UV-RGS: Relightable 3D Gaussian Splatting from Unposed Views under Varied Illuminations

Wei Feng, Chi Huang, Qi Zhang*, Qian Zhang, Nan Li

[1]School of Computer Science and Technology, Tianjin University, China
[2]Key Research Center for Surface Monitoring and Analysis of Relics, State Administration of Cultural Heritage
{wfeng, chi_huang, qizhang118, qianz, linan94}@tju.edu.cn

Abstract

The latest advancements in scene relighting have been predominantly driven by inverse rendering with 3D Gaussian Splatting (3DGS). However, existing methods remain overly reliant on precise camera parameters under static illumination conditions, which is prohibitively expensive and even impractical in real-world scenarios. In this paper, we propose a novel learning from Unposed views under Varied illuminations Relightable 3D Gaussian Splatting (dubbed UV-RGS), to address this challenge by jointly optimizing camera poses, 3DGS representations, surface materials, and environment illuminations (i.e., unknown and varied lighting conditions in training) using only unposed views under varied lightings. Firstly, UV-RGS presents a viewpoint dividing strategy to group inputs into constituent units, enabling each unit can perform similar poses and illuminations. Next, for each unit, to get the constituent model, UV-RGS establishes an incrementally pose learning module to estimate coarse camera parameters, which also enjoy a proxy-view refinement to alleviate the sparse view learning. Additionally, for all constituent unit models, we introduce a holistic model learning strategy that integrates progressive unit aggregation component and the 3DGS coupled with camera poses joint optimization, which realizes the scene high-fidelity perception by the physical-based rendering. Extensive experiments on both real-world and synthetic challenging datasets demonstrate the effectiveness of UV-RGS, achieving the state-of-the-art performance for scene inverse rendering by learning 3DGS from only unposed views under varied illuminations.

Introduction

Scene relighting stands as a long-standing challenge in computer vision and computer graphics. Neural Radiance Fields (NeRF) address this via inverse rendering, leveraging MLPs to predict scene geometry and material properties (Rudnev et al. 2022). However, the substantial computational demands of neural networks hinder training and rendering efficiency, limiting practical deployment.

Recently, 3D Gaussian Splatting (3DGS) has emerged as a compelling alternative, offering highly efficient training and real-time rendering (Kerbl et al. 2023). Current 3DGS-based relighting approaches primarily follow two paradigms. The first couples appearance embeddings derived from input images with Gaussian primitives, which rendering with a specific reference embedding then produces novel views under altered illumination (Zhang et al. 2024a; Dahmani et al. 2024; Wang, Wang, and Qi 2024). The second paradigm augments each Gaussian primitive with learnable material properties (e.g., albedo and metallic), enabling intrinsic scene decomposition during optimization (Liang et al. 2024; Sun et al. 2025; Zhang et al. 2025). This decomposition, combined with the physical-based rendering (PBR) formulation, facilitates relighting under novel illumination. Crucially, the latter approach models scene lighting using parameterized representations such as learnable Spherical Harmonic (SH) coefficients or environment maps. This enables precise illumination control and offers significant potential for explainable scene relighting. Consequently, our work focuses on this inverse rendering paradigm leveraging 3DGS for decomposition of intrinsic scene properties.

Methods combining 3DGS with inverse rendering for relighting have recently demonstrated impressive results (Kaleta et al. 2024). Notable examples include R3DG (Gao et al. 2024), which enables photo-realistic object and scene relighting via a point-based ray tracing with scene decomposition. GaussianShader (Jiang et al. 2024) integrates shading functions with 3D Gaussian representation

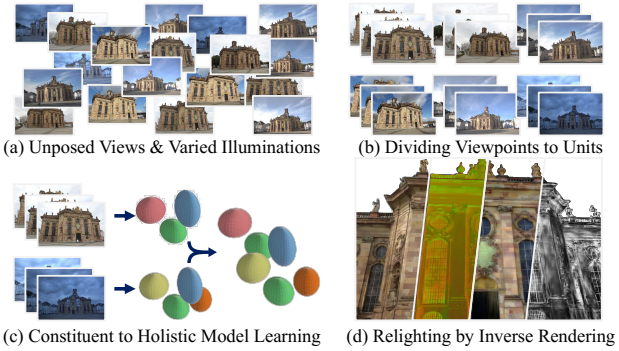

Figure 1: With inputting unposed views under varied illuminations, UV-RGS achieves the state-of-the-art relighting quality by dividing inputs to units and the constituent to holistic model joint learning strategies.

*Corresponding Author

for handling reflective surfaces. GS-IR (Liang et al. 2024) facilitates high-quality relighting by the depth-derivation-based regularization for normal estimation and a baking-based occlusion to model indirect lighting. SVG-IR (Sun et al. 2025) implements inverse rendering using spatially-varying Gaussian splatting via a physically-based indirect illumination modeling and a curve Gaussian rendering. However, a critical limitation unites these approaches: their stringent dependence on precisely calibrated camera poses and the requirement for image capture under strictly controlled and constant illumination. These prerequisites prove prohibitively restrictive in real-world unconstrained environments, often making data collection impractical or infeasible. Moreover, while methods relying on appearance embeddings can accommodate arbitrary illumination during optimization using auxiliary networks, they remain fundamentally constrained by the need for highly accurate camera pose estimation, which poses significant barriers to deployment in the practical wild scenarios.

In this paper, we present the Relightable 3D Gaussian Splatting from only Unposed views with Varied illuminations (dubbed UV-RGS) that relighting scene through the inverse rendering. To the best of our knowledge, our method is the first work to introduce the 3DGS to adapt unposed views under varied illuminations inputs for inverse rendering, which can simultaneously estimate scene geometry, materials, and environment light. We illustrate the framework of the UV-RGS in Fig. 1 and Fig. 2, where the whole pipeline consists of the Viewpoint Dividing, Constituent Model Learning, and Holistic Model Learning. (1) For Viewpoint Dividing, we present a hierarchical strategy to group the multi-view multi-illumination images into the constituent units. Each unit employs the approximately similar viewpoint under the static illuminations, which alleviates the perturbation of various lighting for 3DGS. (2) For Constituent Model Learning, we propose an incremental pose optimization module to learn camera poses for each constituent unit. Also, a proxy-view construction strategy is elaborated to deal with the few-shot views 3DGS during constituent model learning. (3) For Holistic Model Learning, we aggregate all constituent unit models and joint optimize camera poses coupled with 3DGS for scene inverse rendering, which enjoys an efficient fusion and a global learning strategies. Till here, UV-RGS realize to optimize 3DGS from unposed views and varied illuminations.

We conduct extensive evaluations of UV-RGS and the state-of-the-art baselines, where UV-RGS effectively outperforms various methods to achieve the SOTA performance on the NeRF-OSR (Rudnev et al. 2022) and TensoIR Synthetic (Jin et al. 2023) datasets from unposed views under varied illuminations with relightable 3DGS optimization. The main contributions are as follows:

- We advocate the idea of UV-RGS to optimize a relightable 3DGS inverse rendering model only from unposed views under varied illuminations.
- We propose a hierarchical viewpoint dividing strategy, which coupled with constituent to holistic incremental learning module, to alleviate 3DGS from unposed and varied unknown illuminations.
- We elaborate on extra effectiveness optimization of UV-RGS, the proxy-view training strategy, for constructing additional supervisions by pseudo-labels, to facilitate 3DGS from constituent units with few-shot inputs.
- We conduct comprehensive analysis on UV-RGS, which achieves the SOTA performance on several challenge datasets, demonstrating the effectiveness and practicability of our method.

Related Work

Radiance Fields from Unposed Views Learning radiance fields from unposed views has garnered significant attention due to its practical importance. Methods based on NeRF's powerful implicit representation (Mildenhall et al. 2020; Zhang et al. 2024b) treat camera poses as learnable parameters, enabling gradient computation from photometric discrepancies between predicted and rendered images (Wang et al. 2024; Lin et al. 2021). These gradients encapsulate both pose estimation errors and rendering inaccuracies. The widespread adoption of deep learning frameworks has spurred considerable interest in unposed-view NeRF research (Cheng et al. 2023; Kim, Choi, and Kim 2023; Zhang et al. 2023).

Recently, 3DGS has gained prominence for real-time rendering (Kerbl et al. 2023). However, unlike NeRF, 3DGS faces challenges in unposed learning due to its lack of MLP's powerful fitting capability and the requirement for carefully designed densification strategies (Kang et al. 2025; Cai et al. 2024; Li et al. 2024). Consequently, research remains limited, with preliminary attempts like CF-3DGS (Fu et al. 2024) and SFGS (Ji and Yao 2025) demonstrating pose-free optimization only on video frames with minimal viewpoint variation. While TriGS (Huang et al. 2025) extends unposed 3DGS to sparse settings, it still assumes static illumination. Furthermore, real-world scenarios involve uncalibrated cameras and varying illumination conditions, which none of the aforementioned methods address effectively.

In contrast, our hierarchical viewpoint dividing strategy mitigates illumination discrepancies, and combined with constituent-to-holistic incremental learning, enables robust scene relighting without pose priors.

Radiance Fields under Varied Illumination Learning radiance fields from images captured under varied illuminations presents another significant challenge in practical applications. For both NeRF and 3DGS, the predominant solution involves encoding an appearance embedding from the input images (Martin-Brualla et al. 2021; Wang, Wang, and Qi 2024). This embedding is then fused with spatial point color representations, typically via an MLP or other feature extractor, introducing additional parameters to model varying appearance - an approach often termed "NeRF/3DGS in-the-wild" (Zhang et al. 2024a; Dahmani et al. 2024). While effective for synthesizing novel views under different lighting, these methods lack the ability to parameterize scene illumination (Xu, Mei, and Patel 2024). Consequently, they are fundamentally limited in achieving interpretable relight-

ing, offering only visual alterations to scene lighting rather than a physically grounded decomposition. Although methods like NeRF-OSR (Rudnev et al. 2022) introduce multiple sub-MLPs to predict surface color, material properties, and density separately, they share a critical limitation with other "in-the-wild" approaches: they rely on precisely calibrated camera poses. Without accurate poses, high-fidelity modeling of scene geometry and appearance, especially under complex illumination, becomes exceedingly difficult.

Different from the above researches, we present a novel relightable 3D Gaussian Splatting framework that achieves high-fidelity inverse rendering directly from unposed images captured under varied illumination.

Preliminary

Standard Rendering for 3DGS 3D Gaussian splatting defined all Gaussians by a full 3D covariance matrix Σ in world space, which centered at point μ as follows:

$$\mathcal{G}(x) = e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(x-\mu)}, \tag{1}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{3\times3}$ is the anisotropic covariance matrix. $\mu \in \mathbb{R}^3$ denotes the mean vector. Specifically, the covariance matrix $\mathbf{\Sigma} = \mathbf{RSS}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}$ can be factorized into a scaling matrix $\mathbf{S}$ and rotation matrix $\mathbf{R}$. Notably, each view has the camera external parameters $\mathbf{T} \in SE(3)$ and camera internal parameters $\mathbf{K} \in \mathbb{R}^{3\times3}$. The 2D covariance matrix $\mathbf{\Sigma}^{'}$ is as:

$$\mathbf{\Sigma}^{'} = \mathbf{JW\Sigma W}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{J}$ is the Jacobian of the affine approximation of the projective transformation. $\mathbf{W}$ is the rotation matrix. Moreover, each Gaussian has the color $\mathbf{c}$ by a spherical harmonic (SH) coefficient. Finally, the pixel color $\hat{C}$ is rendered as:

$$\hat{C} = \sum_{i\in\mathcal{N}} \mathcal{G}_i\alpha_i\mathbf{c}_i \prod_{j=1}^{i-1}(1-\mathcal{G}_i\alpha_j), \tag{3}$$

where $\prod_{j=1}^{i-1}(1-\mathcal{G}_i\alpha_j)$ is the accumulated transmittance T_i. $\mathcal{N}$ is the set of Gaussians that the current ray traces.

Inverse Rendering for 3DGS For UV-RGS, we follow (Liang et al. 2024) to employ the physical-based rendering (PBR) formulation to replace the standard volume rendering equation for modeling the ambient light interaction with complex surface (e.g., material and geometry properties). The details are as follows:

$$L_{\mathrm{o}}(\mathbf{z},\omega_{\mathrm{o}}) = \int_{\Omega} L_{\mathrm{i}}(\mathbf{z},\omega_{\mathrm{i}}) f_{\mathrm{r}}(\omega_{\mathrm{o}},\omega_{\mathrm{i}})(\omega_{\mathrm{i}}\cdot\mathbf{n})d\omega_{\mathrm{i}}, \tag{4}$$

where L_{i} and L_{o} are the radiance in incoming and outgoing directions, respectively. $f_{\mathrm{r}} = (1-m)\frac{a}{\pi} + \frac{DFG}{4(\mathbf{n}\cdot\omega_{\mathrm{i}})(\mathbf{n}\cdot\omega_{\mathrm{o}})}$ depicts the formulated bidirectional reflectance distribution function (BRDF), where $\mathbf{a} \in \mathbb{R}^3$ and $m \in \mathbb{R}$ are albedo and metallic of the surface. $\mathbf{z}$ and $\mathbf{n}$ denote the surface point and the corresponding normal. And the microfacet distribution function D, Fresnel reflection F, and geometric shadowing factor G are related to the surface roughness $\rho \in \mathbb{R}$. Ω is the hemispherical domain.

Method

Method Overview

This work aims to learn 3DGS model from unposed views under varied illuminations. Specifically, we address the challenging task of optimizing the 3DGS inverse rendering model for relighting. Given a set of RGB images $\{\mathbf{I}_i\}_{i=1}^{N}$ of a scene captured from unposed views, yet unknown and varied illuminations, motivated by the fascinating performance of 3DGS, we present a novel relighting framework UV-RGS, which can decompose the scene's intrinsic properties, including materials, normal, and illumination, to relight the scene by the inverse rendering. As illustrated in Fig. 2, UV-RGS consists of three well-designed modules. Firstly, we present viewpoint dividing strategy to group input views into constituent units, which is directed by the illumination representation and camera correlation for alleviating various unknown lighting problem. Secondly, we propose pose incremental learning for each constituent unit to estimate camera parameters, which also constructs proxy-view pseudo-labels to refine the camera poses and constituent models. Thirdly, UV-RGS will progressively aggregate all constituent models and joint optimize 3DGS coupled with camera poses by physical-based rendering to adapt scene inverse rendering. Till here, UV-RGS realize to decompose scene intrinsic properties by a relightable 3DGS only from unposed views under varied illuminations.

Viewpoint Dividing

Scene Inverse Rendering with 3DGS broadly falls into two categories. The first employs traditional volume rendering to learn geometry first, then followed by PBR-based scene decomposition. The second directly utilizes PBR for scene decomposition. Notably, the latter often yields suboptimal results because of introducing excessive learnable parameters during early training stages. Consequently, the former remains the dominant approach. Therefore, a critical challenge arises under varied illumination: 3D Gaussians must represent varying colors for the same viewpoint. However, standard volume rendering struggles to converge under such conflicting supervision signals. To address this, we propose a two-level grouping (i.e., illumination and camera) strategy to divide inputs into smaller units. Firstly, each unit contains images under nearly identical illumination, mitigating the impact of illumination variation. Secondly, images within a unit share closely similar image content, facilitating accurate camera pose estimation (Fu et al. 2024).

Dividing by Illumination Representation For the illumination grouping, our extensive experiments reveal a key insight: when image content is viewpoint-similar (e.g., multiple images depicting the frontal facade of Ludwigskirche with minor pose differences), the distributions of the H-channel values in the HSV color mode are remarkably consistent under the same illumination. This observation aligns with findings reported at (Feng et al. 2025). Leveraging this, we compute an image's H-channel distribution vector $\mathbf{f}_{\mathrm{h}} \in \mathbb{R}^K$ by grouping all values into K equally spaced bins and extract H-channel's deep feature $\mathbf{f}_{\mathrm{d}}$ by DINOv2 (Oquab et al. 2024). We then calculate pairwise similarities between

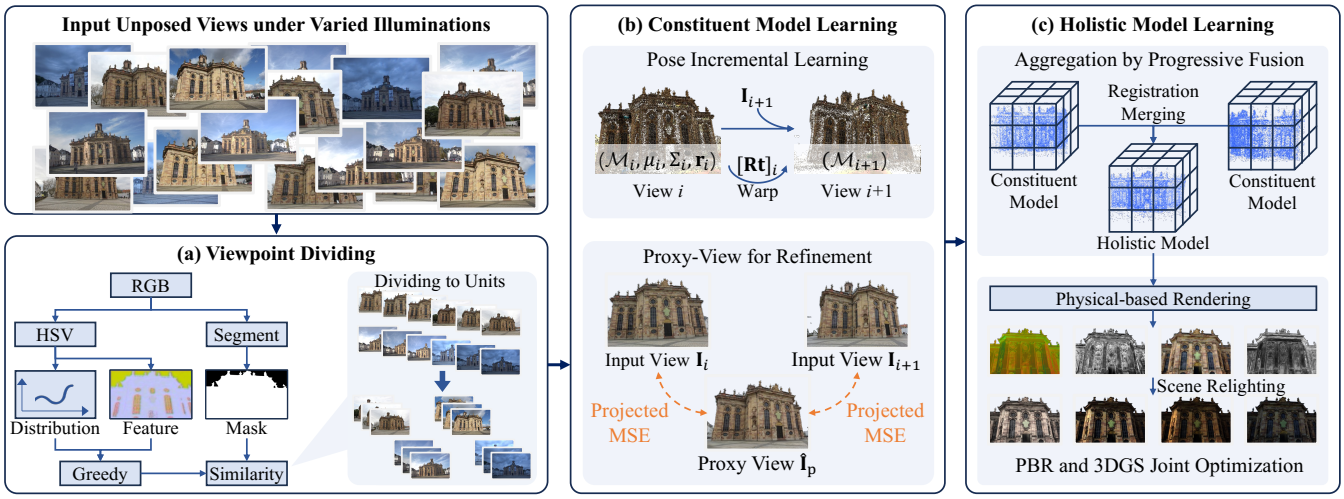

Figure 2: The framework of UV-RGS. (a) We first present a two-level viewpoint dividing strategy to group input views into units within extremely similar poses and lightings. (b) Then UV-RGS incrementally optimize coarse poses by warping 3DGS to next unit. Moreover, UV-RGS constructs proxy-view of units to establish pseudo-labels for constituent model refinement. (c) Finally, we progressively aggregate all constituent models and integrate PBR formulation for holistic model learning, which can learn 3DGS from only unposed views under varied illuminations.

images based on both $\mathbf{f}_\mathrm{h}$ and $\mathbf{f}_\mathrm{d}$. Finally, a greedy algorithm (Pokutta, Singh, and Torrico 2020) partitions all images into G groups of mutually consistent illumination, mitigating the impact of lighting variation on volume rendering.

Dividing by Camera Correlation Existing methods (e.g., CF-3DGS-based variants) estimate relative poses by rotating Gaussians's position, but this critically fails for non-sequential frames. To overcome this, we further subdivide each illumination group into camera units containing images with maximally similar viewpoint content. Specifically, we first segment image foregrounds using SAM2 (Ravi et al. 2024) to obtain masks. Then we compute the cosine similarity between these masks. And groups are partitioned into units based on this similarity. Each unit contains images with highly consistent illumination and viewpoint. Crucially, units within the same illumination group share L overlapping images, facilitating the later aggregation. Notably, computing similarity on foreground masks rather than RGB images is essential, as even minor illumination differences within a group could distort full-image similarity, potentially lowering the score for camera-similar images. This viewpoint dividing strategy yields units containing images closely matched in both illumination and camera.

Constituent Model Learning

Pose Incremental Learning Traditional 3DGS optimizes geometric and color properties of Gaussian primitives by computing and backpropagating the difference between rendered and ground-truth (GT) images. Nevertheless, its CUDA architecture is meticulously designed for efficient gradient computation and propagation. Inspired by CF-3DGS (Fu et al. 2024), UV-RGS introduce learnable 6-DoF relative camera poses between viewpoints, which allows replacing viewpoint switching with warping the Gaussian model to compute the errors of rendering, while incorporating global joint optimization for simultaneous model and pose refinement. This approach resolves CF-3DGS's limitation to dense video frames inputs and eliminates the need for specialized CUDA kernels, offering general applicability.

Concretely, we initialize all camera poses as identity matrices with zero translation. Relative poses between consecutive views are parameterized as learnable rotation $\mathbf{R} \in \mathbb{R}^{3\times3}$ and translation $\mathbf{t} \in \mathbb{R}^3$ matrices (e.g., $[\mathbf{Rt}]_1$-$[\mathbf{Rt}]_4$ matrices for 5 input views $\mathbf{I}_1$-$\mathbf{I}_5$). We first train a radiance field model $\mathcal{M}_1$ from $\mathbf{I}_1$ using standard volume rendering for 2K iterations. $\mathcal{M}_1$'s Gaussian splats (positions), shapes (rotation/scaling), view directions, and spherical harmonics query directions are then transformed via the relative pose $[\mathbf{Rt}]_1$ to the next viewpoint. Rendering this transformed model (instead of switching poses on a fixed model) enables direct gradient updates to $[\mathbf{Rt}]_1$. We then jointly optimize $\mathbf{I}_2$, $\mathcal{M}_1$, and $[\mathbf{Rt}]_1$ for 2K iterations to obtain $\mathcal{M}_2$ and $[\mathbf{Rt}]_2$. The detailed warping learning is as follows:

$$(\mathcal{M}_{i+1}, [\mathbf{Rt}]_i) = \mathrm{WARP}(\mathcal{M}_i, \mu_i, \mathbf{\Sigma}_i, \mathbf{r}_i), \tag{5}$$

where $\mathrm{WARP}(\cdot)$ is warping optimization. It can iterate to obtain all constituent models and camera poses within units.

Proxy-View for Refinement Although the pose incremental learning framework can reconstruct constituent models from unposed views under the static illumination within a unit, viewpoint dividing may yield few-shot units (e.g., only 5 images). This scarcity risks overfitting both the constituent model's rendering quality and its camera poses precision. To mitigate this challenge, we introduce proxy views during constituent model learning to provide additional pseudo-supervisions for refining model quality and pose estimation.

Specifically, given an input image set within a unit (e.g., $\mathbf{I}_1$-$\mathbf{I}_5$), we compute pairwise Euclidean distances between

all learned camera poses and select the top-N most distant view pairs (empirically N=3). For each camera pose pair ($\mathbf{T}_i$, $\mathbf{T}_j$), we interpolate their 6-DoF poses to generate a proxy view pose $\mathbf{T}_\mathrm{p}$. Using 2D-3D correspondences and the proxy view's rendered depth $\hat{D}_\mathrm{p}$, we can project the proxy-rendered image onto views $\mathbf{T}_i$ and $\mathbf{T}_j$, yielding projected images $\tilde{\mathbf{I}}_{\mathrm{p}\to i}$ and $\tilde{\mathbf{I}}_{\mathrm{p}\to j}$. The detailed transformation as:

$$\tilde{\mathbf{p}}_i = \pi(\hat{D}_i, \mathbf{K}_i, \mathbf{T}_i, \pi^{-1}(\mathbf{T}_\mathrm{p}, \hat{D}_\mathrm{p}, \mathbf{K}_\mathrm{p}, \mathbf{p}_\mathrm{p})), \tag{6}$$

where $\pi(\cdot)$ is the projection function. Notably, if the constituent model and poses are well-optimized, the photometric error between ($\tilde{\mathbf{I}}_{\mathrm{p}\to i}$, $\mathbf{I}_i$) and ($\tilde{\mathbf{I}}_{\mathrm{p}\to j}$, $\mathbf{I}_j$) approaches zero. We thus leverage gradients from this photometric loss to jointly refine the constituent model and camera poses, effectively regularizing against overfitting. This completes the optimization of all constituent models and camera poses.

Holistic Model Learning

While each unit and its corresponding constituent model obtained earlier characterize only local scene regions, we now incrementally fuse them to ultimately achieve physical-based rendering (PBR) under various illumination.

Aggregation by Progressive Fusion UV-RGS first fuses units pairwise under identical illumination, then merges the results across varying illuminations to form a complete 3DGS scene representation, with each fusion involving registration and merging.

Given two constituent models under the same illumination: (1) Registration. Leveraging the L common viewpoints required by our Viewpoint Dividing strategy, we align their coordinate systems using relative pose differences between shared views. Inconsistent relative poses are averaged to define the inter-model transformation. (2) Merging. To avoid coordinate misalignment and excessive storage from a direct union of Gaussian primitives, we first prune Gaussians beyond radius r from existing primitives and filter out those with opacity α below the threshold h. The 3D space containing all remaining Gaussians is partitioned into an $S \times S \times S$ voxel grid. Within each voxel, Gaussian positions (i.e., μ) and opacities (i.e., α) are merged via weighted averaging:

$$\mu_\mathrm{fused} = \frac{\sum_{i=1}^{Q} \alpha_i \mu_i}{\sum_{i=1}^{Q} \alpha_i}, \quad \alpha_\mathrm{fused} = \frac{1}{Q} \sum_{i=1}^{Q} \alpha_i, \tag{7}$$

where Q is the maximum index of Gaussian primitive in each voxel. Notably, the covariance matrix (i.e., $\mathbf{\Sigma}$) and spherical harmonics coefficients (SH) from the Gaussian with the highest alpha are retained. This progressively fuses all same-illumination constituent models pairwise into unified representations, leaving L illumination-variant models.

Moreover, we refine each of the same-illumination models via volume rendering for 1K iterations separately. After that, these models are then registered using ICP point cloud registration, followed by a merging procedure identical to the same-illumination case (e.g., position/alpha averaging). This completes the integration of all constituent models into a single and unified 3DGS representation of the full scene.

PBR and 3DGS Joint Optimization After fusion, we replace the traditional volume rendering formulation with physical-based rendering, enabling the holistic model to decompose scene from various illuminations. Specifically, we first use volume rendering to obtain per-pixel material properties: albedo, metallic, and roughness, respectively. These properties are then fed into Eq. 4 to compute the final rendered color. The material computation process for a 3D point corresponding to the rendered pixel is detailed below:

$$\hat{\theta} = \sum_{i \in \mathcal{N}} \mathcal{G}_i \alpha_i \theta_i \prod_{j=1}^{i-1} (1 - \mathcal{G}_i \alpha_j), \tag{8}$$

where θ can be depicted albedo, metallic, and roughness. At this stage, UV-RGS will also optimize camera relative poses and 3DGS model, then we can achieve a relightable 3DGS optimized from unposed views under varied illuminations.

Training Framework

Assembling all loss terms, for constituent model learning, the loss function is as follows:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_\mathrm{pho} + \lambda \mathcal{L}_{\mathrm{D-SSIM}}, \tag{9}$$

where $\mathcal{L}_\mathrm{pho}$ is the photometric loss between GT and rendered pixel color. $\mathcal{L}_{\mathrm{D-SSIM}}$ is a D-SSIM term. And $\lambda = 0.2$ in all tests. Also, the pseudo-labels of proxy-view are built on Eq. (9) between the rendered color of proxy-view and GT color of input view. For holistic model learning, the loss function is as the same as Eq. (9).

Experiment

Experimental Setup

Datasets and Metrics We evaluate UV-RGS and baselines on ***NeRF-OSR*** (Rudnev et al. 2022) and ***TensoIR Synthetic*** (Jin et al. 2023), which provide input views under varied illuminations. We follow the authors to set test set that mentioned in the published papers. For rendering quality, we follow (Wang et al. 2004) and (Zhang et al. 2018) to select PSNR, SSIM, and LPIPS metrics. Normal reconstruction accuracy is measured by Mean Angular Error (MAE) following (Liang et al. 2024) on TensoIR Synthetic. We also calculate the relative pose and depth errors using RPE (Bian et al. 2023) and MDAE (Zhang et al. 2024b).

Baselines We compare to four SOTA relightable radiance field baselines. ***TensoIR*** (Jin et al. 2023) is an inverse rendering approach based on tensor factorization and neural fields, which jointly achieves radiance field reconstruction and physically-based model estimation. ***GSIR*** (Liang et al. 2024) leverages forward mapping volume rendering to achieve photorealistic novel view synthesis and relighting results. ***SVG-IR*** (Sun et al. 2025) advocates on spatially-varying Gaussian splatting to yeild SOTA inverse rendering. ***NeRF-OSR*** (Rudnev et al. 2022) is a relightable NeRF under varied illuminations in outdoor scenes, which suffers from precise posed input views. Moreover, we also compare with ***CF-3DGS*** (Fu et al. 2024), ***GGRT*** (Li et al. 2024), and ***SFGS*** (Ji and Yao 2025) to evaluate the pose estimation.

Implementation Details We implement our framework based on previous work (Kerbl et al. 2023) in Python using

Method	NeRF OSR				TensoIR Synthetic			
	PSNR ↑	SSIM ↑	LPIPS ↓	MDAE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MDAE ↓
TensoIR	17.12	0.599	0.513	0.383	28.71	0.868	0.149	0.364
GSIR	18.50	0.711	0.354	0.301	29.18	0.883	0.130	0.267
SVG-IR	19.09	0.758	0.344	0.238	29.90	0.899	0.123	0.201
NeRF-OSR	17.61	0.620	0.485	0.372	28.82	0.873	0.141	0.366
UV-RGS	**21.69**	**0.884**	**0.208**	**0.197**	**31.95**	**0.953**	**0.053**	**0.177**

Table 1: The quantitative results of all methods for relighting novel view and geometry reconstruction in NeRF-OSR and TensoIR Synthetic datasets, respectively.

the PyTorch and wrote custom CUDA kernels for rasterization. For UV-RGS, we use Adam optimizer (Kingma and Ba 2015) for training. As for all baselines, we defer to the paper and source codes to conduct experiments with the same settings as UV-RGS, which are under varied illuminations and 10K iterations with about 1.5 hour on a single RTX 3090. Especially, we follow NeRF-OSR (Rudnev et al. 2022) and CF-3DGS (Fu et al. 2024) to modify TensoIR, GSIR, and SVG-IR for the unposed and varied illuminations inputting. All experiments were three times the average results.

Comparing under Real-World Scenes

As illustrated in Table 1, the quantitative results of UV-RGS outperform all baselines in all metrics by a large margin. Specifically, compared with the strong baseline SVG-IR, UV-RGS has increased by 2.60, 0.126, and 0.136 with respect to PSNR, SSIM, and LPIPS, respectively. The following conclusions can be made by analyzing the above results: (1) Although NeRF-OSR has achieved satisfactory results from varied illuminations in the published paper, it is worse than GSIR and SVG-IR with unposed views under varied illuminations in most instances. As NeRF-OSR neglects to model complex illumination (i.e., indirect lighting) with only albedo modeling. (2) UV-RGS has outperformed all baselines significantly, because the viewpoint dividing strategy alleviates the disturbance of varied unknown illuminations. Besides, it benefits from the constituent to holistic model learning to estimate camera poses, which also enjoys the pose and 3DGS model joint optimization to realize the high-fidelity scene modeling. These conclusions demonstrate that the hierarchical viewpoint dividing and the incremental optimization can couple with the proxy-view training strategy, to learn 3DGS from unposed views under varied unknown illuminations. As for the qualitative analysis of the scene relighting, decomposition, and geometry reconstruction, UV-RGS outperforms the SOTA relightable inverse rendering 3DGS models significantly from unposed and various lightings views, which are shown in Fig. 3. This further illustrates the superiority of the hierarchical viewpoint dividing and progressive incremental optimization framework to augment 3DGS for real-world scene inverse rendering. Notably, we omit quantitative analysis of scene decomposition and geometry on NeRF-OSR due to the absence of GT albedo, metallic, roughness, and normal labels.

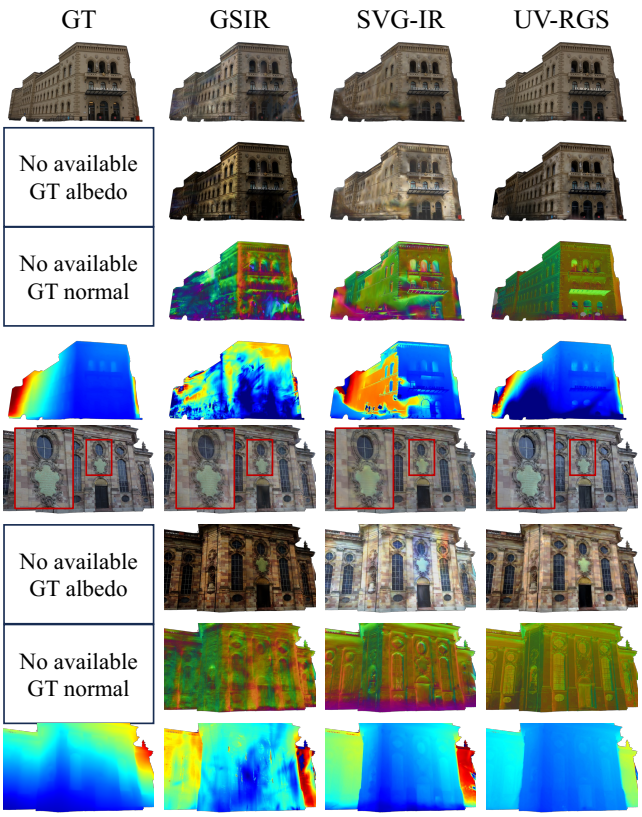

Figure 3: Visualization of relighting novel view, albedo, normal, and depth rendering in NeRF-OSR dataset.

Comparing under Synthetic Scenes

We evaluate UV-RGS and baselines on TensoIR Synthetic with unposed views under varied illuminations.

As illustrated in Table 1 and 2, the quantitative results perform a similar trend to NeRF-OSR dataset. Specifically, compared with SVG-IR, UV-RGS has improved PSNR, SSIM, LPIPS, and MAE with 3.52, 0.071, 0.102, and 0.423 in NVS, respectively. Also, the consistent performance improvements can also be observed in albedo rendering. This phenomenon further verifies the effectiveness of the presented hierarchical viewpoint dividing strategy for alleviating varied illuminations, and the incremental constituent to holistic 3DGS model and camera poses joint learning

Metrics		TensoIR	GSIR	SVG-IR	NeRF-OSR	UV-RGS
Normal	MAE ↓	5.750	5.326	5.197	5.415	**4.774**
NVS	PSNR ↑	29.01	29.34	30.16	29.07	**33.68**
	SSIM ↑	0.875	0.892	0.909	0.887	**0.980**
	LPIPS ↓	0.162	0.150	0.146	0.155	**0.044**
AR	PSNR ↑	28.90	29.26	29.91	28.96	**32.44**
	SSIM ↑	0.871	0.887	0.904	0.879	**0.906**
	LPIPS ↓	0.154	0.135	0.138	0.148	**0.051**

Table 2: The quantitative results of all methods for normal estimation, novel view synthesis (NVS), and albedo rendering (AR) in TensoIR Synthetic dataset, respectively.

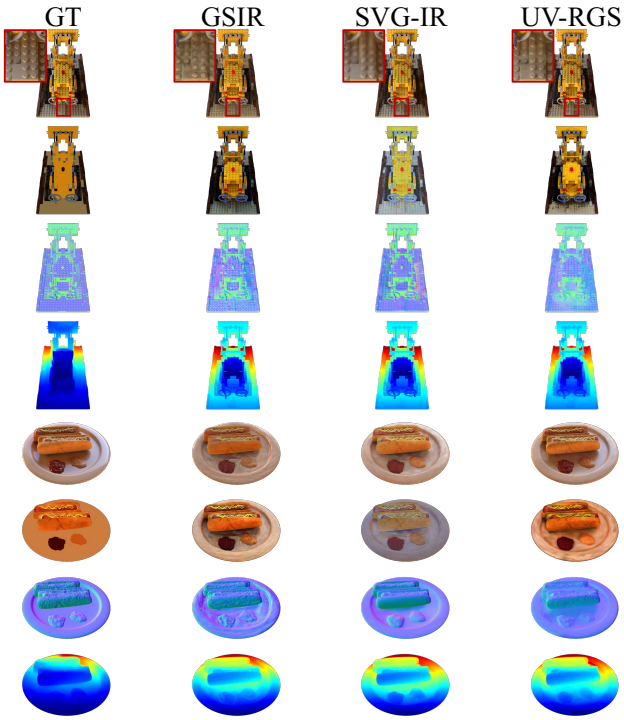

Figure 4: Visualization of scene relighting, decomposition, normal, and depth rendering in TensoIR Synthetic.

framework. For the qualitative analysis of the scene relighting, decomposition, and geometry reconstruction, as shown in Fig. 4, UV-RGS outperforms the SOTA methods consistently. This illustrates the robustness and generalization of UV-RGS to perform satisfactory inverse rendering in both synthetic and real-world scenarios.

Method Analysis

Effect of Key Components of UV-RGS We ablate the key component of our approach with novel view relighting, here eliminating Viewpoint Dividing (VD), Constituent Model Learning (CML), and Holistic Model Learning (HML), respectively. According to the results in Table 3, we observe that the performances from the first to the last increase significantly on scene relighting task. It demonstrates the presented key components (i.e., VD, CML, and HML) can effectively assist UV-RGS to fulfill learning relightable 3DGS from unposed views under varied illuminations.

Variants	VD	CML	HML	PSNR↑	SSIM↑	LPIPS↓
1				17.55	0.616	0.477
2			✓	19.96	0.758	0.339
3	✓	✓		20.35	0.740	0.328
4	✓	✓	✓	**21.69**	**0.884**	**0.208**

Table 3: Ablation analysis of UV-RGS from unposed views and varied illuminations.

Method	PSNR↑	SSIM↑	LPIPS↓	RPE_R ↓	RPE_t ↓
CF-3DGS	17.55	0.616	0.477	3.395	3.661
GGRT	18.57	0.695	0.419	2.008	1.899
SFGS	19.15	0.744	0.396	1.392	1.446
UV-RGS	**21.69**	**0.884**	**0.208**	**1.073**	**1.180**

Table 4: Comparison of unposed baselines and UV-RGS.

Evaluation of Camera Poses Estimation In Table 4, we compare UV-RGS with SOTA unposed baselines (i.e., CF-3DGS (Fu et al. 2024), GGRT (Li et al. 2024), and SFGS (Ji and Yao 2025)) under varied illuminations. We observe that all baselines suffer from varied lighting conditions, which can not achieve as better results as they did in posed inputs view settings. However, UV-RGS can meet the problem by the viewpoint dividing and constituent to holistic model learning strategies to learn 3DGS from unposed views under varied illuminations. It assists 3DGS to deploy in real-world applications significantly.

Conclusion

Current relightable 3DGS for scene inverse rendering struggles with pre-calibrated camera poses and static illuminations. This work presents a relightable 3DGS from only unposed views under varied illuminations (dubbed UV-RGS). We first present viewpoint dividing strategy to hierarchically group input views for alleviating varied unknown illuminations. Specifically, UV-RGS constructs constituent to holistic model optimization pipeline, which not only realize the camera poses incremental learning, but also integrate proxy-view pseudo-labels to refine the quality of 3DGS. Moreover, the presented progressive fusion strategy assists UV-RGS to realize 3DGS and camera poses joint optimization on scene inverse rendering. We evaluate UV-RGS and state-of-the-art methods in three challenging scenarios. The experiments verify the effectiveness and robustness of UV-RGS. We also plan to extend our strategy to other challenging applications of 3DGS (e.g., real-time relighting).

Acknowledgements

This work was supported in part by the National Key R&D Program of China (Grant 2023YFF0906200), in part by the Natural Science Foundation of China (Grants 625B2129 and 62406222), and in part by the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center.

References

Bian, W.; Wang, Z.; Li, K.; and Bian, J.-W. 2023. NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4160–4169.

Cai, X.; Wang, Y.; Fan, Z.; Deng, H.; Wang, S.; Li, W.; Li, D.; Luo, L.; Wang, M.; and Xu, J. 2024. Dust to Tower: Coarse-to-Fine Photo-Realistic Scene Reconstruction from Sparse Uncalibrated Images. In *arXiv preprint arXiv: 2412.19518*, 1–13.

Cheng, Z.; Esteves, C.; Jampani, V.; Kar, A.; Maji, S.; and Makadia, A. 2023. LU-NeRF: Scene and Pose Estimation by Synchronizing Local Unposed NeRFs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18266–18275.

Dahmani, H.; Bennehar, M.; Piasco, N.; Roldão, L.; and Tsishkou, D. 2024. SWAG: Splatting in the Wild Images with Appearance-Conditioned Gaussians. In *Proceedings of the 18th European Conference on Computer Vision*, 325–340.

Feng, W.; Ye, K.; Zhang, Q.; Zhang, Q.; and Li, N. 2025. 2D Gaussian splatting for outdoor scene decomposition and relighting. In *International Joint Conference on Artificial Intelligence*, 1–9.

Fu, Y.; Wang, X.; Liu, S.; Kulkarni, A.; Kautz, J.; and Efros, A. A. 2024. COLMAP-Free 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20796–20805.

Gao, J.; Gu, C.; Lin, Y.; Li, Z.; Zhu, H.; Cao, X.; Zhang, L.; and Yao, Y. 2024. Relightable 3D Gaussians: Realistic Point Cloud Relighting with BRDF Decomposition and Ray Tracing. In *Proceedings of the 18th European Conference on Computer Vision*, 73–89.

Huang, C.; Zhang, Q.; Zhang, Q.; Li, N.; Gong, Y.; Wang, X.; and Feng, W. 2025. TriGS: Tri-consistency 3D Gaussian Splatting from Sparse and Unposed Views. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 699–708.

Ji, B.; and Yao, A. 2025. SfM-Free 3D Gaussian Splatting via Hierarchical Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21654–21663.

Jiang, Y.; Tu, J.; Liu, Y.; Gao, X.; Long, X.; Wang, W.; and Ma, Y. 2024. GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5322–5332.

Jin, H.; Liu, I.; Xu, P.; Zhang, X.; Han, S.; Bi, S.; Zhou, X.; Xu, Z.; and Su, H. 2023. TensoIR: Tensorial Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 165–174.

Kaleta, J.; Kania, K.; Trzcinski, T.; and Kowalski, M. 2024. LumiGauss: High-Fidelity Outdoor Relighting with 2D Gaussian Splatting. In *arXiv preprint arXiv: 2412.19518*, 1–13.

Kang, G.; Yoo, J.; Park, J.; Nam, S.; Im, H.; Shin, S.; Kim, S.; and Park, E. 2025. SelfSplat: Pose-Free and 3D Prior-Free Generalizable 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22012–22022.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42: 139:1–139:14.

Kim, I.; Choi, M.; and Kim, H. J. 2023. UP-NeRF: Unconstrained Pose-Prior-Free Neural Radiance Fields. In *arXiv preprint arXiv: 2311.03784*, 1–13.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 1–15.

Li, H.; Gao, Y.; Wu, C.; Zhang, D.; Dai, Y.; Zhao, C.; Feng, H.; Ding, E.; Wang, J.; and Han, J. 2024. GGRt: Towards Pose-Free Generalizable 3D Gaussian Splatting in Real-Time. In *Proceedings of the 18th European Conference on Computer Vision*, 325–341.

Liang, Z.; Zhang, Q.; Feng, Y.; Shan, Y.; and Jia, K. 2024. GS-IR: 3D Gaussian Splatting for Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21644–21653.

Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. BARF: Bundle-Adjusting Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5721–5731.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S. M.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7210–7219.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the 16th European Conference on Computer Vision*, 405–421.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 24: 1–32.

Pokutta, S.; Singh, M.; and Torrico, A. 2020. On the Unreasonable Effectiveness of the Greedy Algorithm: Greedy Adapts to Sharpness. In *Proceedings of the 37th International Conference on Machine Learning*, 7772–7782.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Rudnev, V.; Elgharib, M.; Smith, W. A. P.; Liu, L.; Golyanik, V.; and Theobalt, C. 2022. NeRF for Outdoor Scene Relighting. In *Proceedings of the 17th European Conference on Computer Vision*, 615–631.

Sun, H.; Gao, Y.; Xie, J.; Yang, J.; and Wang, B. 2025. SVG-IR: Spatially-Varying Gaussian Splatting for Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16143–16152.

Wang, W.; Gleize, P.; Tang, H.; Chen, X.; Liang, K. J.; and Feiszli, M. 2024. ICON: Incremental CONfidence for Joint Pose and Radiance Field Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5406–5417.

Wang, Y.; Wang, J.; and Qi, Y. 2024. WE-GS: An In-the-wild Efficient 3D Gaussian Representation for Unconstrained Photo Collections. In *arXiv preprint arXiv: 2406.02407*, 1–12.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13: 600–612.

Xu, J.; Mei, Y.; and Patel, V. M. 2024. Wild-GS: Real-Time Novel View Synthesis from Unconstrained Photo Collections. In *arXiv preprint arXiv: 2406.10373*, 1–15.

Zhang, D.; Wang, C.; Wang, W.; Li, P.; Qin, M.; and Wang, H. 2024a. Gaussian in the Wild: 3D Gaussian Splatting for Unconstrained Image Collections. In *Proceedings of the 18th European Conference on Computer Vision*, 341–359.

Zhang, J.; Zhan, F.; Yu, Y.; Liu, K.; Wu, R.; Zhang, X.; Shao, L.; and Lu, S. 2023. Pose-Free Neural Radiance Fields via Implicit Pose Regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3511–3520.

Zhang, Q.; Huang, C.; Zhang, Q.; Li, N.; and Feng, W. 2024b. Learning Geometry Consistent Neural Radiance Fields from Sparse and Unposed Views. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8508–8517.

Zhang, Q.; Huang, C.; Zhang, Q.; Li, N.; and Feng, W. 2025. SU-RGS: Relightable 3D Gaussian Splatting from Sparse Views under Unconstrained Illuminations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 26859–26868.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.