

WildSplat: Feedforward Gaussian Splatting from Unposed In-the-Wild Images

Xiyu Zhang^{1,2,*,†}, Jingyu Zhuang^{2,*}, Hongjia Zhai¹, Zizheng Yan²,
Jinwei Chen², Guofeng Zhang^{1,†}, and Qingnan Fan^{2,†}

¹ State Key Lab of CAD&CG, Zhejiang University

² vivo BlueImage Lab

*Equal contribution. †Corresponding author.

‡Work done during an internship at vivo BlueImage Lab.

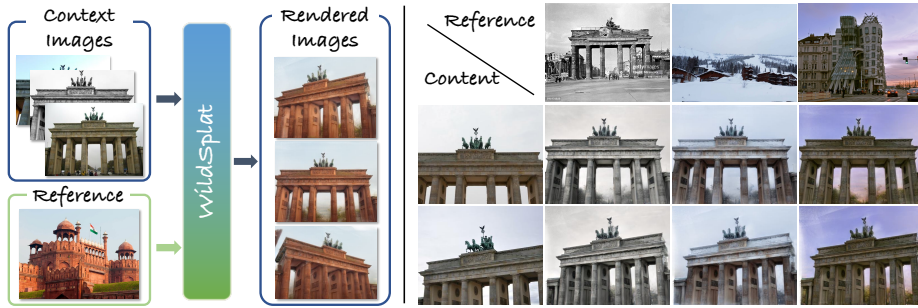


Fig. 1: Given a set of unposed in-the-wild images with varying appearances and a reference image, **WildSplat** can reconstruct a high-quality 3D scene in a single feedforward pass. The reconstructed scene can then be rendered with the target appearance specified by the reference image.

Abstract. While feedforward 3D reconstruction excels at efficient novel view synthesis, it typically falters when faced with scenes under varying illumination. To this end, we introduce **WildSplat**, the first feedforward 3D Gaussian Splatting framework capable of appearance-conditioned novel-view synthesis for unposed in-the-wild images. To handle inconsistent photometric conditions, we propose a dual-branch architecture that explicitly decouples geometry from appearance. The geometry branch extracts an appearance-invariant 3D structure and jointly predicts camera poses. To govern the rendering appearance, the appearance branch injects target appearance cues into the content features via a globally pre-modulated cross-attention mechanism. To further prevent feature entanglement, we introduce a joint multi-reference training strategy that stabilizes the training process. Extensive experiments show that WildSplat surpasses existing optimization-based and feedforward methods, achieving state-of-the-art performance in in-the-wild novel view synthesis and appearance editing from sparse inputs in a single forward pass.

Keywords: Novel View Synthesis · 3D Gaussian Splatting · Feedforward Reconstruction · In-the-Wild

1 Introduction

Novel view synthesis from multiple images is a longstanding problem in computer vision and graphics, forming the foundation for many downstream applications, such as virtual reality, augmented reality, and robotics. Recently, Neural Radiance Fields (NeRF) [21] and 3D Gaussian Splatting (3DGS) [12] have achieved photorealistic rendering by representing scenes with neural networks or 3D Gaussians. However, these methods assume static scene geometry, materials, and lighting—an assumption usually violated for in-the-wild photo collections. Such images, typically sourced from the internet, capture the same scene under varying conditions, exhibiting significant appearance differences in lighting, weather, and camera exposure.

To handle such complex scenarios, the community has historically relied on optimization-based approaches. Pioneering methods like NeRF-W [20] and subsequent appearance-aware 3DGS techniques [8, 13, 36, 41] address this by optimizing per-image appearance embeddings to decouple the shared scene structure from these photometric variations. While these methods achieve impressive appearance interpolation, they strictly require accurately precomputed camera poses, and their performance heavily depends on dense image collections and lengthy per-scene optimization. When only sparse views are available, optimizing per-image embeddings becomes severely under-constrained, leading to inaccurate geometry and a complete lack of reconstruction efficiency.

Meanwhile, some methods leverage large-scale multi-view datasets to learn generalizable priors and predict 3D representations in a single forward pass [11, 32, 33, 37]. Without requiring exhaustive per-scene optimization, these feedforward methods enable efficient and photorealistic novel-view synthesis. However, extending such models to in-the-wild photo collections remains highly challenging. When the appearance of the input images varies significantly, existing feedforward frameworks struggle to disentangle underlying 3D geometry from illumination changes, resulting in noticeable appearance inconsistencies in the rendered results.

To this end, we introduce **WildSplat**, the first feedforward 3D Gaussian Splatting framework designed for in-the-wild scenes with varying appearances. By conditioning the reconstruction on a designated reference image, WildSplat enables photorealistic novel-view synthesis without exhaustive per-scene optimization. Specifically, we propose a dual-branch architecture to explicitly decouple geometry from appearance. This design allows us to reconstruct 3D Gaussians by separating appearance-invariant geometry from target-conditioned appearance. The geometry branch is responsible for predicting this invariant geometry, alongside robust content features that encode the structural information of the scene. To control the rendering appearance, an appearance encoder extracts features from the reference image and injects the appearance information into the content features via a cross-attention-based appearance injector. Furthermore, to stabilize the training process, we introduce a multi-reference training paradigm. Within a single training batch, our framework constructs an appearance-invariant geometry but simultaneously renders it into multiple views under distinct appearance conditions. This strategy enables the model to effectively accommodate photometric

variations, ensuring accurate appearance modeling and photorealistic rendering for diverse target views.

In summary, our main contributions are as follows:

1. We propose **WildSplat**, the first feedforward 3D Gaussian Splatting framework enabling appearance-conditioned novel-view synthesis from sparse, unposed, in-the-wild image collections.
2. We design a novel architecture that explicitly injects target appearance features into the content features from the geometry backbone via global pre-modulated cross-attention, enabling appearance-conditioned synthesis in a single forward pass.
3. We introduce a multi-reference training paradigm tailored for photometrically inconsistent data, encouraging the model to learn robust 3D priors and accurately model real-world appearances.
4. We conduct extensive experiments to demonstrate that WildSplat achieves state-of-the-art performance on in-the-wild datasets, significantly outperforming both existing feedforward approaches and optimization-based methods.

2 Related Work

2.1 Novel View Synthesis

Novel view synthesis aims to infer scene structure and appearance based on a given set of input views. Traditional works focus on explicit [9, 17, 31] or discrete volumetric [14, 27, 29] representations of the underlying geometry, while recent methods employ radiance field-based representations [21] or 3D Gaussian Splatting [12] to achieve unprecedented photorealistic rendering quality. Neural Radiance Fields (NeRF) [21] and related methods [1–3] have pioneered high-quality novel view synthesis by training neural networks to learn neural fields that implicitly represent scene geometry and appearance. However, their reliance on computationally expensive volume rendering limits real-time performance. In contrast, 3D Gaussian Splatting (3DGS) [12] explicitly represents scenes with millions of anisotropic Gaussians and achieves real-time photorealistic novel view rendering through differentiable Gaussian projection. Its explicit and efficient representation has further benefited downstream tasks such as visual localization, surface reconstruction, and open-vocabulary panoptic mapping [39, 40, 42].

The aforementioned novel view synthesis methods assume that scene geometry, materials, and lighting are static; however, in-the-wild images collected from the Internet often violate this assumption. To address this issue, NeRF-W [20] learns per-image appearance and transient embeddings to model photometric variations and disentangle transient elements. GS-W [41] employs 3DGS to represent the scene and introduces separate intrinsic and dynamic features for each Gaussian to capture the static scene along with dynamic variations. WildGaussians [13] leverages robust DINO [23] features to eliminate occluders and integrates an appearance modeling module within 3DGS to predict color affine transformations. Similarly, Splatfacto-W [34] integrates per-image appearance embeddings and

per-Gaussian color features into the rasterization process. Furthermore, WildGS [36] decomposes appearance into image-based global and local embeddings and leverages a triplane representation to achieve explicit local appearance modeling. Despite these advances, most in-the-wild NeRF and 3DGS methods typically rely on Structure-from-Motion tools [26] to obtain camera poses and still involve redundant optimization processes, which significantly reduce reconstruction efficiency.

2.2 Feedforward Reconstruction

Since gradient backpropagation in per-scene optimization is computationally expensive and time-consuming, some methods attempt to learn powerful priors from large-scale datasets to reconstruct scenes and enable novel view synthesis via a single feedforward pass. pixelNeRF [38] introduces a fully convolutional architecture that conditions on input images to predict pixel-aligned features for radiance field reconstruction. Subsequent feedforward NeRF models further improve performance by incorporating feature matching [5, 6], Transformers [10, 22, 25], and 3D volumetric representations [35]. However, due to the inherently pixel-wise volume rendering, existing feedforward NeRF models still suffer from slow rendering speeds.

Recently, a growing number of feedforward models based on 3DGS have been proposed to avoid the expensive volume rendering in NeRF. pixelSplat [4] learns a Transformer-based model to regress Gaussian parameters from two input views by extracting cross-view-aware features and predicting probabilistic depth distributions for depth sampling. Different from existing data-driven approaches, MVSPat [7] constructs a cost volume to exploit multi-view correspondence information for improved geometry learning. Without requiring accurate camera poses as input, NoPoSplat [37] anchors one input camera coordinate system as the canonical space, predicts Gaussians for all views within this space, and encodes camera intrinsics as token embeddings to enable accurate scene scale prediction. AnySplat [11] distills camera and geometry priors from a pretrained VGGT [32] model to encode input images into high-dimensional features, which are then decoded into Gaussian parameters and camera poses. Despite the impressive progress of feedforward reconstruction models, they rely on a strict assumption of photometric consistency across input views and struggle to disentangle the underlying 3D geometry from illumination variations. When in-the-wild images are used as inputs, appearance inconsistencies caused by varying weather and lighting conditions can severely degrade the visual consistency of the reconstructed scenes produced by existing feedforward methods.

3 Methods

In Fig. 2, we present **WildSplat**, a novel feedforward 3D Gaussian Splatting framework designed for unposed scenes with varying appearances. Given a set of context images exhibiting various appearances and a designated reference

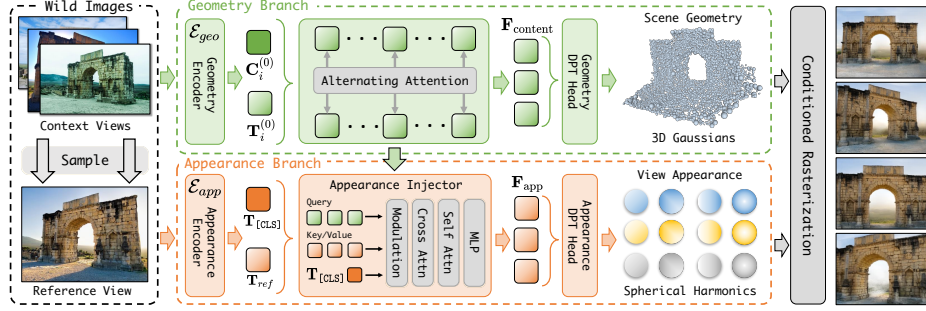


Fig. 2: Overview of WildSplat. Given context images with varying appearances and a reference view, the Geometry Branch extracts content features and predicts scene geometry along with explicit 3D Gaussian attributes. Concurrently, the Appearance Branch employs an Appearance Injector, conditioned on a reference appearance image, to predict view-consistent color attributes. The decoupled geometry and conditioned appearance are then combined for conditioned rasterization to synthesize novel views with the designated reference appearance.

image, WildSplat decouples the scene into appearance-invariant geometry and conditioned appearance under the target appearance. The problem formulation is introduced in Sec. 3.1, followed by a detailed description of our WildSplat in Sec. 3.2. Finally, the multi-reference training strategy is presented in Sec. 3.3.

3.1 Problem formulation

Unlike standard novel view synthesis settings that rely on strict multi-view photometric consistency, we focus on unposed in-the-wild scenarios where the context images $\{I_i\}$ frequently exhibit photometric variations. To resolve the resulting photometric ambiguities, our framework explicitly decouples the geometry and appearance and conditions the appearance on a target reference image I_{ref} , which provides the designated appearance cues. Consequently, WildSplat jointly predicts the camera poses $\{p_i\}$ of the input images, the underlying 3D Gaussian geometry attributes $\mathcal{G}$, and the conditioned appearance attributes $\mathcal{C}_{ref}$. This formulation ensures that the reconstructed 3D representation geometrically aligns with the context views while strictly matching the appearance of I_{ref} . Formally, WildSplat is defined as:

$$f : \{I_i\}, I_{ref} \mapsto (\mathcal{G}, \mathcal{C}_{ref}, \{p_i\}), \quad (1)$$

where $\mathcal{G}$ consists of Gaussian positions $\boldsymbol{\mu} \in \mathbb{R}^{G \times 3}$, rotations $\mathbf{q} \in \mathbb{R}^{G \times 4}$, scales $\mathbf{s} \in \mathbb{R}^{G \times 3}$, and opacities $\boldsymbol{\alpha} \in \mathbb{R}^G$. $\mathcal{C}_{ref} \in \mathbb{R}^{G \times 3 \times (k+1)^2}$ contains spherical harmonics coefficients of degree k conditioned on the reference image I_{ref} , and G is the number of Gaussians.

3.2 Decoupled Geometry and Appearance Modeling

Previous feedforward methods typically predict 3D Gaussian geometry and color attributes jointly from an entangled representation. To prevent the target appearance conditioning from inadvertently altering the underlying geometric structure, we explicitly decouple these properties and design a dual-branch framework to separate the prediction of these attributes. Specifically, the geometry branch extracts appearance-invariant geometry that remains consistent across arbitrary reference images, while the appearance branch, via an appearance injector module, predicts decoupled appearance from the conditioned reference image.

Geometry Branch. We build our geometry branch upon the VGGT-like architecture [32]. Initially, N context images are processed independently by a DINOv2 [23] based geometry encoder $\mathcal{E}_{\text{geo}}$, yielding dense patch-level tokens $\mathbf{T}_i^{(0)} = \mathcal{E}_{\text{geo}}(I_i) \in \mathbb{R}^{P \times D}$ for each view $i \in \{1, \dots, N\}$. To enable camera pose estimation, a set of learnable camera tokens $\mathbf{C}_i^{(0)} \in \mathbb{R}^{5 \times D}$ is concatenated with the patch tokens. These combined tokens are refined through L alternating attention blocks. Let $\mathbf{T}_i^{(k)}$ and $\mathbf{C}_i^{(k)}$ denote the patch and camera tokens for the i -th view after the k -th layer, respectively.

To encapsulate the appearance-invariant structural information of the scene, we extract and aggregate multi-scale representations from a subset of transformer layers $\mathcal{K} = \{l_1, \dots, l_K\}$. We formally define these aggregated representations as the content features, which serve as a foundational geometric scaffold prior to any appearance conditioning:

$$\mathbf{F}_{\text{content}} = \left\{ \mathbf{T}_i^{(k)} \mid k \in \mathcal{K}, i = 1 \dots N \right\}. \quad (2)$$

Based on these enriched representations, we employ specialized heads to regress 3D Gaussian properties. Specifically, DPT-based heads $\mathcal{H}$ map the aggregated features $\mathbf{F}_{\text{content}}$ to pixel-aligned depth maps $\{d_i\}$ and Gaussian attributes $\{s_i, q_i, \alpha_i\}$. Simultaneously, the refined camera tokens $\mathbf{C}_i^{(L)}$ are processed by a camera head to predict poses $\{p_i\}$. The explicit 3D Gaussian centers μ_i are then derived via a back-projection layer:

$$\begin{aligned} \{d_i\} &= \mathcal{H}_{\text{depth}}(\mathbf{F}_{\text{content}}), \quad p_i = \mathcal{H}_{\text{cam}}(\mathbf{C}_i^{(L)}), \\ \{s_i, q_i, \alpha_i\} &= \mathcal{H}_{\text{attr}}(\mathbf{F}_{\text{content}}), \quad \mu_i = \text{backproj}(d_i, p_i). \end{aligned} \quad (3)$$

Finally, based on the predicted camera pose, we transform the 3D Gaussians from the viewpoint into the reconstructed global coordinates, thereby obtaining a viewpoint-independent representation of the scene geometry: $\mathcal{G} = \{\boldsymbol{\mu}, \mathbf{s}, \mathbf{q}, \boldsymbol{\alpha}\}$.

Appearance Branch. The appearance branch aims to predict appearance conditioned on the target reference image. Given the target reference image I_{ref} , we first employ another DINOv2 based appearance encoder $\mathcal{E}_{\text{app}}$ to extract the appearance information:

$$\mathbf{T}_{[\text{CLS}]}, \mathbf{T}_{\text{ref}} = \mathcal{E}_{\text{app}}(I_{\text{ref}}), \quad (4)$$

where $\mathbf{T}_{[\text{CLS}]} \in \mathbb{R}^D$ corresponds to the [CLS] token that represents global appearance, and $\mathbf{T}_{ref} \in \mathbb{R}^{P_a \times D}$ represents local dense appearance tokens, with P_a being the number of patches and D the feature dimension. Then, we use a cross-attention mechanism to integrate the target appearance into the content features $\mathbf{F}_{\text{content}}$. The content features $\mathbf{F}_{\text{content}}$ extracted from the geometry branch act as the queries (Q), while the reference appearance tokens $\mathbf{T}_{ref}$ serve as the keys (K) and values (V).

However, the content features $\mathbf{F}_{\text{content}}$ from the geometry branch inherently retain the original photometric properties of the context views. To explicitly mitigate the influence of the original appearances, we introduce a global pre-modulation step before the cross-attention operation. We leverage the global descriptor $\mathbf{T}_{[\text{CLS}]}$ to holistically shift the content features into the target appearance domain. Specifically, an AdaLN-Zero layer [24] is employed to decode $\mathbf{T}_{[\text{CLS}]}$ into three parameters: a scale γ , a shift β , and an adaptive gating η . These parameters are then used to modulate the appearance information within the content feature $\mathbf{F}_{\text{content}}$ as follows:

$$\tilde{\mathbf{T}}_i^{(k)} = \gamma \odot \text{LayerNorm}(\mathbf{T}_i^{(k)}) + \beta, \quad [\gamma, \beta, \eta] = \text{MLP}_{\text{cond}}(\mathbf{T}_{[\text{CLS}]}) \quad (5)$$

where $\odot$ denotes element-wise multiplication. The appearance-modulated feature $\tilde{\mathbf{T}}^{(k)}$ is then utilized as the query in the cross-attention layer to aggregate fine-grained, condition-specific details from T_{ref} . This is followed by a self-attention layer to ensure spatial regularization. The complete decoupled injection process is formulated as:

$$\bar{\mathbf{T}}_i^{(k)} = \mathbf{T}_i^{(k)} + \eta \odot \text{CrossAttn}(\tilde{\mathbf{T}}_i^{(k)}, \mathbf{T}_{ref}), \quad \hat{\mathbf{T}}_i^{(k)} = \bar{\mathbf{T}}_i^{(k)} + \text{SelfAttn}(\bar{\mathbf{T}}_i^{(k)}), \quad (6)$$

$$\mathbf{T}_{\text{app},i}^{(k)} = \hat{\mathbf{T}}_i^{(k)} + \text{MLP}(\hat{\mathbf{T}}_i^{(k)}). \quad (7)$$

Finally, the updated appearance features $\mathbf{F}_{\text{app}} = \{\mathbf{T}_{\text{app},i}^{(k)} \mid k \in \mathcal{K}, i = 1, \dots, N\}$ are routed to the appearance head $\mathcal{H}_{\text{color}}$ to predict the conditioned 3D Gaussian color attributes $\mathcal{C}_{ref}$:

$$\mathcal{C}_{ref} = \mathcal{H}_{\text{color}}(\mathbf{F}_{\text{app}}). \quad (8)$$

Combined with the appearance-invariant geometry attributes $\mathcal{G} = \{\boldsymbol{\mu}, \mathbf{s}, \mathbf{q}, \boldsymbol{\alpha}\}$, the fully decoupled 3D representation $\{\mathcal{G}, \mathcal{C}_{ref}\}$ is ready for conditioned rasterization. **Conditioned Rasterization.** With the decoupled geometry and appearance attributes, we perform a conditioned rasterization to ensure the rendered novel views maintain a globally consistent 3D structure while seamlessly integrating the conditioned appearance. Formally, given the decoupled 3D representation $\{\mathcal{G}, \mathcal{C}_{ref}\}$, and camera poses $\{p_i\}$ of target views, the conditioned representation is rendered to synthesize the predicted outputs $\{\hat{I}_i\}$:

$$\{\hat{I}_i\} = \mathcal{R}(\mathcal{G}, \mathcal{C}_{ref}, \{p_i\}), \quad (9)$$

where $\mathcal{R}$ denotes the tile-based differentiable rasterization process of 3D Gaussian Splatting.

Algorithm 1 Geometry-guided View Sampling**Input:** Matrices S (scale) & R (ratio), num views N , sample count K , thresholds τ_s, τ_r, τ_n **Output:** Selected indices I

```

1:  $V \leftarrow (S \geq \tau_s) \wedge (R \geq \tau_r)$  // Valid mask
2:  $C \leftarrow \{i \mid \sum_j V_{i,j} > \tau_n\}$  // Valid starting rows
3: Initialize queue  $Q \leftarrow [c]$  and set  $I \leftarrow \{c\}$  where  $c \sim \text{Uniform}(C)$ 
4: while  $|I| < N$  and  $Q \neq \emptyset$  do
5:    $u \leftarrow Q.\text{pop}()$ 
6:    $S_k \leftarrow \text{Sample up to } K \text{ elements from } \{v \mid V_{u,v} \wedge v \notin I\}$ 
7:    $Q.\text{push}(S_k)$  and  $I \leftarrow I \cup S_k$ 
8: end while
9: return  $I$ 

```

3.3 Training Paradigm

Geometry-guided View Sampling. In feedforward training, input views must share sufficient visual overlap to establish reliable geometric correspondences. Standard sampling strategies, however, rely solely on camera extrinsics or sequential information. These approaches may fail on in-the-wild datasets, where varying focal lengths mean that even minor pose shifts can drastically alter the captured content. To guarantee a valid multi-view context, we introduce an overlap-based sampling strategy. Leveraging SfM points, we formulate an overlap matrix and a scale matrix to quantify the ratio of shared 3D points and the minimum depth ratio between views, respectively. Based on predefined thresholds, we first discard view pairs lacking adequate spatial overlap or scale consistency. Then, initializing from a strongly connected candidate view, we iteratively aggregate valid neighboring views. The complete procedure is detailed in Algorithm 1.

Joint Multi-Reference Supervision. To stabilize the training process and explicitly prevent the entanglement of geometry and appearance, we introduce a joint multi-reference training strategy. During a training iteration, after extracting the appearance-invariant geometry $\mathcal{G}$, the intermediate content features $\mathbf{F}_{\text{content}}$, and the camera poses $\{p_i\}$ from a batch of sampled multi-view images, we explicitly supervise the appearance conditioning. To capture the diverse appearances within the batch, we randomly sample a subset of M images, $\{I_m\}$, as target appearance references and ground truth. The appearance branch processes these M references to extract conditioned appearance attributes $\{\mathcal{C}_{ref,m}\}$. Finally, we render images from the viewpoint of reference images explicitly conditioned on their respective appearance information. The rendered M images $\{\hat{I}_m\}$ are subsequently used for end-to-end supervision. To prevent appearance leakage and ensure the geometry remains illumination-agnostic, we apply color jittering to the input images and random cropping to the references prior to feature extraction.

Table 1: Quantitative comparison of novel view synthesis on MegaScenes [30]. We evaluate models under sparse input configurations (4, 8, and 12 views) using a fixed test set. Performance is measured using PSNR ($\uparrow$), SSIM ($\uparrow$), and LPIPS ($\downarrow$).

Method	4 views			8 views			12 views		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
FSGS [43]	12.88	0.352	0.463	13.28	0.371	0.454	13.43	0.376	0.453
GS-W [41]	9.54	0.265	0.539	10.76	0.294	0.500	9.81	0.311	0.526
WildGaussian [13]	14.38	0.385	0.458	15.32	0.436	0.433	15.68	0.448	0.452
AnySplat [11]	14.58	0.559	0.313	16.16	0.594	0.274	16.31	0.601	0.273
WorldMirror [19]	15.33	0.609	0.288	16.08	0.633	0.281	16.05	0.629	0.287
Ours	17.91	0.635	0.235	18.65	0.668	0.218	19.20	0.683	0.210

Table 2: Quantitative comparison of novel view synthesis on Phototourism [28]. We evaluate models under sparse input configurations (4, 8, and 12 views) using a fixed test set. Performance is measured using PSNR ($\uparrow$), SSIM ($\uparrow$), and LPIPS ($\downarrow$).

Method	4 views			8 views			12 views		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
FSGS [43]	13.63	0.476	0.413	14.85	0.532	0.366	15.39	0.573	0.346
GS-W [41]	9.40	0.344	0.512	10.19	0.354	0.490	12.63	0.399	0.437
WildGaussian [13]	14.95	0.485	0.436	17.02	0.567	0.371	17.65	0.616	0.353
AnySplat [11]	15.61	0.616	0.279	17.35	0.681	0.240	18.00	0.700	0.228
WorldMirror [19]	15.38	0.664	0.273	16.71	0.699	0.252	17.24	0.724	0.234
Ours	19.57	0.727	0.178	21.15	0.771	0.146	21.47	0.788	0.137

Training Objectives. The joint objective function is defined as the summation of the reconstruction losses over these rendered images:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \lambda_1 \mathcal{L}_{\text{pose}}. \quad (10)$$

$\mathcal{L}_{\text{rgb}}$ is the photometric loss, which is formulated as a weighted combination of multiple loss terms:

$$\mathcal{L}_{\text{rgb}} = \lambda_2 \mathcal{L}_{\text{MSE}} + \lambda_3 \mathcal{L}_{\text{SSIM}} + \lambda_4 \mathcal{L}_{\text{LPIPS}}, \quad (11)$$

$\mathcal{L}_{\text{pose}}$ is the distillation loss in [11] to regularize the camera pose estimation and mitigate the adverse effects of inconsistent inputs during training. The corresponding balancing weights are set to $\lambda_i = \{10, 1.0, 0.05, 0.05\}$, respectively.

4 Experiments

4.1 Experimental Setup

Dataset. We categorize our training datasets into two types: consistent appearance and diverse appearance. To establish geometric priors, we use the DL3DV dataset [16], which provides strict photometric consistency. For diverse appearance, we incorporate MegaScenes [30] and MegaDepth [15]. Because these

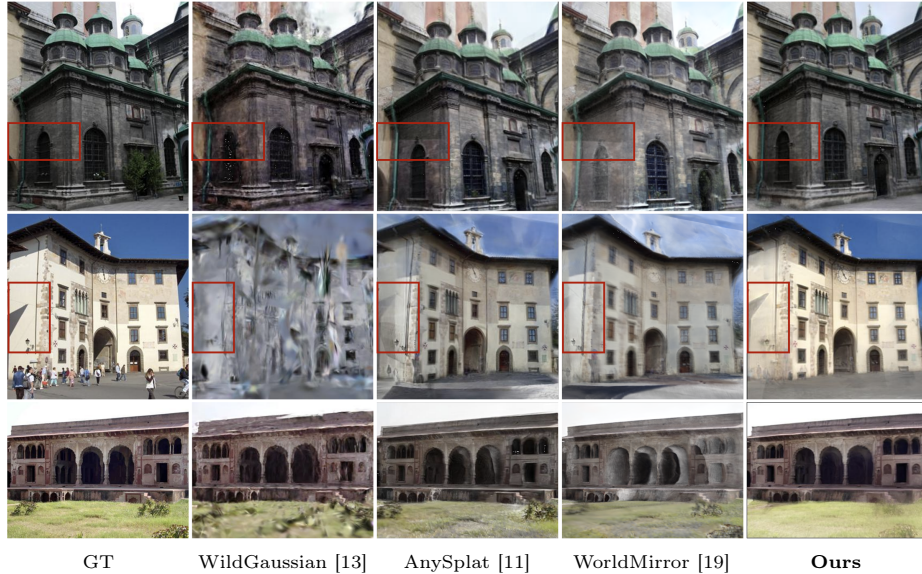


Fig. 3: Qualitative comparison on the MegaScenes dataset [30]. We visualize novel view synthesis results under the 12-view setting. Due to sparse input views, the optimization-based method WildGaussian produces many artifacts. Feedforward methods (AnySplat and WorldMirror) synthesize building exteriors with inconsistent colors (as highlighted in the red box) due to varying lighting conditions in the input images. In contrast, our method reconstructs realistic renderings with consistent appearance.

in-the-wild datasets lack multiple appearances from identical viewpoints, we augment sampled videos from DL3DV using an off-the-shelf video relighting model [18]. This generates a multi-view, multi-appearance dataset with paired data under various synthetic lighting conditions, facilitating appearance disentanglement. For view sampling, we adopt interval-based sampling for DL3DV and geometry-guided view sampling for MegaDepth and MegaScenes.

To evaluate novel-view synthesis in scenes exhibiting complex appearance variations, following previous works [13, 20, 41], we test our method on four scenes from the Phototourism [28] dataset. In addition, we further use ten scenes sampled from the MegaScenes [30] test set to increase scene diversity. All evaluations are performed under 4-, 8-, and 12-view configurations using a fixed test set.

Baselines. We compare our approach against recent state-of-the-art methods based on Gaussian Splatting. We categorize these baselines into two main paradigms: feedforward and optimization-based methods. For feedforward approaches, we evaluate against AnySplat [11] and WorldMirror [19]. At evaluation time, we employ an evaluation-time pose alignment [37] in which the Gaussian parameters are kept fixed while the target camera pose is optimized using a photometric loss, applied consistently to both the baselines and our method.



Fig. 4: Qualitative comparison on the Phototourism dataset [28]. We visualize novel-view synthesis results under the 12-view setting. Compared with existing methods, our method successfully renders photorealistic and appearance-consistent images.

For optimization-based methods, we consider WildGaussian [13], GS-W [41], and FSGS [43]. Since these optimization-based approaches explicitly depend on accurate camera poses and dense point clouds, we use VGGT [32] to estimate the required geometric priors. We further ensure that all target views are correctly aligned to the input image coordinate system to enable a fair and controlled comparison. To prevent photometric information leakage, we follow the evaluation protocol established by NeRF-W [20]. Specifically, we use the left half of the ground-truth image as the reference and compute quantitative metrics solely on the right half.

Implementation details. We implement our framework in PyTorch using the Adam optimizer. Our geometry branch comprises a DINOv2-large encoder ($D = 1024$) and $L = 24$ alternating-attention layers. The appearance branch consists of a pretrained DINOv2-base encoder coupled with a 4-layer appearance

Table 3: Ablation Study. We conduct an ablation study on the Phototourism dataset [28] to evaluate the contribution of individual components. Quantitative results, including PSNR ($\uparrow$), SSIM ($\uparrow$), and LPIPS ($\downarrow$), are reported for sparse input configurations of 4, 8, and 12 views.

Setting	4 views			8 views			12 views		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o Multi-Ref Supervision	16.88	0.668	0.221	18.77	0.723	0.184	19.35	0.746	0.177
w/o Geo-guided Sampling	17.99	0.671	0.220	19.24	0.709	0.192	19.58	0.731	0.182
w/o Pre-Modulation	18.25	0.690	0.226	19.70	0.738	0.189	20.11	0.757	0.175
Full	18.52	0.705	0.212	19.90	0.748	0.180	20.15	0.766	0.169

injector. For initialization, we load pretrained VGGT weights into the geometry backbone, apply zero-initialization to the AdaLN-Zero layers, and initialize all remaining parameters randomly. We train the model across four NVIDIA RTX H20 GPUs in two stages. First, we conduct a 30k-iteration warm-up on the DL3DV dataset with the appearance injector disabled to stabilize the Gaussian attribute heads. This is followed by 60k iterations of end-to-end training on the complete dataset. More experimental details are provided in the supplementary material.

4.2 Main Results

Qualitative Results. Fig. 3 and Fig. 4 show the visual comparisons between our method and the baselines on MegaScenes [30] and Phototourism [28], respectively. The results demonstrate that, given a single reference image for appearance guidance, our method achieves the best rendering performance with consistent appearance. Although WildGaussian can optimize an appearance embedding using a reference image, it produces many artifacts under sparse input views and requires long per-scene optimization times. Both AnySplat and WorldMirror are feedforward methods that cannot decouple geometry and appearance. Consequently, due to varying lighting conditions in the input images, the reconstructed buildings exhibit mixed and inconsistent appearances.

Quantitative Results. Tabs. 1 and 2 report the PSNR, SSIM, and LPIPS of all methods on MegaScenes and Phototourism, respectively. Since both our method and optimization-based methods require a reference image to specify the reconstruction appearance, we split each ground-truth image in half. The left half is used as the reference input, while the right half is used to compute the metrics for the rendered results. The results clearly demonstrate the superiority of our method in all metrics. Moreover, as the number of input images increases, the performance of our method steadily improves, without the drastic fluctuations observed in WorldMirror. Optimization-based methods (FSGS, GS-W, WildGaussian) produce the worst results due to a lack of sufficient scene images for optimization. Feedforward methods, including AnySplat and WorldMirror, are hampered by the diversity of lighting conditions in the input images, resulting in appearances that are inconsistent with the target scene.



Fig. 5: Appearance transfer across diverse conditions. The left column displays the target content scenes, and the top row provides various reference images captured under different lighting, weather, and times of day. Even when the references depict different building types, our model effectively transfers the overall color tone and appearance conditions while strictly preserving the underlying geometry.

Pose Evaluation. We further evaluate camera pose estimation on 20 scenes sampled from MegaScenes, with each scene evaluated using 24 views. As reported in Tab. 4, we compare methods using AUC@5, AUC@10, AUC@30, and ATE metrics. Our method improves over AnySplat across the AUC metrics and achieves results close to VGGT, a dedicated and strong pose prior. These results show that our geometry branch is able to learn meaningful camera poses.

Table 4: Pose evaluation. We report the metrics on MegaScenes and highlight the best and second-best results in **bold** and underline, respectively.

Method	AUC@5 $\uparrow$	AUC@10 $\uparrow$	AUC@30 $\uparrow$	ATE $\downarrow$
VGGT	0.314	0.451	0.632	1.525
AnySplat	0.249	0.392	0.596	1.564
Ours	<u>0.310</u>	<u>0.443</u>	<u>0.624</u>	<u>1.559</u>

4.3 Ablation Study

To evaluate the design choices of WildSplat, we conduct an ablation study on the Phototourism dataset [28] using 4-, 8-, and 12-view settings, with quantitative results summarized in Tab. 3. Decoupling geometry from appearance is essential for in-the-wild scenes. Removing the joint multi-reference training strategy (w/o Multi-Ref Supervision) results in the most significant performance drop across all metrics. Using only a single reference view per iteration tends to entangle the 3D geometry with specific lighting conditions, causing unstable optimization during training. By simultaneously rendering multiple appearance conditions, our method maintains training stability and disentangles the underlying geometry and appearance. Furthermore, traditional pose-ranking methods often struggle with the extreme variations in viewpoint and focal length typical of unconstrained photo collections. By explicitly accounting for spatial overlap and scale consistency, the geometry-guided sampling provides the network with more informative and reliable context views, significantly enhancing the quality of novel view



Fig. 6: Appearance interpolation. The grid demonstrates continuous interpolation across distinct reference appearances. By blending the extracted features, we achieve faithful and seamless transitions between the source images.

synthesis. Finally, omitting the global modulation scheme (w/o Pre-Modulation) degrades performance, demonstrating that using the global [CLS] token to pre-modulate intermediate content features provides an appearance prior that reduces interference from inconsistent appearance in the content features. Our full model achieves the best performance, indicating that these components work together to enable robust, appearance-conditioned synthesis from sparse inputs.

4.4 Appearance Editing

Since our method factorizes the reconstruction process into illumination-agnostic geometry extraction and appearance-aware conditioning, we can adjust the reference image to inject different appearance features into the same scene. As shown in Fig. 5, we present qualitative results of our method using reference images under varying weather conditions, times of day, and even across different building types. Our method successfully captures the overall color tone from the reference image and transfers it to the target scene. Thanks to the decoupled geometry and color modeling architecture, combined with a multi-reference training paradigm tailored for photometrically inconsistent data, our method can distinguish between the background and the main building in the reference image, accurately transferring their colors to the corresponding regions in the target scene. Fig. 6 further illustrates appearance interpolation between given reference images. By interpolating the extracted features, we achieve smooth visual blending, resulting in seamless transitions across different lighting and style conditions. Finally, our feedforward model eliminates the need for redundant

per-scene optimization, making the approach highly efficient and practical for real-world scene style transfer tasks.

5 Conclusion

In this work, we presented WildSplat, the first feedforward 3D Gaussian Splatting framework designed for unposed in-the-wild images exhibiting severe appearance variations. To address the inherent photometric inconsistencies of such data, we proposed a dual-branch architecture that explicitly decouples geometry from appearance. The geometry branch leverages a VGGT-based backbone to jointly estimate camera poses and extract an illumination-agnostic, shared 3D geometry. Our appearance branch employs an appearance injector module that utilizes feature pre-modulation and cross-attention to seamlessly inject target reference information into the content features. Extensive experiments demonstrate that WildSplat achieves state-of-the-art photorealistic novel view synthesis for scenes with various appearances, while also enabling faithful appearance editing.

Acknowledgements

This work was partially supported by NSF of China (No. 62425209).

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In: *Int. Conf. Comput. Vis.* pp. 5855–5864 (2021)
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 5470–5479 (2022)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-NeRF: Anti-aliased grid-based neural radiance fields. In: *Int. Conf. Comput. Vis.* pp. 19697–19705 (2023)
4. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 19457–19467 (2024)
5. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: MVSNerF: Fast generalizable radiance field reconstruction from multi-view stereo. In: *Int. Conf. Comput. Vis.* pp. 14124–14133 (2021)
6. Chen, Y., Xu, H., Wu, Q., Zheng, C., Cham, T.J., Cai, J.: Explicit correspondence matching for generalizable neural radiance fields. *IEEE Trans. Pattern Anal. Mach. Intell.* (2025). <https://doi.org/10.1109/TPAMI.2025.3598711>
7. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In: *Eur. Conf. Comput. Vis.* pp. 370–386. Springer (2024)
8. Dahmani, H., Bennehar, M., Piasco, N., Roldao, L., Tsishkou, D.: SWAG: Splatting in the wild images with appearance-conditioned gaussians. In: *Eur. Conf. Comput. Vis.* pp. 325–340. Springer (2024)

9. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. pp. 11–20. SIGGRAPH '96, Association for Computing Machinery, New York, NY, USA (1996)
10. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 4970–4980 (2023)
11. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: AnySplat: Feed-forward 3D gaussian splatting from unconstrained views. *ACM Trans. Graph.* **44**(6), 1–16 (2025)
12. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
13. Kulhanek, J., Peng, S., Kukulova, Z., Pollefeys, M., Sattler, T.: WildGaussians: 3D gaussian splatting in the wild. In: *Adv. Neural Inform. Process. Syst.* (2024)
14. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. *Int. J. Comput. Vis.* **38**(3), 199–218 (2000)
15. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2041–2050 (2018)
16. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 22160–22169 (2024)
17. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: *Int. Conf. Comput. Vis.* pp. 7708–7717 (2019)
18. Liu, Y., Luo, C., Tang, Z., Li, Y., Yang, Y., Ning, Y., Fan, L., Peng, J., Zhang, Z.: TC-Light: Temporally coherent generative rendering for realistic world transfer. *arXiv preprint arXiv:2506.18904* (2025)
19. Liu, Y., Min, Z., Wang, Z., Wu, J., Wang, T., Yuan, Y., Luo, Y., Guo, C.: World-Mirror: Universal 3D world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726* (2025)
20. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the wild: Neural radiance fields for unconstrained photo collections. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 7210–7219 (2021)
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
22. Miyato, T., Jaeger, B., Welling, M., Geiger, A.: GTA: A geometry-aware attention mechanism for multi-view transformers. *arXiv preprint arXiv:2310.10375* (2023)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
24. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *Int. Conf. Comput. Vis.* pp. 4195–4205 (2023)
25. Sajjadi, M.S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., et al.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 6229–6238 (2022)
26. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 4104–4113 (2016)
27. Seitz, S.M., Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. *Int. J. Comput. Vis.* **35**(2), 151–173 (1999)

28. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM SIGGRAPH 2006 Papers, pp. 835–846. Association for Computing Machinery (2006)
29. Szeliski, R., Golland, P.: Stereo matching with transparency and matting. *Int. J. Comput. Vis.* **32**(1), 45–61 (1999)
30. Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B., Snavely, N.: MegaScenes: Scene-level view synthesis at scale. In: *Eur. Conf. Comput. Vis.* pp. 197–214. Springer (2024)
31. Waechter, M., Moehrle, N., Goesele, M.: Let there be color! large-scale texturing of 3d reconstructions. In: *Eur. Conf. Comput. Vis.* pp. 836–850. Springer (2014)
32. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 5294–5306 (2025)
33. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUST3R: Geometric 3D vision made easy. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 20697–20709 (2024)
34. Xu, C., Kerr, J., Kanazawa, A.: Splatfacto-W: A Nerfstudio implementation of gaussian splatting for unconstrained photo collections. *arXiv preprint arXiv:2407.12306* (2024)
35. Xu, H., Chen, A., Chen, Y., Sakaridis, C., Zhang, Y., Pollefeys, M., Geiger, A., Yu, F.: MuRF: Multi-baseline radiance fields. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 20041–20050 (2024)
36. Xu, J., Mei, Y., Patel, V.: Wild-GS: Real-time novel view synthesis from unconstrained photo collections. *Adv. Neural Inform. Process. Syst.* **37**, 103334–103355 (2024)
37. Ye, B., Liu, S., Xu, H., Xueting, L., Pollefeys, M., Yang, M.H., Songyou, P.: NoPoSplat: No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207* (2024)
38. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 4578–4587 (2021)
39. Zhai, H., Zhang, Q., Pan, X., Zhang, X., Dong, Y., Zhang, H., Xu, D., Zhang, G.: OnlinePG: Online open-vocabulary panoptic mapping with 3D gaussian splatting. *arXiv preprint arXiv:2603.18510* (2026), CVPR 2026
40. Zhai, H., Zhang, X., Zhao, B., Li, H., He, Y., Cui, Z., Bao, H., Zhang, G.: SplatLoc: 3D gaussian splatting-based visual localization for augmented reality. *arXiv preprint arXiv:2409.14067* (2024)
41. Zhang, D., Wang, C., Wang, W., Li, P., Qin, M., Wang, H.: Gaussian in the wild: 3D gaussian splatting for unconstrained image collections. In: *Eur. Conf. Comput. Vis.* pp. 341–359. Springer (2024)
42. Zhang, X., Bao, C., Chen, Y., Zhai, H., Dong, Y., Bao, H., Cui, Z., Zhang, G.: AtlasGS: Atlanta-world guided surface reconstruction with implicit structured gaussians. *arXiv preprint arXiv:2510.25129* (2025), NeurIPS 2025
43. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: FSGS: Real-time few-shot view synthesis using gaussian splatting. In: *Eur. Conf. Comput. Vis.* pp. 145–163. Springer (2024)