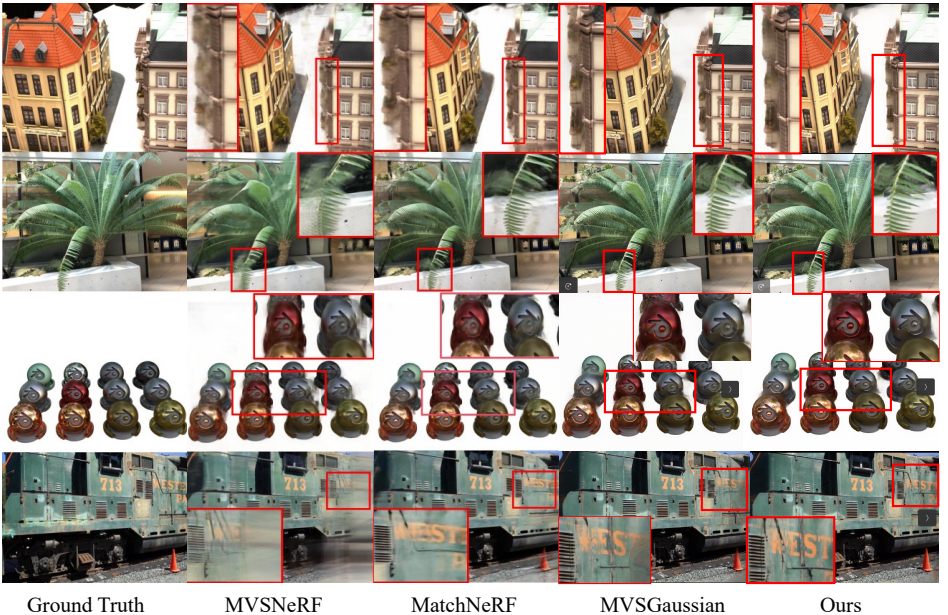

Figure 3: **Qualitative comparison with state-of-the-art methods [6, 8, 23] using 3 input views on DTU [15], Real Forward-facing [24], NeRF Synthetic [25], and Tanks and Temples [18], arranged top to bottom.**

4 Experiments

4.1 Experimental Setup

Datasets. As in [6, 23], we train our model using the DTU training set [15] and evaluate on the DTU test set, applying the same dataset split configuration used in [23]. We evaluate our model directly on the Real Forward-facing [24], NeRF Synthetic [25], and Tanks and Temples [18] datasets, without further training. For each scene, we select the same views as target views as established in [23], and the remaining images may serve as source views

based on their distance from the target view. The quality of synthesized views is measured by broadly adopted PSNR, SSIM [31], and LPIPS [38] metrics.

Implementation Details Our framework is built on MVSGaussian [23], which employs a two-stage cascaded framework. Following [23], we train the model with 3 source views and additionally test its performance with a varying number of input views (2, 3, 4). For depth estimation, we sample 64 and 8 depth hypotheses for the coarse and fine stages, respectively. For loss function, we set $\beta_s = 0.1$, $\beta_p = 0.05$, $\gamma^{(1)} = 0.5$ and $\gamma^{(2)} = 1$ in Eq. 13. The model is trained using the Adam optimizer [17] on a single RTX 4090 GPU.

Method	2-view			3-view		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MVSNeRF [6]	24.03	0.914	0.192	26.63	0.931	0.168
MatchNeRF [6]	25.03	0.919	0.181	26.91	0.934	0.159
PixelSplat [4]	14.01	0.662	0.389	-	-	-
MVSplat [4]	13.94	0.473	0.385	14.29	0.508	0.370
MVPGS [53]	17.55	0.823	0.147	20.65	0.877	0.099
MVSGaussian [23]	23.83	0.918	0.121	27.03	0.959	0.084
Ours	24.27	0.936	0.113	27.87	0.962	0.077

Table 1: **Quantitative results of generalization on DTU [15].**

Baselines. We compare our framework with state-of-the-art generalizable Gaussian methods [4, 7, 23, 33], as well as generalizable NeRF methods [5, 8]. We use their reported results if available. However, due to the difference between pipelines and experimental settings (*i.e.,* views, datasets, and resolutions), some results are not available. In such cases, we generate the results using their released code and pre-trained model on other datasets. Note that we also re-train our baseline model (MVSGaussian [23]) from scratch on our machine to eliminate environmental differences for fair comparison. For all tables, we use color to indicate rankings: red for first place, orange for second, and yellow for third.

Method	Views	Real Forward-facing [24]			NeRF Synthetic [25]			Tanks and Temples [18]		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MVSNeRF [6]	2	20.22	0.763	0.287	20.56	0.856	0.243	18.92	0.756	0.326
MatchNeRF [6]		20.59	0.775	0.276	20.57	0.864	0.200	19.88	0.773	0.334
MVSplat [4]		15.32	0.370	0.422	12.59	0.665	0.282	16.35	0.490	0.355
MVPGS [53]		18.53	0.607	0.280	-	-	-	16.33	0.687	0.373
MVSGaussian [23]		22.58	0.817	0.192	23.56	0.929	0.122	21.05	0.830	0.222
Ours		22.32	0.818	0.201	23.83	0.930	0.109	21.58	0.832	0.221
MVSNeRF [6]	3	21.93	0.795	0.252	23.62	0.897	0.176	20.87	0.823	0.260
MatchNeRF [6]		22.43	0.805	0.244	23.20	0.897	0.164	20.80	0.793	0.300
MVPGS [53]		19.91	0.696	0.229	-	-	-	18.60	0.607	0.467
MVSGaussian [23]		23.87	0.856	0.169	25.31	0.943	0.098	22.70	0.872	0.187
Ours		23.97	0.857	0.167	26.14	0.946	0.079	23.30	0.879	0.178
MVPGS [53]	4	21.28	0.750	0.180	-	-	-	20.27	0.679	0.371
MVSGaussian [23]		24.01	0.864	0.175	24.81	0.938	0.106	21.85	0.860	0.209
Ours		24.16	0.867	0.174	25.29	0.941	0.087	22.35	0.866	0.204

Table 2: **Quantitative results of generalization on Real Forward-facing [24], NeRF Synthetic [25], and Tanks and Temples [18] datasets.**

Method	Views	Real Forward-facing [24]			NeRF Synthetic [25]			Tanks and Temples [18]		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
PixelSplat [4]	2	22.99	0.810	0.190	15.77	0.755	0.314	19.40	0.689	0.223
MVSGaussian [23]		22.87	0.823	0.159	24.00	0.928	0.121	21.81	0.842	0.176
Ours		22.87	0.825	0.156	24.51	0.929	0.098	22.34	0.847	0.172

Table 3: **Comparison with PixelSplat [4].** Given the high memory requirements of PixelSplat [4], the evaluations are conducted on low-resolution (512×512) images.

4.2 Generalization Results

We first demonstrate the results on the DTU [15] test set. As shown in Fig. 3, our results preserve more scene details with fewer artifacts. Quantitative results in Table 1 show that we achieve the best scores in all settings except for 2-view PSNR. This is mainly because MatchNeRF [8] adopts a pre-trained model [32] that is specially designed to capture accurate optical flow between two input images as a match prior.

To further assess the generalization ability of our method, we conduct experiments on Real Forward-facing [24], NeRF Synthetic [25], and Tanks and Temples [18], with results shown in Fig. 3 and Table 2. Additional visualizations are provided in the supplementary materials. Note that, due to their fixed architectures, we exclude 4-view results for MVSNeRF [6] and MatchNeRF [8] and provide only 2-view results for MVSplat [7]. For PixelSplat [4] with large memory consumption, we compare our method with it in lower-resolution images separately in Table 3. As MVPGS [33] requires per-scene training assisted by monocular depth and masks, we re-train it a fixed number of times for fair comparison.

As demonstrated in Table 2, our method achieves top-ranking performance with 3 and 4 input views. A slight performance drop at 4 views is observed for both our method and MVS-Gaussian [23], likely due to limited view overlap, as discussed in prior MVS studies [26, 39]. Overall, our approach generalizes well without any additional training, benefiting from enhanced feature extraction across 2D and 3D spaces.

4.3 Depth Estimation Results

Following [5, 23], we also evaluate depth map estimation. As illustrated in Fig. 4, our method produces sharper and more accurate depth boundaries, despite being trained without ground-truth supervision. Quantitative comparisons in Table 4 further confirm our superiority. An accurate depth map, as an intermediate output in our model, indicates that our method generates more precise encoded features for Gaussian parameters decoder. These refined features form a solid foundation for the subsequent estimation of Gaussian parameters, enhancing the reliability and quality of the rendered scene.

4.4 Ablation Study

As shown in Table 5, we perform ablation studies on the DTU [15] test set (3-view input) to evaluate the effectiveness of each module. Using only Coordinate-Guided Attention (CGA) yields limited improvements (No.1), while Cross-Dimensional Attention (CDA) significantly boosts PSNR (No.2). Combining CGA and CDA (No.4) further improves SSIM and LPIPS, indicating complementary effects. Cross-Scale Fusion (CSF) alone also enhances results (No.3), and combining it with CGA (No.5) or CDA (No.6) leads to consistent gains. Finally, integrating all three modules (No.7) achieves the best overall performance. Corresponding qualitative comparisons are available in the supplementary material.

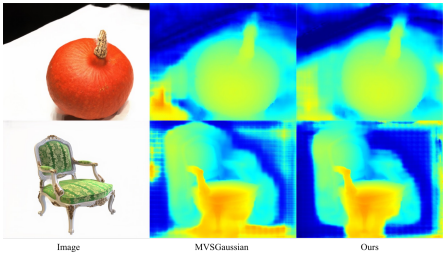

Figure 4: **Qualitative comparison of depth maps with MVSGaussian [23].**

Method	Novel view		
	Abs err ↓	Acc(2)↑	Acc(10)↑
PixelNeRF [34]	47.8	0.039	0.187
IBRNet [30]	324	0.000	0.866
MVSNeRF [6]	7.00	0.717	0.866
MVSGaussian [23]	4.06	0.820	0.936
Ours	3.79	0.832	0.943

Table 4: **Quantitative results of depth estimation on DTU [15].** "Abs err" denotes the mean absolute error (mm), while "Acc(X)" indicates the percentage of pixels with an error below X mm.

	CGA	CDA	CSF	PSNR↑	SSIM↑	LPIPS↓
baseline	-	-	-	27.03	0.959	0.084
No.1	✓	-	-	27.07	0.959	0.084
No.2	-	✓	-	27.34	0.959	0.084
No.3	-	-	✓	27.31	0.960	0.080
No.4	✓	✓	-	27.42	0.960	0.078
No.5	✓	-	✓	27.58	0.960	0.079
No.6	-	✓	✓	27.79	0.961	0.077
No.7	✓	✓	✓	**27.87**	**0.962**	**0.077**

Table 5: **Ablation studies on combination of modules.** The PSNR, SSIM and LPIPS are the metrics on DTU datasets [15]. The baseline method is MVSGaussian [23].

5 Conclusion

In this work, we propose C^3-GS, a generalizable Gaussian Splatting framework that enhances feature encoding with context-aware, cross-dimensional, and cross-scale information, without relying on additional supervisory signals. By leveraging lightweight feature aggregation and flexible multi-view constraints, C^3-GS achieves consistent improvements in cross-dataset generalization over existing state-of-the-art methods. Extensive experiments validate the effectiveness of our approach. However, several important challenges remain unaddressed. First, our current framework does not explicitly handle challenging real-world conditions such as domain shifts across heterogeneous camera systems or degraded inputs caused by motion blur and image noise. Second, our method, like other 3D cost volume–based approaches, may face limitations in wide-baseline scenarios and under view extrapolation, where target views lie outside the range of input observations. Addressing these scenarios can be explored in future work.

Acknowledgements This work was supported by the China Scholarship Council under Grant No. 202208440157.

References

[1] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. *Transactions on Machine Learning Research*, 2023.

[2] Helin Cao and Sven Behnke. DiffSSC: Semantic LiDAR scan completion using denoising diffusion probabilistic models. *arXiv preprint arXiv:2409.18092*, 2024.

[3] Helin Cao and Sven Behnke. SLCF-Net: Sequential LiDAR-camera fusion for semantic scene completion using a 3D recurrent U-Net. In *IEEE International Conference on Robotics and Automation*, pages 2767–2773, 2024.

[4] 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.

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14124–14133, 2021.

[6] Kuangyi Chen, Jun Zhang, and Friedrich Fraundorfer. Evloc: Event-based visual localization in lidar maps via event-depth registration. In *IEEE International Conference on Robotics and Automation*, 2025.

[7] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *Proceedings of the European Conference on Computer Vision*, pages 370–386, 2024.

[8] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[9] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022.

[10] Pascal Fua and Yvan G Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16(1):35–56, 1995.

[11] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[13] Yuxi Hu, Taimeng Fu, Guanchong Niu, Zixiao Liu, and Man-On Pun. 3d map reconstruction using a monocular camera for smart cities. *The Journal of Supercomputing*, 78(14):16512–16528, 2022.

[14] Yuxi Hu, Jun Zhang, Zhe Zhang, Rafael Weilharter, Yuchen Rao, Kuangyi Chen, Runze Yuan, and Friedrich Fraundorfer. Icg-mvsnet: Learning intra-view and cross-view relationships for guidance in multi-view stereo. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2025.

[15] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014.

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

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36 (4):1–13, 2017.

[19] Yuqi Li, Junhao Dong, Zeyu Dong, Chuanguang Yang, Zhulin An, and Yongjun Xu. Srkd: Towards efficient 3d point cloud segmentation via structure-and relation-aware knowledge distillation. *arXiv preprint arXiv:2506.17290*, 2025.

[20] Yuqi Li, Chuangang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. *arXiv:2507.02939*, 2025.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[22] Tianqi Liu, Xinyi Ye, Min Shi, Zihao Huang, Zhiyu Pan, Zhan Peng, and Zhiguo Cao. Geometry-aware reconstruction and fusion-refined rendering for generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7654–7663, 2024.

[23] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 37–53, 2025.

[24] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, 38 (4):1–14, 2019.

[25] 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. In *Proceedings of the European Conference on Computer Vision*, 2020.

[26] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo a unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022.

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[28] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 501–518, 2016.

[29] Stanislaw Szymanowicz, Chrisitian 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.

[30] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[31] 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.

[32] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022.

[33] Wangze Xu, Huachen Gao, Shihe Shen, Rui Peng, Jianbo Jiao, and Ronggang Wang. Mvpgs: Excavating multi-view priors for gaussian splatting from sparse input views. In *Proceedings of the European Conference on Computer Vision*, pages 203–220, 2024.

[34] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *Proceedings of the European Conference on Computer Vision*, pages 674–689. Springer, 2020.

[35] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 767–783, 2018.

[36] 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*, 2021.

[37] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020.

[38] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[39] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21508–21518, 2023.

[40] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19680–19690, 2024.