Epipolar-Free 3D Gaussian Splatting for Generalizable Novel View Synthesis

Zhiyuan Min[1] **Yawei Luo**[1,*] **Jianwen Sun**[2] **Yi Yang**[1]
[1]Zhejiang University [2]Central China Normal University

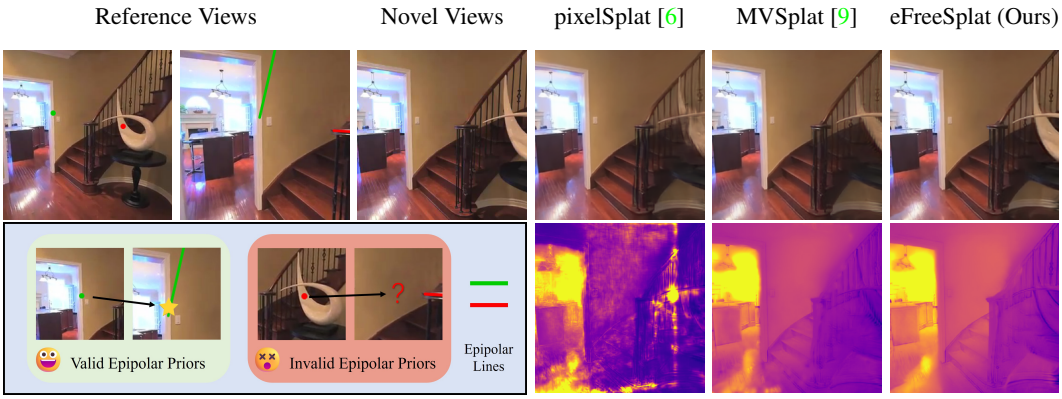

Fig. 1. Epipolar priors can be unreliable across extremely sparse views, especially in non-overlapping or occluded areas. Our model, eFreeSplat, generalizes to novel scenes without relying on epipolar priors, offering superior appearance and geometric perception.

Abstract

Generalizable 3D Gaussian splitting (3DGS) can reconstruct new scenes from sparse-view observations in a feed-forward inference manner, eliminating the need for scene-specific retraining required in conventional 3DGS. However, existing methods rely heavily on epipolar priors, which can be unreliable in complex real-world scenes, particularly in non-overlapping and occluded regions. In this paper, we propose eFreeSplat, an efficient feed-forward 3DGS-based model for generalizable novel view synthesis that operates independently of epipolar line constraints. To enhance multiview feature extraction with 3D perception, we employ a self-supervised Vision Transformer (ViT) with cross-view completion pre-training on large-scale datasets. Additionally, we introduce an Iterative Cross-view Gaussians Alignment method to ensure consistent depth scales across different views. Our eFreeSplat represents an innovative approach for generalizable novel view synthesis. Different from the existing pure geometry-free methods, eFreeSplat focuses more on achieving epipolar-free feature matching and encoding by providing 3D priors through cross-view pretraining. We evaluate eFreeSplat on wide-baseline novel view synthesis tasks using the RealEstate10K and ACID datasets. Extensive experiments demonstrate that eFreeSplat surpasses state-of-the-art baselines that rely on epipolar priors, achieving superior geometry reconstruction and novel view synthesis quality. Project page: https://tatakai1.github.io/efreesplat/.

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1251

1 Introduction

Rendering novel views from sparse observations has long been a challenging research task in the 3D vision community. Recently, generalizable novel view synthesis (GNVS) techniques have emerged as a promising solution. These models, trained on large-scale multiview datasets, can directly synthesize novel views of new scenes from a few observations, eliminating the need for scene-specific retraining. Notable works in this vein include NeRF-based GNVS [37, 38, 55, 64] and Light Field Network-based GNVS [12, 48, 49]. An enabling factor in their generalizability is the use of epipolar priors, which help determine the precise location of a pixel in one image on the corresponding epipolar line in another viewpoint [17, 70]. More recently, generalizable 3D Gaussian splatting methods, such as pixelSplat [6] and MVSplat [9], have been proposed. These methods leverage the benefits of a primitive-based 3D representation, offering fast and memory-efficient rendering along with an interpretable 3D structure for generalizable view synthesis. Like previous approaches, most 3DGS-based GNVS methods [6, 9, 61] depend on epipolar priors to achieve high-quality and fast cross-scene novel view rendering.

Despite significant advancements utilizing epipolar priors, a new and underexplored issue has emerged in GNVS: epipolar priors prove unreliable in non-overlapping and occluded regions of complex real-world scenes, where corresponding points on epipolar lines are absent. As depicted in Fig. 1, epipolar lines (marked in green) effectively identify geometric correspondences in multiview overlapping areas. Conversely, epipolar lines (marked in red) become invalid in those non-overlapping regions, leading to unreliable geometric reconstructions. Moreover, sampling on invalid epipolar lines and employing attention mechanism will produce a lot of redundant calculations [6, 38, 49].

A newly proposed geometry-free 3D reconstruction method [56], which captures multiview consistent knowledge from a versatile model pre-trained on cross-view data, has inspired our development of a novel GNVS method that circumvents the dependence on epipolar priors through data-driven 3D priors. Leveraging this insight, we propose eFreeSplat, an efficient feed-forward 3D Gaussian Splatting model for GNVS that operates independently of epipolar line priors. eFreeSplat is built upon 3DGS [23] originally designed for single-scene NVS and extends its advantages to GNVS. The overview of our method is illustrated in Fig. 2. To capture 3D structural information across sparse views without unreliable epipolar priors, we utilize a self-supervised pre-training model for 3D cross-view completion [59, 60]. This model uses a Vision Transformer (ViT) [11] encoder and cross-attention decoder to predict parts of the masked images from reference views. In eFreeSplat, the pre-training model retains all patches, effectively capturing spatial relationships and acting as a "*cross-view mutual perceiver*". This approach provides robust geometric biases for global 3D representation via cross-view completion pre-training on large-scale datasets [25, 35, 40, 44, 45].

Experimentally, we found that without an explicit 3D constraint, the scale of predicted depth maps of per-pixel 3D points from different views tends to be inconsistent [4, 53], leading to artifacts or pixel displacement in images from novel views. To address the issue of inconsistent depth scales across different views, we introduce an Iterative Cross-view Gaussians Alignment (ICGA) technique to eFreeSplat. ICGA is based on the fact that the features of most surface points projected onto the camera planes of different views remain consistent. Specifically, we obtain the warped features for each view based on the predicted depths via U-Net. We then calculate the fine depths for the next iteration via the correlation between the warped features and the features from other views. Unlike the plane-sweep stereo approach [9, 62, 63], our updating and alignment strategy does not require numerous depth candidates, thereby reducing computational and storage costs.

The main contributions of this paper are summarized as follows:

- We introduce eFreeSplat, a method with novel insights into GNVS that operates without relying on epipolar priors in the process of multi-view geometric perception. eFreeSplat demonstrates robustness in generalizing to new scenarios with sparse and non-overlapping observations.
- To ensure depth scale consistency across different viewpoints without explicit epipolar constraints, we propose an Iterative Cross-view Gaussians Alignment method, which alleviates artifacts and pixel displacement issues in renderings.
- eFreeSplat achieves competitive cross-scene rendering performance on the RealEstate10K [72] and ACID [26] datasets, surpassing state-of-the-art approaches such as pixelSplat [6] and MVSplat [9].

https://doi.org/10.52202/079017-1251

2 Related Work

Single-Scene 3DGS. 3D Gaussian Splatting (3DGS) [23] marks a significant shift in 3D scene representation. It employs millions of learnable 3D Gaussians to explicitly map spatial coordinates to pixel values, enhancing rendering efficiency and quality via a rasterization-based splatting approach, and boosting various downstream tasks [34, 36]. Unlike early 3D neural representation methods [37, 39, 46] that require intensive computations and large memory usage (*e.g.*, neural fields [2, 3, 67] and volume rendering [27, 65, 66]) , 3DGS enables real-time rendering and editability with minimized computational demands [8]. Existing single-scene 3DGS-liked methods [10, 18, 23] demand dense views for each scene via the expensive per-scene gradient back-propagation process. In our work, we employ a single feedforward network to deduce the parameters of Gaussian primitives using merely two images.

Cross-Scene Generalizable 3DGS. Cross-scene generalizable 3DGS learns robust priors from large-scale scenarios to predict Gaussian primitive parameters and render novel view images using sparse inputs. pixelSplat [6] and LatentSplat [61] leverage the epipolar transformer [17] to find cross-view correspondences and learn per-pixel Gaussian depth distributions. However, this can fail in non-overlapping and occluded areas, leading to inaccurate geometry and surface reconstructions. Splatter Image [50] merges Gaussian primitives from single-view regressions but lacks cross-view information, limiting its multiview applications. GPS-Gaussian [71] and MVSplat [9] improve feature matching with cost volumes for better geometries; however, GPS-Gaussian is limited to human body reconstruction with depth ground truth, and MVSplat, using plane-sweep stereo [28, 29, 62, 63], still relies on the epipolar priors [13, 15, 63]. Triplane-Gaussian [73] encodes single-view images into latent 3D point clouds and triplane features, outputting 3D Gaussian properties via MLP decoders. However, it focuses on single-view reconstruction, with rendering quality dependent on initial geometry. Our method bypasses 3D priors through sampling along epipolar lines or cost volumes, instead using cross-view competition pre-training [59, 60] on large-scale datasets [25, 35, 40, 44, 45].

Solving 3D Tasks using Geometry-free Methods. Priors are crucial for visual tasks to provide generalized features [14, 30, 31, 32, 33]. Capitalizing on the geometric priors, methods based on re-projection features [21, 51, 64], cost volume [7, 19, 22, 63], and image warping [5] have performed well in downstream 3D activities. However, these methods rely on task-specific designs and struggle with complex scenarios, such as occlusions or non-overlapping views. Recently, some geometry-free alternatives have been proposed to this challenge. SRT [43] and GS-LRM [68] are epipolar-free GNVS methods that boldly eschew any explicit geometric inductive biases. SRT encodes patches from all reference views using a Transformer encoder and decodes the RGB color for target rays through a Transformer decoder. GS-LRM's network, composed of a large number of Transformer blocks, implicitly learns 3D representations. However, due to the lack of targeted scene encoding, these methods are either limited to specific datasets or suffer from unacceptable computational efficiency and carbon footprint. Some pose-free GNVS methods [20, 43, 54] are also epipolar-free. These methods, lacking known camera poses, find it challenging to perform epipolar line sampling. They often reduce task complexity through specially designed feature representations (e.g., Learned 3D Neural Volume in LEAP [20] and Triplane in PF-LRM [54]), but this reduction comes at the cost of decreased model generalization. Different from the above methods, our method focuses on data-driven 3D priors and does not require any time-consuming and complex structured feature representations, such as cost volumes. CroCo [59], a self-supervised pre-training method for 3D vision tasks, uses cross-view completion to recover occluded parts of an image from different viewpoints without any 3D inductive biases, significantly enhancing downstream 3D vision tasks. DUSt3R [56] introduces a novel paradigm for dense and unconstrained stereo 3D reconstruction from arbitrary image collections, operating without prior information about camera calibration. These geometry-free pioneers pave the way for more adaptable and efficient 3D vision systems capable of performing accurately across diverse and challenging environments.

3 Methodology

3.1 Overview

Our objective is to predict per-pixel 3D Gaussian [23] primitives $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \boldsymbol{SH}_i\}_{i=1}^{M}$ using N reference views images $\{\boldsymbol{I}_j\}_{j=1}^{N}$, camera intrinsics matrices $\{\boldsymbol{K}_j\}_{j=1}^{N}$ and poses matrices $\{\boldsymbol{P}_j\}_{j=1}^{N}$

https://doi.org/10.52202/079017-1251

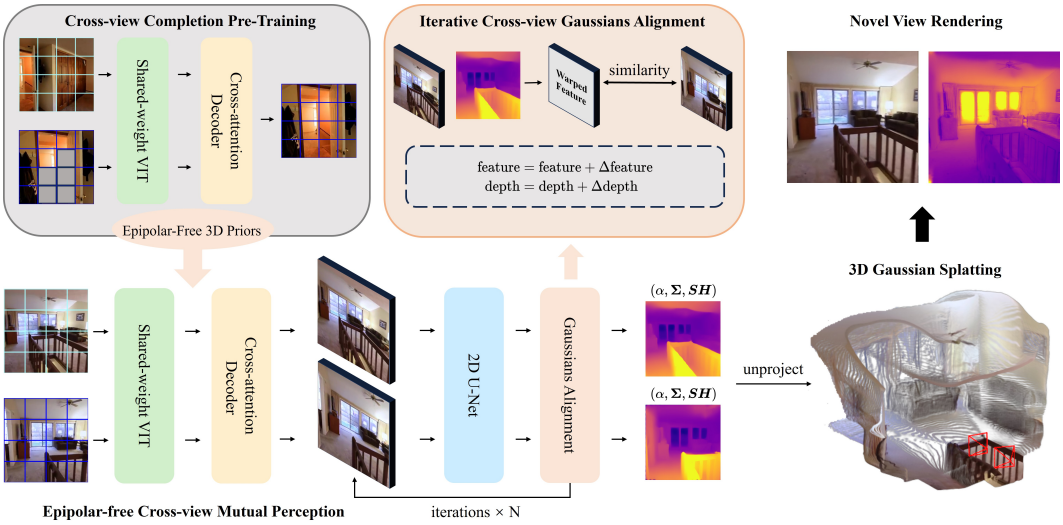

Fig. 2. Overview of eFreeSplat. (a) Epipolar-free Cross-view Mutual Perception leverages self-supervised cross-view completion pre-training [60] to extract robust 3D priors. The ViT [11] with shared weights processes the reference images, followed by a cross-attention decoder to generate multiview feature maps, forming 3D perception without epipolar priors. (b) Iterative Cross-view Gaussians Alignment module iteratively refines Gaussian attributes through a 2D U-Net. The process involves warped features to align corresponding features and depths, ensuring consistent depth scales across different views. (c) The final step involves employing rasterization-based volume rendering [23] to generate high-quality geometry and realistic novel view images.

in a single feed-forward inference. The 3D Gaussian primitives include position $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$, opacity α, and spherical harmonics $\boldsymbol{SH}$ for colors. Given a $H \times W$ sized reference image, the number of 3D Gaussian primitives can be calculated as $M = N \times H \times W$. The position of the 3D Gaussians $\boldsymbol{\mu}_i$ determines the geometric shape of the scene, which corresponds to pixel $\mathbf{u}$ is calculated using the camera origin $\boldsymbol{o}$, the ray direction $\boldsymbol{d}_{\mathbf{u}}$, and the predicted depth d:

$$\boldsymbol{\mu}_i = \boldsymbol{o} + d \cdot \boldsymbol{d}_{\mathbf{u}}, \tag{1}$$

where $\boldsymbol{d}_u$ is calculated by the camera intrinsic and pose matrix: $\boldsymbol{d}_{\mathbf{u}} = \boldsymbol{P}\boldsymbol{K}^{-1}[\mathbf{u}, 1]^T$. However, when the number of reference views is extremely sparse, predicting accurate depths d and reconstructing high-quality geometric structures and appearances become particularly challenging. Particularly in non-overlapping and occluded areas, prevalent methods [6, 9, 12] based on epipolar line sampling fail to introduce valid geometric priors.

In this paper, we propose eFreeSplat, a generalizable 3D Gaussian Splatting model from sparse reference views[2] that operates independently of epipolar line priors. As illustrated in Fig. 2, the pre-trained ViT model based on cross-view completion via self-supervised training [59, 60] in large-scale datasets provides robust geometric priors, serving as our Epipolar-free Cross-view Mutual Perception (Sec. 3.2). Unlike recent works [6, 9, 61], which directly combine per-view 3D Gaussians, we propose Iterative Cross-view Gaussians Alignment (ICGA) in Sec. 3.3. This module iteratively updates the position and features of Gaussians by calculating the similarity between warped features and corresponding features, alleviating the issues of local geometric inaccuracies caused by inconsistent depth scales. In Sec. 3.4, we predict the centers of the 3D Gaussians by unprojecting the aligned depth maps while calculating other 3D Gaussian parameters based on the aligned features.

3.2 Epipolar-free Cross-view Mutual Perception

To realize the cross-view mutual preception without relying on the epipolar prior, we extract cross-view image features using a shared-weight ViT $\mathcal{E}_{\boldsymbol{\theta}_1}$ and a cross-attention decoder $\mathcal{D}_{\boldsymbol{\theta}_2}$, both pre-trained on large-scale cross-view completion tasks in a self-supervised manner [60]. Following

[2]In our experiments, the number of reference views $N = 2$, which is consistent with previous methods [6, 9]. For convenience, all subsequent discussions will assume a 2-views input scenario.

https://doi.org/10.52202/079017-1251

the methodologies of CroCo v2 [60] and ViT [11], both images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are divided into $2n$ non-overlapping patches via a linear projection, with each patch measuring 16×16 pixels. Additionally, relative positional embeddings [47] are added to the RGB patches before inputted into a series of stacked Transformer modules for encoding tokens $\boldsymbol{\varepsilon}_j$:

$$\{\boldsymbol{\varepsilon}_j\}_{j=1}^{2} = \{\mathcal{E}_{\boldsymbol{\theta_1}}(\boldsymbol{I}_j)\}_{j=1}^{2}. \tag{2}$$

After encoding $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ via ViT independently, the cross-attention decoder $\mathcal{D}_{\boldsymbol{\theta_2}}$ takes $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ conditioned on each other for cross-view features $\boldsymbol{\mathcal{F}}_j \in \mathbb{R}^{C\times H\times W}$:

$$\boldsymbol{\mathcal{F}}_1 = f\left(\mathcal{D}_{\boldsymbol{\theta_2}}\left(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\right)\right), \quad \boldsymbol{\mathcal{F}}_2 = f\left(\mathcal{D}_{\boldsymbol{\theta_2}}\left(\boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_1\right)\right). \tag{3}$$

The structure of the cross-attention decoder consists of alternating multi-head self-attention blocks and multi-head cross-attention blocks. The mapping function f refers to unflattening the tokens back to the original image size. The multi-head self-attention blocks learn token representations from the first viewpoint, while the multi-head cross-attention blocks facilitate cross-view information exchange conditioned on the token representations from the second view.

The CroCo model [59, 60], as a variant of masked image modeling [1, 16, 58] that leverages cross-view information from the same scene to capture the spatial relationship between two images, can significantly enhance performance on 3D downstream tasks. Based on cross-view completion self-supervised pre-training on large-scale datasets, our epipolar-Free cross-view mutual perception method provides robust 3D priors information by understanding the spatial relationship between the two images [59]. Due to the randomness of the masking process during pre-training, the pre-trained model is capable of reasoning about non-overlapping and occluded areas, which is hard for traditional geometric methods to achieve. Therefore, our epipolar-Free mutual perception possesses a more global and robust feature-matching inductive bias compared to methods [6, 9, 12, 61] that rely on epipolar line sampling [17] or the plane-sweep stereo approach [63].

3.3 Iterative Cross-view Gaussians Alignment

To address the issue of inconsistent depth scales across different views, we utilize cross-view feature matching information to align and update per-pixel Gaussians' centers and features iteratively.

Firstly, we predict per-pixel Gaussians' depths $\boldsymbol{d}$ and features $\boldsymbol{\mathcal{G}}$ via a 2D U-Net [42] mapping U with cross-view attention, similar to [9]:

$$\boldsymbol{d}_1, \boldsymbol{\mathcal{G}}_1, \boldsymbol{d}_2, \boldsymbol{\mathcal{G}}_2 = U(\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2). \tag{4}$$

Next, to establish cross-view correspondences, we endeavor to make the features of each 3D Gaussian point projected onto the known camera planes to be as similar as possible. Taking the first view as an example, we calculate the warped features $\boldsymbol{\mathcal{G}}_{1,2}$ of the first view on the second view's features map via the predicted coarse depth $\boldsymbol{d}_1$:

$$\mathcal{W}_{1,2} = \boldsymbol{K}_2\boldsymbol{R}_2\left(\boldsymbol{R}_1^{-1} - \frac{\left(\boldsymbol{R}_2^{-1}\mathbf{t}_2 - \boldsymbol{R}_1^{-1}\mathbf{t}_1\right)\mathbf{n}_1^T}{\boldsymbol{d}_1}\right)\boldsymbol{K}_1^{-1}, \tag{5}$$

$$\boldsymbol{\mathcal{G}}_{1,2}(\mathbf{u}) = \boldsymbol{\mathcal{G}}_2(\mathcal{W}_{1,2}[\mathbf{u}, 1]^T), \tag{6}$$

where $\mathcal{W}$ denotes the homographic warping matrix. $\mathbf{u}$ represents a pixel location in the first view. $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ are the rotation and translation parameters of the camera pose $\boldsymbol{P}_i$. $\mathbf{n}_i$ refers to the normal vector of the target plane. We compute the similarity $\boldsymbol{\mathcal{S}}^1, \boldsymbol{\mathcal{S}}^2$ between the warped feature map $\boldsymbol{\mathcal{G}}_{1,2}$ and the corresponding feature map $\boldsymbol{\mathcal{G}}_1$ based on $\Delta\boldsymbol{d}_{\cos}$. $\boldsymbol{\mathcal{S}}^2$ is obtained by the dot product of $\boldsymbol{\mathcal{G}}_1$ and $\boldsymbol{\mathcal{G}}_{1,2}$, where C denotes the feature dimension of the 3D Gaussian primitives.

$$\boldsymbol{\mathcal{S}}^1 = (\boldsymbol{\mathcal{G}}_1 - \boldsymbol{\mathcal{G}}_{1,2})^2, \quad \boldsymbol{\mathcal{S}}^2 = \frac{\boldsymbol{\mathcal{G}}_1 \cdot \boldsymbol{\mathcal{G}}_{1,2}}{\sqrt{C}}. \tag{7}$$

Finally, we update the coarse per-pixel 3D Gaussian features and predicted depths.

$$\Delta\boldsymbol{\mathcal{G}}_1 = \varphi([\,\boldsymbol{\mathcal{G}}_1 \parallel \boldsymbol{\mathcal{S}}^1\,]) \cdot \boldsymbol{\mathcal{S}}^2, \quad \Delta\boldsymbol{d}_1 = \boldsymbol{d}_1 \cdot \boldsymbol{\mathcal{S}}^2, \tag{8}$$

$$\boldsymbol{\mathcal{G}}_1 = \boldsymbol{\mathcal{G}}_1 + \Delta\boldsymbol{\mathcal{G}}_1, \quad \boldsymbol{d}_1 = \boldsymbol{d}_1 + \Delta\boldsymbol{d}_1, \tag{9}$$

https://doi.org/10.52202/079017-1251

Table 1. Quantitative comparisons. We evaluate our method by rendering three novel view images from two reference viewpoints for each scene. The performance is determined by averaging across all scenes. The dataset's training and testing split follows the protocol established by pixelSplat [6]. The inference time includes both scene encoding and rendering time, tested on a single RTX-4090 GPU.

Methods	RealEstate10K [72]			ACID [26]			Inference Time (s)
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
Du et al. [12]	24.78	0.820	0.213	26.88	0.799	0.218	1.578
GPNR [49]	24.11	0.793	0.255	25.28	0.764	0.332	13.180
pixelSplat [6]	25.89	0.858	0.142	28.14	0.839	0.150	0.100
MVSplat [9]	<u>26.39</u>	**0.869**	<u>0.128</u>	<u>28.25</u>	<u>0.843</u>	<u>0.144</u>	**0.046**
eFreeSplat	**26.45**	<u>0.865</u>	**0.126**	**28.30**	**0.851**	**0.140**	<u>0.061</u>

where $[\cdot\|\cdot]$ refers to the concatenation operation of tensors. We employ the mapping function $\varphi : \mathbb{R}^{2C\times H\times W} \mapsto \mathbb{R}^{C\times H\times W}$ through lightweight convolutional blocks.

The updated features and depths serve as inputs for Eq. (4) (5) and (6), bootstrapping the next iteration of Gaussian updates. Our cross-view Gaussians alignment method, during each iteration, involves establishing a match for target pixel $\mathbf{u}_1$ in the first view with matching pixel $\mathbf{u}_2$ in the second view. This process is akin to considering all neighboring pixels of the projected pixel $\mathbf{u}_2'$ based on the current coarse depth due to the locality inductive bias inherent in convolutions. During each querying process, the discrepancy between $\mathbf{u}_2'$ and the true matching $\mathbf{u}_2$ progressively decreases, thereby harmonizing the consistency of depth scales across multiple views.

3.4 Gaussian Parameters Prediction

We calculate the per-view Gaussians' centers $\boldsymbol{\mu}$ based on the refined depths and camera parameters using Eq. (1). We predict additional Gaussian primitives: $\boldsymbol{\Sigma}, \alpha, \boldsymbol{SH}$, via an additional U-Net. Following other 3DGS-based methods [6, 9, 23], the covariance matrix $\boldsymbol{\Sigma}$ is composed of a scaling matrix and a rotation matrix. The spherical harmonic coefficients $\boldsymbol{SH}$ are used to compute RGB values given a direction. Since we have harmonized the depth scale across different viewpoints, we directly merge all views' Gaussian primitives $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \boldsymbol{SH}_i\}_{i=1}^{N\times H\times W}$.

4 Experiments

4.1 Experimental Settings

Datasets. eFreeSplat is trained on RealEstate10K [72] and ACID [26]. The RealEstate10K dataset consists of home tour videos, providing a wealth of scenes and a variety of viewpoint changes. The ACID dataset contains aerial landscape videos, featuring expansive views and complex terrains. Both datasets provide estimated camera parameters. Following pixelSplat [6], we use the provided training and testing splits and evaluate three novel view images on each test scene.

Evaluation Metrics and Training Losses. We employ standard image quality metrics to validate and compare our results quantitatively: pixel-level PSNR, patch-level SSIM [57], and feature-level LPIPS [69]. During the training phase, the loss is composed of a linear combination of MSE and LPIPS loss, with loss weights of 1 and 0.05, respectively. Since existing methods conduct experiments at 256×256, we also set the resolution of our training and testing images for fair comparison.

Comparison Methods. We compared four feed-forward methods for sparse view novel view synthesis. Du et al. [12] and GPNR [49] are the methods based on light field rendering that combines features on epipolar lines aggregated by the epipolar transformer. pixelSplat [6] and MVSplat [9] are the latest 3DGS-based models based on epipolar sampling and multi-plane sweeping, respectively. Our method compared the qualitative and quantitative results with these four methods.

Implementation details. The ViT-B vision transformer [11] and cross-attention decoder [59] have been pretrained by CroCo v2 [60], which underwent self-supervised cross-view completion training on large-scale datasets [25, 35, 40, 44, 45]. The Iterative Alignment and Updating strategy

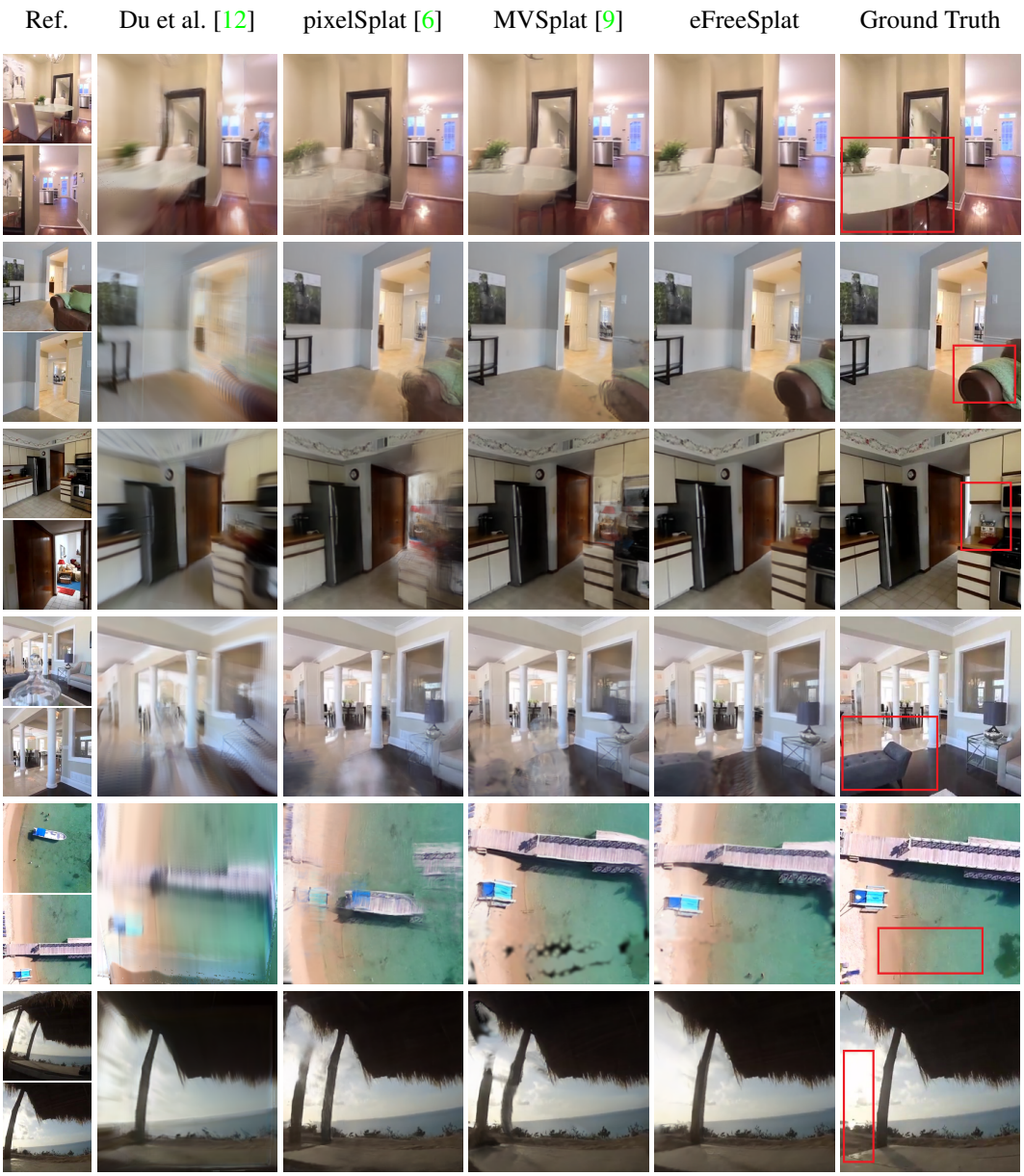

Fig. 3. We provide qualitative comparisons on the RealEstate10K (first four rows) and the ACID (last two rows). Compared to baselines, our method produces fewer artifacts in rendering results (red boxes). Moreover, our approach can perform better in non-overlapping areas (1st, 2nd, 5th and 6th rows) and occluded areas (3th and 4th rows) without relying on unreliable epipolar priors.

is implemented through 2 iterations. All models are trained on 4 RTX-4090 GPUs for $300,000$ iterations using the Adam optimizer [24]. More details are provided in Appendix C.

4.2 Comparative Studies

Image quality comparison. We report quantitative results against baselines [6, 9, 12, 49] on the RealEstate10K and ACID datasets in Tab. 1. Our method, eFreeSplat, outperforms the SOTA method, MVSplat [9] by 0.06dB in PSNR on the RealEstate10K dataset and by 0.05dB on the ACID dataset. The evaluation metrics for all baselines are derived from experimental results published in the papers on pixelSplat [6] and MVSplat [9].

https://doi.org/10.52202/079017-1251

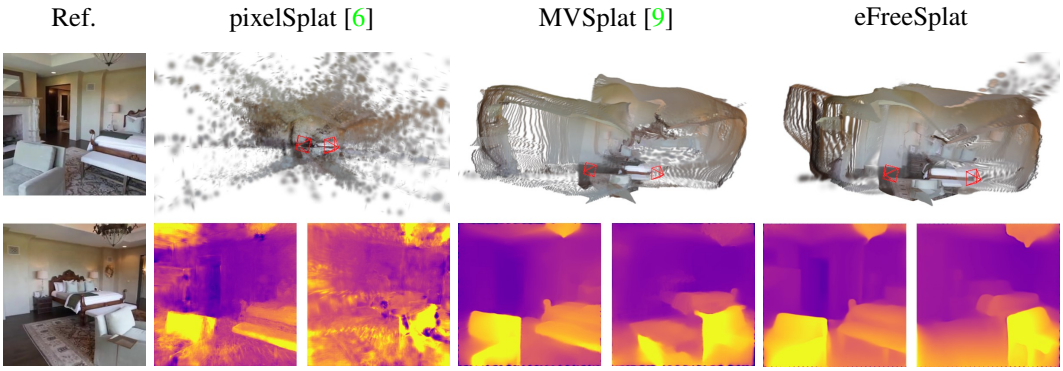

Fig. 4. Comparison results about 3D Gaussians (top) and predicted depth maps of the reference viewpoints (bottom). Compared to SOTA 3DGS-based methods [6, 9], our method achieves higher quality in 3D Gaussian Splatting and produces smoother depth maps.

Table 2. In more challenging scenarios, we classify the RealEstate10K dataset [72] into three subsets based on the overlap size of the reference images: scenes with an overlap below 0.7, 0.6, and 0.5.

Methods	Overlap 0.7			Overlap 0.6			Overlap 0.5		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
pixelSplat [6]	25.05	0.852	0.145	24.79	0.849	0.149	24.96	0.846	0.149
MVSplat [9]	25.11	0.854	0.139	24.70	0.841	0.146	24.64	0.840	0.150
eFreeSplat	**25.72**	**0.861**	**0.132**	**25.48**	**0.859**	**0.135**	**25.46**	**0.853**	**0.139**

Our method's qualitative comparison with baselines is illustrated in Fig. 3. Our rendering results show fewer artifacts or object deformations, especially in non-overlapping or occluded areas. Competitive methods like pixelSplat [64] and MVSplat [9], based on sampling along the epipolar lines, produce unreliable reconstructions in these challenging areas. It demonstrates that eFreeSplat provides more robust 3D priors than epipolar priors, offering global 3D perception even in challenging areas.

Geometry quality comparison. As illustrated in Fig. 4, our method produces higher-quality 3DGS reconstructions and smoother priors without the epipolar priors. pixelSplat [6], despite additional finetuning via depth regularization during training, exhibits noticeable artifacts in its reconstructed 3DGS and depth maps. MVSplat [9] generates competitive depth maps by building a cost volume representation [63], which directly merges per-view Gaussians, resulting in significant point cloud shifts. Our method, which does not rely on sampling along epipolar lines or additional depth regularization finetuning, surpasses current SOTA methods in 3DGS reconstruction quality. Please refer to Appendix A for additional comparison and analysis.

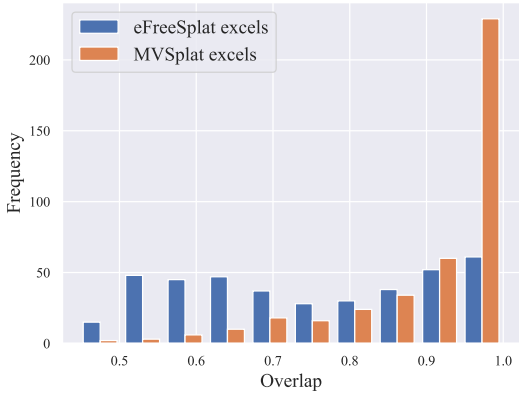

Fig. 5. Our method reconstructs more reliable results than MVSplat when the reference views overlap is low. In the histogram, the blue bars represent the frequency at which our method exceeds MVSplat in rendering quality under the current overlap conditions, while the orange bars indicate the opposite.

Performance with Low-overlapped observations. In this section, we analyze the differences between our method and 3DGS-based methods when the reference viewpoints have a lower overlap. First, we counted the number of scenes where our method and MVSplat [9] outperform each other in PSNR on the RealEstate10K dataset, selecting the top 400 scenes with the largest PSNR differences for

https://doi.org/10.52202/079017-1251

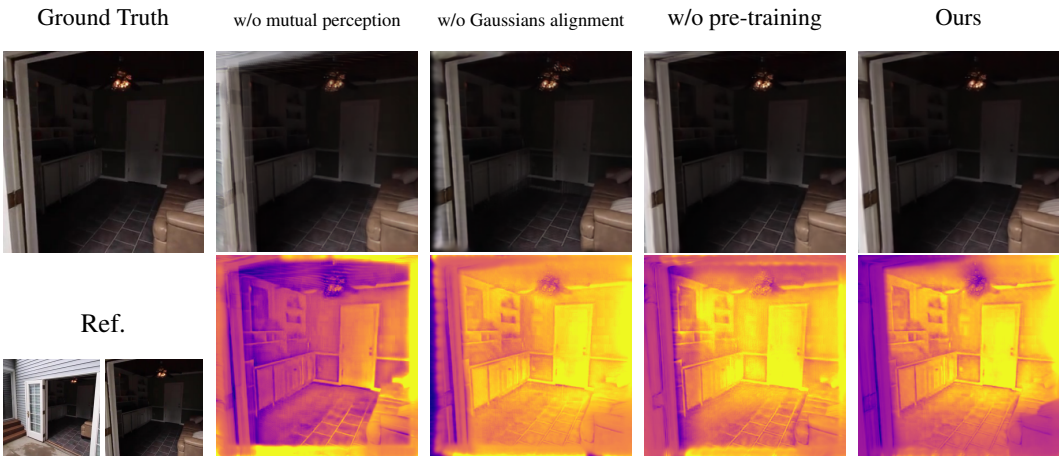

Fig. 6. Ablations. The first row displays the novel viewpoint images, while the last row shows the reference viewpoints and the depth maps of the novel views. Our full model renders higher-quality RGB images and smoother depth maps.

Table 3. Ablations. All ablation experiments were conducted by training and evaluating on the RealEstate10K dataset [72]. Each ablation model was derived from our full model by removing the corresponding modules.

Model	PSNR↑	SSIM↑	LPIPS↓
eFreeSplat (Full)	**26.45**	**0.865**	**0.126**
w/o mutual perception	22.04	0.723	0.212
w/o Gaussians alignment	23.03	0.758	0.187
w/o pre-training weights	24.81	0.829	0.153

each. As shown in Fig. 5, our method performs better in scenes with more minor viewpoint overlaps, while MVSplat excels when the overlap is close to 1. In Tab. 2, our method outperforms other 3DGS baselines [6, 9] in settings with more minor overlaps by 3.1% ↑ in PSNR and 8.6% ↓ in LPIPS. It confirms the robustness of our method in non-overlapping areas. However, methods based on epipolar priors have advantages in scenes where reference viewpoints are closer, and reconstruction quality declines as the overlap decreases.

4.3 Ablation Studies

As shown in Tab. 3 and Fig. 6, we conducted ablation studies on the eFreeSplat model on the RealEstate10K dataset. We will detail the analysis in the following three subsections.

Importance of epipolar-free cross-view mutual perception. Epipolar-free cross-view mutual perception extracts cross-view image features using a shared-weight ViT [11] and a cross-attention decoder. According to Tab. 3, this module's absence results in a 4.41dB decrease in PSNR. In Fig. 6, the absence of cross-view mutual perception results in significant offsets in the depth map and noticeable artifacts.

Importance of iterative cross-view Gaussians alignment. Iterative cross-view Gaussian alignment updates per-pixel Gaussian features and depths through warped U-Net features, thereby aligning the cross-view 3D Gaussian point clouds. The lack of Gaussian alignment can lead to pixel displacement or unreliable local geometric details (*e.g.*, the lamp's position in Fig 6). Additionally, we conducted extra experiments with 1 to 3 iterations. As shown in Fig. 7, using 2 iterations significantly reduces artifacts and inconsistent depth in novel view rendering. This validates that the iterative mode helps align the depth scale across multiple views. When the iteration count increases to 3, there is no notable improvement in reconstruction and rendering quality. For further analysis and results, please refer to Appendix B.

Importance of self-supervised cross-view completion pre-training. In Fig. 6, the absence of cross-view completion pre-training weights results in unaccuracy depth maps. Self-supervised pre-

https://doi.org/10.52202/079017-1251

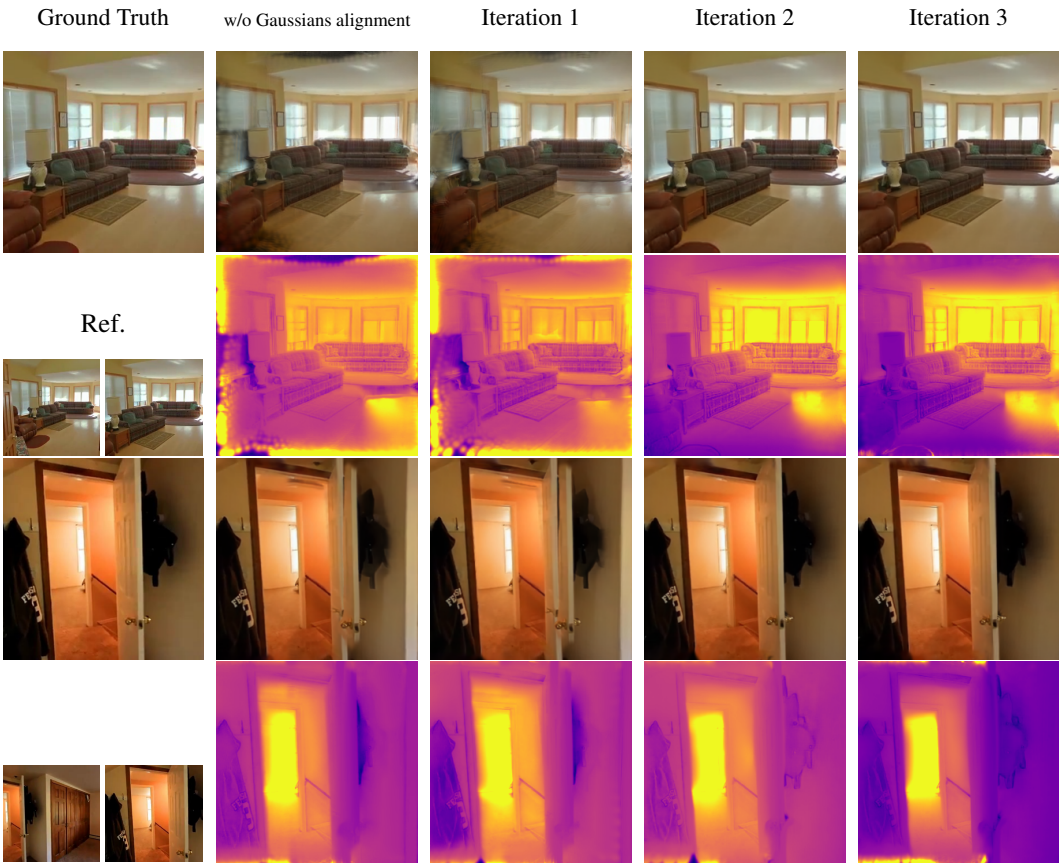

Fig. 7. Ablation of gaussians alignment module. Additional iterations can significantly aid in aligning the depth scale and reducing artifacts that occur during novel view rendering.

training by cross-view completion [60] on large-scale datasets allows our model to perceive spatial correspondences, thereby enabling it to predict more reliable and smoother depth maps.

5 Conclusion

Our work introduces eFreeSplat, a novel generalizable 3D Gaussian Splatting model tailored for novel view synthesis across new scenes, designed to function independently of epipolar constraints that might be unreliable when large viewpoint changes occur. By leveraging a Vision Transformer architecture self-supervised pre-trained by cross-view completion [60] on large-scale datasets, eFreeSplat excels in handling sparse and challenging viewing conditions that traditional methods [17, 63] struggle with. This model's ability to unify the consistency of depth scales across different views marks a significant improvement over existing techniques, effectively addressing issues like artifacts and misalignment in rendered images. Our experiments have demonstrated that our method provides high-quality geometric reconstructions and novel viewpoint images. In settings with a large baseline from 2-view inputs, it outperforms the latest state-of-the-art methods [6, 9] that rely on epipolar priors.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (62293554, 62206249, U2336212), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01073), Ningbo Innovation "Yongjiang 2035" Key Research and Development Programme (2024Z292).

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael G. Rabbat, and Nicolas Ballas. Masked Siamese Networks for Label-efficient Learning. *CoRR*, abs/2204.07141, 2022. 5

[2] 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. *CoRR*, abs/2103.13415, 2021. 3

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-aliased Neural Radiance Fields. *CoRR*, abs/2111.12077, 2021. 3

[4] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian D. Reid. Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. *CoRR*, abs/1908.10553, 2019. 2

[5] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV*, pages 25–36. Springer, 2004. 3

[6] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction. *CoRR*, abs/2312.12337, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 16, 18

[7] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVS-NeRF: Fast Generalizable Radiance Field Reconstruction from Multi-view Stereo. *CoRR*, abs/2103.15595, 2021. 3

[8] Guikun Chen and Wenguan Wang. A Survey on 3D Gaussian Splatting. *CoRR*, abs/2401.03890, 2024. 3

[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. *CoRR*, abs/2403.14627, 2024. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 16, 18

[10] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. GaussianPro: 3D Gaussian Splatting with Progressive Propagation. *CoRR*, abs/2402.14650, 2024. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020. 2, 4, 5, 6, 9, 18

[12] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to Render Novel Views from Wide-baseline Stereo Pairs. *CoRR*, abs/2304.08463, 2023. 2, 4, 5, 6, 7, 15

[13] Ziyue Feng, Leon Yang, Pengsheng Guo, and Bing Li. CVRecon: Rethinking 3D Geometric Feature Learning For Neural Reconstruction. *CoRR*, abs/2304.14633, 2023. 3

[14] Yuan Gan, Ruijie Quan, and Yawei Luo. Expavatar: High-fidelity avatar generation of unseen expressions with 3d face priors. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 3

[15] 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. *CoRR*, abs/1912.06378, 2019. 3

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked Autoencoders Are Scalable Vision Learners. *CoRR*, abs/2111.06377, 2021. 5

[17] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar Transformers. *CoRR*, abs/2005.04551, 2020. 2, 3, 5, 10

[18] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. *CoRR*, abs/2403.17888, 2024. 3

[19] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A Transformer Architecture for Optical Flow. *CoRR*, abs/2203.16194, 2022. 3

[20] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. *arXiv preprint arXiv:2310.01410*, 2023. 3

[21] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. GeoNeRF: Generalizing NeRF with Geometry Priors. *CoRR*, abs/2111.13539, 2021. 3

[22] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end Learning of Geometry and Context for Deep Stereo Regression. *CoRR*, abs/1703.04309, 2017. 3

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-time Radiance Field Rendering. *CoRR*, abs/2308.04079, 2023. 2, 3, 4, 6

[24] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

https://doi.org/10.52202/079017-1251

[25] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. *CoRR*, abs/2203.11483, 2022. 2, 3, 6, 16

[26] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite Nature: Perpetual View Generation of Natural Scenes from a Single Image. *CoRR*, abs/2012.09855, 2020. 2, 6, 16

[27] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *CoRR*, abs/1906.07751, 2019. 3

[28] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019. 3

[29] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020. 3

[30] Yawei Luo and Yi Yang. Large language model and domain-specific model collaboration for smart education. *Frontiers of Information Technology & Electronic Engineering*, 25(3):333–341, 2024. 3

[31] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 3

[32] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[33] Yawei Luo, Ping Liu, and Yi Yang. Kill two birds with one stone: Domain generalization for semantic segmentation via network pruning. *International Journal of Computer Vision*, 2024. 3

[34] Shaojie Ma, Yawei Luo, and Yi Yang. Reconstructing and simulating dynamic 3d objects with mesh-adsorbed gaussian splatting. *arXiv preprint arXiv:2406.01593*, 2024. 3

[35] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. *CoRR*, abs/1512.02134, 2015. 2, 3, 6, 16

[36] Qiaowei Miao, Yawei Luo, and Yi Yang. Pla4d: Pixel-level alignments for text-to-4d gaussian splatting. *arXiv preprint arXiv:2405.19957*, 2024. 3

[37] 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. *CoRR*, abs/2003.08934, 2020. 2, 3

[38] Zhiyuan Min, Yawei Luo, Wei Yang, Yuesong Wang, and Yi Yang. Entangled view-epipolar information aggregation for generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4906–4916, 2024. 2

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *CoRR*, abs/2201.05989, 2022. 3

[40] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open Challenges in Deep Stereo: the Booster Dataset. *CoRR*, abs/2206.04671, 2022. 2, 3, 6, 16

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution Image Synthesis with Latent Diffusion Models. *CoRR*, abs/2112.10752, 2021. 18

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 5

[43] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 3

[44] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution Stereo Datasets with Subpixel-accurate Ground Truth. In *Pattern Recognition - 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings*, pages 31–42. Springer, 2014. 2, 3, 6, 16

[45] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-view Stereo Benchmark with High-resolution Images and Multi-camera Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2538–2547. IEEE Computer Society, 2017. 2, 3, 6, 16

[46] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning Persistent 3D Feature Embeddings. *CoRR*, abs/1812.01024, 2018. 3

[47] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *CoRR*, abs/2104.09864, 2021. 5

[48] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light Field Neural Rendering. *CoRR*, abs/2112.09687, 2021. 2

[49] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable Patch-based Neural Rendering. *CoRR*, abs/2207.10662, 2022. 2, 6, 7

[50] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *arXiv preprint arXiv:2312.13150*, 2023. 3

[51] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is Attention All That NeRF Needs? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[52] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large Multi-view Gaussian Model for High-resolution 3D Content Creation. *CoRR*, abs/2402.05054, 2024. 18

[53] Lijun Wang, Yifan Wang, Linzhao Wang, Yunlong Zhan, Ying Wang, and Huchuan Lu. Can Scale-consistent Monocular Depth Be Learned in a Self-supervised Scale-invariant Manner? In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12707–12716. IEEE, 2021. 2

[54] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3

[55] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. IBRNet: Learning Multi-view Image-based Rendering. *CoRR*, abs/2102.13090, 2021. 2

[56] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D Vision Made Easy. *CoRR*, abs/2312.14132, 2023. 2, 3

[57] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6

[58] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked Feature Prediction for Self-supervised Visual Pre-training. *CoRR*, abs/2112.09133, 2021. 5

[59] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised Pre-training for 3D Vision Tasks by Cross-view Completion. *CoRR*, abs/2210.10716, 2022. 2, 3, 4, 5, 6, 16, 18

[60] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 17923–17934. IEEE, 2023. 2, 3, 4, 5, 6, 10, 16

[61] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentSplat: Autoencoding Variational Gaussians for Fast Generalizable 3D Reconstruction. *CoRR*, abs/2403.16292, 2024. 2, 3, 4, 5

[62] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying Flow, Stereo and Depth Estimation. *CoRR*, abs/2211.05783, 2022. 2, 3

[63] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *CoRR*, abs/1804.02505, 2018. 2, 3, 5, 8, 10

[64] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. *CoRR*, abs/2012.02190, 2020. 2, 3, 8

[65] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. *CoRR*, abs/2112.05131, 2021. 3

[66] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 3

[67] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and Improving Neural Radiance Fields. *CoRR*, abs/2010.07492, 2020. 3

[68] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. 3

[69] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CoRR*, abs/1801.03924, 2018. 6

[70] Zhengyou Zhang. Determining the Epipolar Geometry and its Uncertainty: A Review. *Int. J. Comput. Vis.*, 27(2):161–195, 1998. 2

[71] 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. *CoRR*, abs/2312.02155, 2023. 3

 https://doi.org/10.52202/079017-1251

[72] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis using Multiplane Images. *CoRR*, abs/1805.09817, 2018. 2, 6, 8, 9, 15, 16

[73] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane Meets Gaussian Splatting: Fast and Generalizable Single-view 3D Reconstruction with Transformers. *CoRR*, abs/2312.09147, 2023. 3

A Additional Experimental results

We provide additional qualitative comparisons against baselines. The visualization results on the RealEstate10K are shown in Fig. 8. Additionally, we provide more geometry reconstruction comparison results, as shown in Fig. 9. Our method reconstructs high-quality 3DGS without using epipolar priors or depth regularization finetuning.

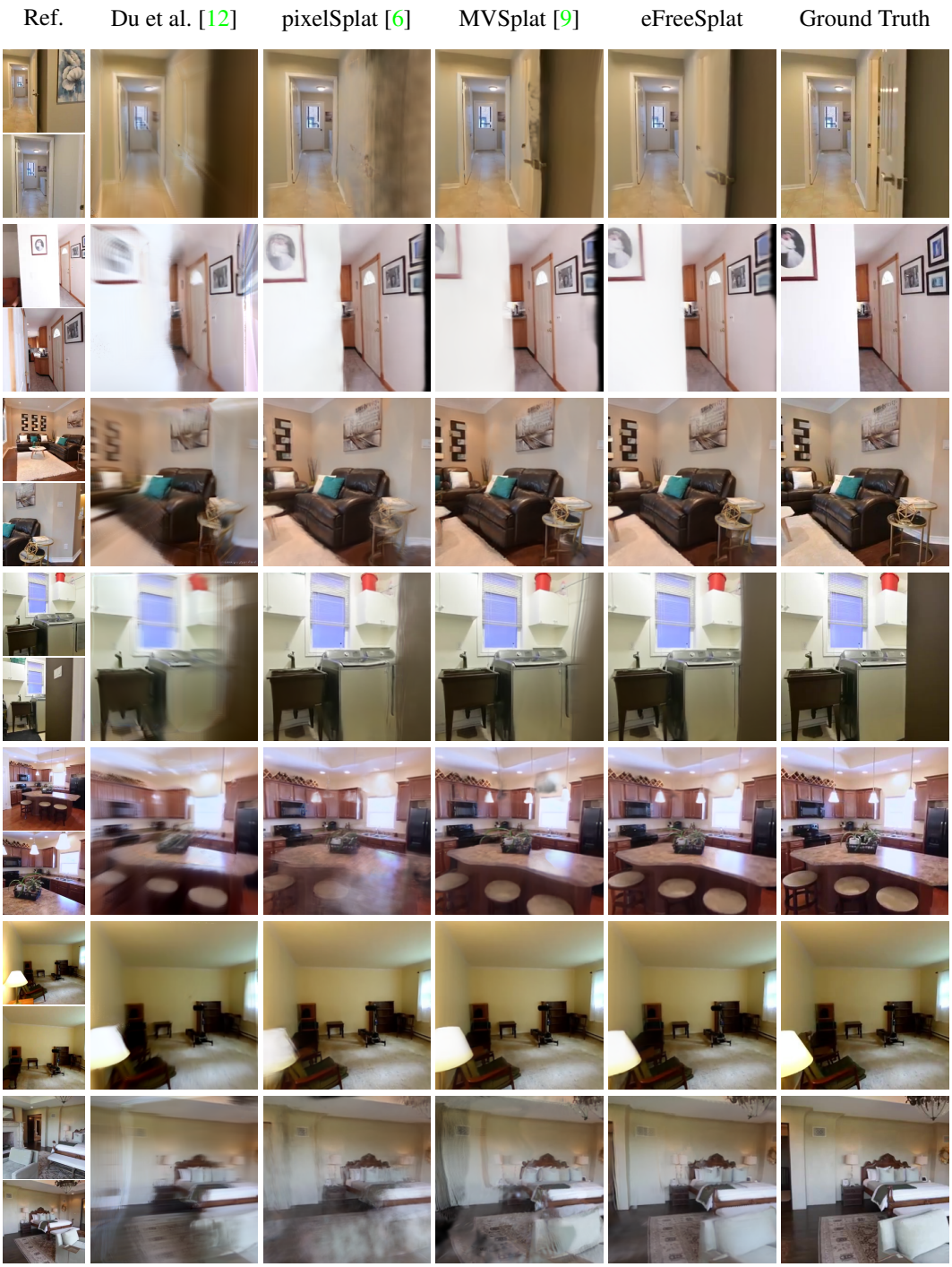

Fig. 8. Additional visualization results on the RealEstate10K [72]. Our method, eFreeSplat, outperforms baselines in rendering results, producing fewer artifacts and scene distortions.

https://doi.org/10.52202/079017-1251

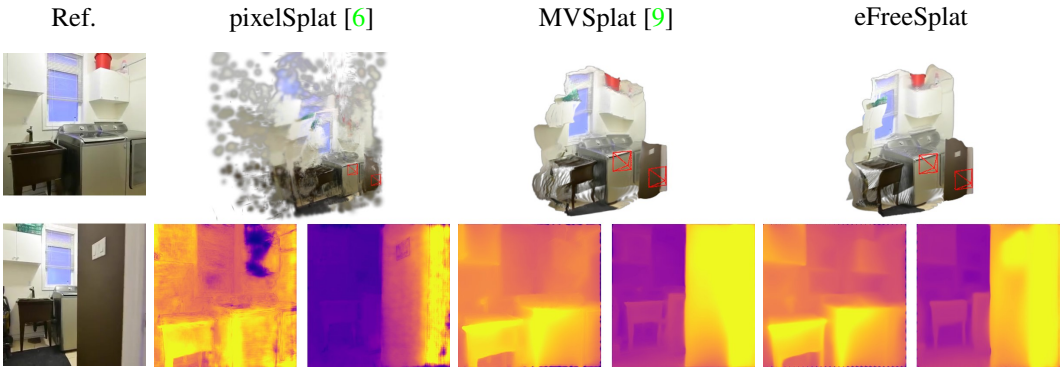

Fig. 9. Additional geometry reconstruction quality comparison results. Our method achieves higher quality in 3D Gaussian Splatting and produces smoother depth maps than pixelSplat [6] and MVSplat [9].

Table 4. Quantitative results of gaussians alignment module under the settings of 1 to 3. "Memory" refers to GPU memory usage, and "Time" indicates the inference time.

Iterations	PSNR↑	SSIM↑	LPIPS↓	Memory(M)	Times(s)
1	23.36	0.768	0.182	2410	0.058
2	**26.45**	**0.865**	0.126	2452	0.061
3	26.40	0.861	**0.126**	2488	0.086

Fig. 10. Visualization of gaussians alignment module under the settings of 1 to 3.

B Additional Experimental Analysis

More ablations. In this section, we provide both quantitative and qualitative results of the Gaussians Alignment module under the settings of 1 to 3 iterations. As shown in Tab. 4 and Fig. 10, setting the iteration count to 2 effectively reduces artifacts caused by inconsistent depth scales. When the iteration count is set to 3, we had to reduce the model's parameter size and batch size to avoid OOM errors, which might be one of the reasons for the lack of significant improvement in image reconstruction metrics.

Failure cases. Our method relies on the 3D prior knowledge provided by CroCo [59] pre-trained weights. However, the input viewpoint overlap in the pre-trained dataset does not exceed 0.75 [60], while the input viewpoint overlap in the RealEstate10K and ACID datasets mainly ranges from 0.9 to 1.0. As shown in Fig. 11, our method renders unreliable results when the input viewpoints are very close, which can be attributed to the distribution bias between the GNVS dataset [26, 72] and the pre-trained dataset [25, 35, 40, 44, 45].

https://doi.org/10.52202/079017-1251

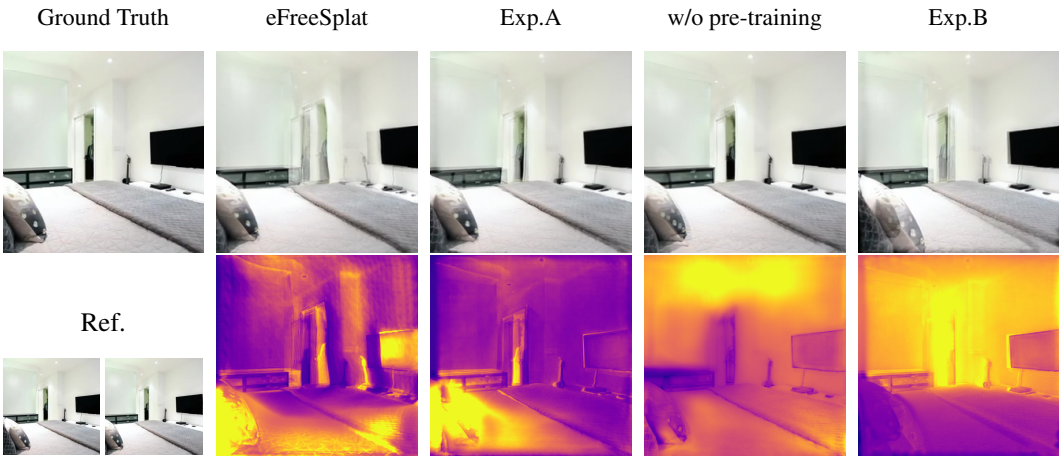

Fig. 11. Failure cases. Our method may produce unstable results in scenarios where the input viewpoints are very close. Exp.A and Exp.B indicate that fine-tuning the CroCo model on Re10k helps mitigate this issue.

Table 5. Exp. A involves fine-tuning the CroCo pretrained weights using the RE10K training set, while Exp. B trains CroCo directly using the RE10K training set without loading the pretrained weights. w/o pre-training refers to neither using CroCo pre-trained weights nor performing fine-tuning.

Model	PSNR↑	SSIM↑	LPIPS↓
raw eFreeSplat	**26.45**	**0.865**	**0.126**
Exp.A	26.32	0.862	0.129
w/o pre-training	24.81	0.829	0.153
Exp.B	**25.12**	**0.839**	**0.144**

Limitations. Our method lacks geometric inductive biases, so our model is data-hungry and sensitive to the training data distribution. Joint training with richer multiview datasets across different scenes could be a viable direction. Additionally, the per-pixel 3D Gaussian mapping struggles to reconstruct parts of the scene that are occluded or missing from input viewpoints, such as an obscured chair. Therefore, introducing high-level features for scene completion might be a future research direction for generalizable 3D Gaussian Splatting work.

Fine-tuning of the CroCo model. We have conducted preliminary explorations to address the aforementioned limitation. We conducted relevant Experiments A and B regarding fine-tuning the CroCo model using the RE10K dataset. Experiment A involved fine-tuning the CroCo pretrained weights with the RE10K training set, while Experiment B involved training CroCo directly with the RE10K training set without loading the pretrained weights. Finally, we retrained eFreeSplat using the new pretrained weights. As shown in Fig. 11 and Tab. 5, the results indicate that pretraining the backbone model on the RE10K training set effectively addresses the model's poor performance in low-overlap scenarios. However, in the RE10K test set, Experiment A's reconstruction metrics were slightly lower than those of the original model, which may be due to insufficient training iterations. We will further investigate the positive impact of fine-tuning the CroCo pretrained model on novel view synthesis and 3D reconstruction in future work.

Potential negative societal impacts. Our model could be misused for unethical purposes, such as creating false evidence or manipulating media, which threatens information integrity and personal privacy. Additionally, the model introduces security risks in contexts like autonomous driving, as it may produce incorrect reconstructions in real and complex scenarios. These concerns underscore the importance of implementing stringent ethical guidelines and security measures when deploying such technology, to prevent misuse and ensure that it is used responsibly.

https://doi.org/10.52202/079017-1251

C Additional Implementation Details

The cross-attention decoder. Following the *CrossBlock* decoder architecture in CroCo [59], the cross-attention decoder comprises a self-attention module and a cross-attention module. Let $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{R}^{N \times C}$ be the tokens of the two viewpoints outputted by a Vision Transformer [11]. The computation process of the decoder is as follows:

$$\begin{aligned}
\bar{\boldsymbol{\varepsilon}}_i &= \text{ LayerNorm } (\boldsymbol{\varepsilon}_i), && i = 1, 2 \\
\boldsymbol{\varepsilon}'_i &= \boldsymbol{\varepsilon}_i + \text{Attention } (\bar{\boldsymbol{\varepsilon}}_i, \bar{\boldsymbol{\varepsilon}}_i, \bar{\boldsymbol{\varepsilon}}_i), && i = 1, 2 \\
\boldsymbol{\varepsilon}''_1 &= \boldsymbol{\varepsilon}'_1 + \text{Attention } (\text{LayerNorm } (\boldsymbol{\varepsilon}'_1), \bar{\boldsymbol{\varepsilon}}_2, \bar{\boldsymbol{\varepsilon}}_2), && \\
\boldsymbol{\varepsilon}''_2 &= \boldsymbol{\varepsilon}'_2 + \text{Attention } (\text{LayerNorm } (\boldsymbol{\varepsilon}'_2), \bar{\boldsymbol{\varepsilon}}_1, \bar{\boldsymbol{\varepsilon}}_1), && \\
\text{output}_i &= \boldsymbol{\varepsilon}''_i + \text{MLP } (\text{LayerNorm } (\boldsymbol{\varepsilon}''_i)), && i = 1, 2
\end{aligned} \tag{10}$$

In Equations 10, Attention is derived from the classic attention computation. The inputs Q, K, V undergo projection transformations using W_q, W_k, W_v:

$$\begin{aligned}
Q' = W_q Q, \ K' = W_k K, \ V' = W_v V, \\
\text{Attention}(Q, K, V) = \text{Linear} \left(\text{softmax} \left(\frac{Q' {K'}^\top}{\sqrt{C}} \right) V' \right).
\end{aligned} \tag{11}$$

The cross-view U-Net. For the Gaussian Alignment Strategy and the prediction of Gaussian primitives, we utilize a 2D Cross-View U-Net inspired by [41, 52], 2024. We concatenate and flatten multiview feature maps for cross-view information exchange, similar to the structure of the U-Net used for cost volume refinement in MVSplat [9]. Specifically, for the Gaussian Alignment Strategy, we apply four times of 2 $\times$ down-sampling and add attention at the 16 $\times$ down-sampled level, with the channel dimensions being [32, 32, 64, 128, 256]. For the prediction of Gaussian primitives, we keep the channel dimension fixed at 32, while the rest of the architecture remains the same as that of the U-Net used in the Gaussian Alignment Strategy.

More training details. Our model is trained and tested on 4 RTX-4090 GPUs using the Adam optimizer with a learning rate 2e-4. The per-GPU batch size during training is 4. Similar to pixelSplat [6], the distance between the two input viewpoints gradually increases throughout training. However, to learn more robust 3D prior information, our setup allows for a maximum viewpoint distance of 60 frames, compared to the 45 frames used by pixelSplat [6] and MVSplat [9].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions and scope are claimed clearly in the abstract and introduction (last paragraph).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the Limitation section in Appendix B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-1251

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We introduce our method and experimental details in the paper. We will open source the code and provide model weights in the future.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our datasets is open-source, and we will open source our code in the future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail all the training and test details in the Experiments and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include statistical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/079017-1251

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Appenidx B.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We follow the license CC-BY 4.0.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

https://doi.org/10.52202/079017-1251

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.