SU-RGS: Relightable 3D Gaussian Splatting from Sparse Views under Unconstrained Illuminations

Qi Zhang, Chi Huang, Qian Zhang, Nan Li, Wei Feng*
College of Intelligence and Computing, Tianjin University, China
{qizhang118, 3020244197, qianz, linan94, wfeng}@tju.edu.cn

Figure 1. With inputting sparse views under unconstrained illuminations, SU-RGS can render novel views with RGB color, albedo, geometry, and materials of the scene, which achieves the state-of-the-art relighting quality with fewer inputs views than baselines (e.g., only 25% of the baselines).

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 densely sampled images under static illumination conditions, which is prohibitively expensive and even impractical in real-world scenarios. In this paper, we propose a novel learning from Sparse views under Unconstrained illuminations Relightable 3D Gaussian Splatting (dubbed SU-RGS), to address this challenge by jointly optimizing 3DGS representations, surface materials, and environment illuminations (i.e., unknown and various lighting conditions in training) using only sparse input views. Firstly, SU-RGS presents a varying appearance rendering strategy, enabling each 3D Gaussian can perform inconsistent color under various lightings. Next, SU-RGS establishes the multi-view semantic consistency by constructing hierarchical semantics pseudo-labels across inter-views, to compensate for extra supervisions and facilitate sparse inverse rendering for confronting unconstrained illuminations. Additionally, we introduce an adaptive transient object perception component that integrates the scene geometry and semantics in a fine-grained manner, to quantify and eliminate the uncertainty of the foreground. Extensive experiments on both synthetic and real-world challenging datasets demonstrate the effectiveness of SU-RGS, achieving the state-of-the-art performance for scene inverse rendering by learning 3DGS from only sparse views under unconstrained illuminations.

1. Introduction

Scene relighting is a long-standing problem in computer vision and graphics, where inverse rendering provides a capable pattern for this task. Impressively, neural radiance fields (NeRF) [1] employ multi-MLP networks to predict 3D physical attributes (e.g., shape and materials) brilliantly from captured images [2]. Nevertheless, the expensive computation of the neural network holds back its training and rendering efficiency.

Recently, 3D Gaussian splatting (3DGS) performs real-time rendering for 3D scene representation [3]. Some researchers attempt to tackle the efficient scene inverse rendering by 3DGS [4]. GS-IR incorporates normal reconstruction and illumination modeling to estimate scene geometry and surface material [5]. R3DG proposes a point-based ray tracing with the bounding volume for physical-based rendering (PBR) [6]. However, the remarkable performance of 3DGS is heavily contingent on two critical requirements: (1) dense input views and (2) static environ-

*Corresponding author.

ment illuminations. These requirements impose significant limitations on the practical applicability of 3DGS inverse rendering in real-world scenarios. In many cases, the availability of input images becomes sparse, often leaving only a few RGB images captured from specific viewpoints [7]. Especially in outdoor scenes, the ambient light will change continuously during the image capturing. Consequently, 3DGS may suffer from overfitting and produce erroneous relighting images [8, 9]. Although we can adopt the traditional image-based relighting methods [10, 11], to transform the various illumination images to the specific lighting separately, the error of relighting and 3DGS modeling can accumulate to result in a terrible rendering. Moreover, the generalization of the mentioned relighting methods relies on the specific training images, which will introduce the domain gap problem inevitably.

The key challenges of relighting 3DGS from sparse views under unconstrained (i.e., unknown and various) illuminations are evident: (1) Each 3D Gaussian needs perform varying appearance representation when meeting various illuminations in the same scene with the same viewpoint. (2) All 3D Gaussians lack sufficient supervision from ground truth (GT) labels when fall into sparse views. One of the directly solutions is to follow [7] and [12] for constructing inter-view photometric pseudo-labels, which are hopeful about mitigating the above problems by plentiful handcraft supervisions. Nevertheless, it may be a failure because inter-view matching pixels could perform different photometric colors due to the unconstrained illuminations, which weakens the effectiveness radically. Based on the strategy, the other concise solution is to integrate physical-based rendering for synthesizing the current view image with inter-view light to produce extra supervisions. However, the error of estimated light parameters and inaccurate Gaussians' representation will accumulate for a bad rendering.

In this paper, we present the Relightable 3D Gaussian Splatting from only Sparse view with Unconstrained illuminations (dubbed SU-RGS) that relighting scene through the inverse rendering. To the best of our knowledge, our method is the first work to introduce the 3DGS from sparse views under unconstrained illuminations for inverse rendering, which can simultaneously estimate scene geometry, materials, and illuminations. We illustrate the framework of the SU-RGS in Fig. 1 and Fig. 2, where the whole pipeline consists of the Varying Appearance Rendering, Sparse Inverse Rendering, and Transient Object Perception. (1) For the Varying Appearance Rendering, we present a various appearance network to integrate the observed direction, the base color of 3D Gaussian, the position of 3D Gaussian, and the image appearance to make each Gaussian perform different colors in the same scene under different views and various illuminations, that generates a passable geometry prior. (2) For the Sparse Inverse Rendering, we establish warping correlations from inter-view images for constructing additional semantics pseudo-labels, which provide sufficient hierarchical semantic gradients of warping pixels to compensate sparse supervisions, that is robustness for the discrepancy of unconstrained illuminations. Till here, we can learn scene geometry, materials and illuminations simultaneously. (3) For the Transient Object Perception, the proposed geometry and semantics cooperative module can quantify and eliminate the transient uncertainties, which considers the relations between geometry and semantics of the mentioned warping pixels adaptively.

We conduct extensive evaluations of SU-RGS and the state-of-the-art baselines, as shown in Fig. 1, where SU-RGS effectively outperforms various methods to achieve the SOTA performance on the NeRF-OSR [13], TensoIR Synthetic [14], and NeRF On-the-go [15] datasets for sparse views under unconstrained illuminations with relightable 3DGS optimization. The main contributions are as follows:

- We advocate the idea of SU-RGS to optimize a relightable 3D Gaussian splatting inverse rendering model only from sparse views under unconstrained illuminations.
- We propose a varying appearance rendering strategy of SU-RGS, for rendering view- and illumination-dependent color for each Gaussian, to optimize 3DGS under various unknown illuminations.
- We elaborate on extra effectiveness supervisions of SU-RGS, the inter-view semantic pseudo-labels, for constructing sufficient supervisions by hierarchical semantic features between re-projected pixels, to realize 3DGS optimization from sparse views.
- We present a novel transient object perception component, which can quantify and eliminates the transient uncertainties of the scene via integrating the geometry and semantics in a fine-grained manner.

2. Related Work

We mainly introduce two aspects of research for 3DGS, i.e., the radiance fields representation and the relightable 3D Gaussian splatting, which are relevant to ours.

Radiance Fields Representation The recent advancements in radiance field have shown immense potential in novel view synthesis (NVS). Neural radiance field (NeRF) is an excellent implicit scene representation strategy for NVS, which employs an MLP network to predict the coordinates' density and color before volume rendering [1]. Specifically, most methods focus on realizing authentic rendering [16–18]. Some researchers attempt to reduce the number of inputting views [19–21]. Moreover, relaxing the precise camera parameters requirements is received wide attention [22–26]. However, the NeRF-based methods suffer from the efficient training and rendering because of the heavy computation of MLP network. Thus, 3D Gaussians splatting comes into scene rendering [3, 27]. Mip-

Splatting [28], VastGaussian [29], and SuGaR [30] achieve satisfactory results of scene representation. [31] and [32] also try to input sparse views. And some works slim Gaussians to decrease the expenditure of model storage [33, 34].

Although the mentioned radiance fields perform impressive novel view rendering, they all represent the scene as a non-Lambertian model, which neglect the surface material and environment light. Thus they suffer from relighting the authentic scene with the accurate and flexible lighting.

Relightable 3D Gaussian Splatting Recently, increasing researches have focused on the relightable radiance fields [35, 36]. 3DGS becomes a key foundation method because of the efficient rendering and scene representation [37, 38]. GS-W [39] and WildGaussians [40] optimize 3DGS in-the-wild data, which is characterized by occlusions, transient objects, and varying illumination challenges. Nevertheless, they neglect to model the scene's materials explicitly (i.e., albedo), which is difficult for the scene inverse rendering with parameterized relighting. More recently, GSIR estimates scene geometry, surface material, and environment illumination from multi-view images under unknown lighting conditions [5]. R3DG introduces a point-based ray tracing strategy for 3DGS optimizing to decompose bidirectional reflectance distribution function (BRDF) and ambient lighting by physically based differentiable rendering [6]. Although these methods achieve impressive results in scene inverse rendering, they suffer from inputting a constant environment lighting, which is rarely seen in real-world scenarios. Notably, LumiGauss [4] realizes inverse rendering with relightable 3DGS under various illuminations. However, it is specifically designed for dense view settings, which remains fundamentally constrained like R3DG [6] with sparse inputting views.

Different from the above researches, we present a novel relightable 3D Gaussian Splatting strategy, which achieves remarkable scene inverse rendering performances only with sparse views under unconstrained illuminations.

3. Preliminary

In this section, we give backgrounds and notations that are necessary for the presentation of our proposed method.

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

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

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 $\mathbf{\Sigma} = \mathbf{RSS}^{\mathsf{T}}\mathbf{R}^{\mathsf{T}}$ 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 $\mathbf{\Sigma}'$ can be defined as follows:

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

where $\mathbf{W} \in \mathbb{R}^{3\times3}$ is the rotation matrix. $\mathbf{J}$ is the Jacobian of the affine approximation of the projection. Moreover, each Gaussian has the color $\mathbf{c}$ by a spherical harmonic (SH) coefficient. Finally, the pixel color $\hat{C}$ is rendered as follows:

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

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.

3DGS with Inverse Rendering For SU-RGS, we follow [5] to employ the physical-based rendering 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.

4. Method

In this section, we first elucidate the motivation and method overview in Section 4.1. Then we report the framework of SU-RGS in Section 4.2 - 4.5.

4.1. Overview

This work aims to enjoy scene relighting by 3DGS from sparse views under unconstrained illuminations. Specifically, we address the challenging task of optimizing the 3DGS inverse rendering model for relighting.

Given a set of RGB images $\{I_i\}_{i=1}^N$ of a scene captured from sparse views (e.g., as low as 6 views), yet unknown and various illuminations. Motivated by the fascinating performance of 3DGS [3], we present a novel relighting framework SU-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, SU-RGS consists of three well-designed modules. First, we propose a varying appearance rendering module, to characterize view- and illumination-dependent color for

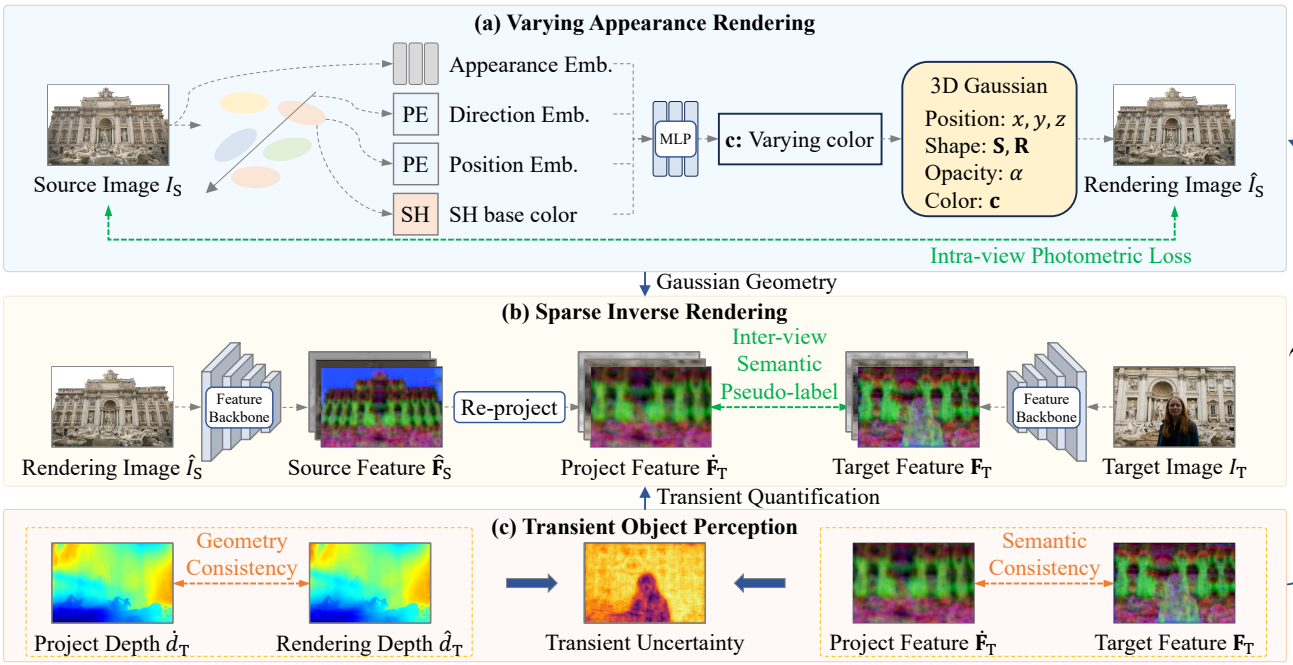

Figure 2. Overview of the framework for SU-RGS. We first present the varying appearance rendering for varying color of each Gaussian under unconstrained illuminations. Then SU-RGS constructs inter-view hierarchical semantics pseudo-labels for sparse inverse rendering. Finally, we employ the inter-view geometry and semantics to metric the transient uncertainty. PE is the positional encoding in NeRF [1].

each Gaussian under unconstrained illuminations, which will provide a general satisfaction scene geometry (i.e., all Gaussians' position) (*cf.* Sec. 4.2). Senondly, we leverage the physical-based rendering formulation, and construct inter-view semantic feature pseudo-labels to compensate scarce supervisions *w.r.t* sparse inputting views, which is robustness for the discrepancy of environment lighting. The scene materials and illuminations can be optimized simultaneously with the 3DGS model (*cf.* Sec. 4.3). Moreover, we also integrate the image semantics and depth maps to discriminate and eliminate the transient objects by a novel geometry and semantics cooperative module (*cf.* Sec. 4.4).

4.2. Varying Appearance Rendering

For the scene relighting via 3DGS, the previous methods adopt inverse rendering through a two-stage strategy [5, 6]. They first optimize 3DGS by the standard volume rendering formulation under a static illumination. Then replacing the rendering equation as the physical-based rendering mechanism to learn the scene materials and illuminations, where freezes Gaussians' geometry from the first stage. However, especially for the first stage, this pipeline fails to generate precise Gaussian geometry under unconstrained illuminations. Because each Gaussian performs varying appearance under various illuminations, only one spherical harmonics (SH) coefficient of a Gaussian will receive the different (even contradictory) color gradients from more than one photometric labels, where are from the same or similar views. It can be misled to converge at local minima, thus yielding terrible geometry and relighting rendering.

Towards this end, we propose a various appearance network module for SU-RGS, to realize the view- and illumination-dependent varying appearances rendering for each Gaussian. In SU-RGS, we adapt an MLP $f_{\mathbf{\Theta}}$ that takes the input as a concatenation of: (a) the observed direction $\mathbf{r} \in \mathbb{R}^3$, (b) the base color of 3D Gaussian $\mathbf{c}' \in \mathbb{R}^3$, (c) the position of 3D Gaussian $\mathbf{x} \in \mathbb{R}^3$, and (d) the image appearance $\mathbf{l} \in \mathbb{R}^M$. Notably, we obtain the base color of 3D Gaussian $\mathbf{c}'$ via the 0-th order SH coefficients, which is the least disturbed color representation under various illuminations. Specifically, the observed direction $\mathbf{r}$ and the position of 3D Gaussian $\mathbf{x}$ will be mapped from $\mathbb{R}^3$ into a higher dimensional space by a positional encoding function.

Accordingly, the view- and the illumination-dependent varying appearance $\dot{\mathbf{c}}$ for each 3D Gaussian is computed as:

$$\dot{\mathbf{c}} = f_{\mathbf{\Theta}}\left(\gamma(\mathbf{r}), \mathbf{c}', \gamma(\mathbf{x}), \mathbf{l}\right), \tag{5}$$

where $\gamma(\cdot)$ is the positional encoding function followed by NeRF [1]. $\mathbf{l}$ is modeled by the generative latent optimization (GLO) [41] in which each image I is assigned a corresponding vector. Till here, all Gaussians have capable to perform varying appearance rendering, which facilitates 3DGS to generate satisfactory geometry (i.e., Gaussians' position) under unconstrained illuminations with the standard volume rendering formulation.

4.3. Sparse Inverse Rendering

The varying appearance rendering strategy favors the 3DGS learning scene geometry passably via the volume rendering. According to this prior, we would relight the scene by integrating the physical-based rendering (PBR) formulation. However, that is insufficient (i.e., overfitting) to fulfill it only with intra-view PBR photometric losses from sparse views (e.g., 24 or fewer views). Specifically, following [7] to construct inter-view PBR pseudo-labels, and rendering with the current camera poses by the inter-view light, it couples the errors of estimated ambient light and Gaussians representation (e.g., geometry and BRDF), which actually further increases the difficulty of optimization. Moreover, deep feature focuses on semantics rather than colors [25]. Thus we construct inter-view hierarchical semantic pseudo-labels, to compensate for sparse supervisions.

The key idea is to narrow the gap of the semantic feature differences between pixels with re-projection relationships. Given a pair of images $\hat{I}_{\mathrm{S}}$ and I_{T} from the source and target view, respectively. We re-project the pixel $\mathbf{p}_{\mathrm{S}}$ from view S to pixel $\dot{\mathbf{p}}_{\mathrm{T}}$ of view T as the follows:

$$\begin{aligned} \bar{P}_{\mathrm{S}} &= \overline{[\mathbf{R}\ \mathbf{t}]}_{\mathrm{S}}^{-1}[\hat{d}_{\mathrm{S}}\mathbf{K}_{\mathrm{S}}^{-1}\bar{\mathbf{p}}_{\mathrm{S}}], \\ \bar{\dot{P}}_{\mathrm{T}} &= \overline{[\mathbf{R}\ \mathbf{t}]}_{\mathrm{T}}\bar{P}_{\mathrm{S}}, \\ \bar{\dot{\mathbf{p}}}_{\mathrm{T}} &= \mathrm{NORM}(\mathbf{K}_{\mathrm{T}}\dot{P}_{\mathrm{T}}), \end{aligned} \tag{6}$$

where $\dot{\mathbf{p}}_{\mathrm{T}}$ depicts the re-projected pixel from S to T. $\hat{d}$ is the depth of pixel $\mathbf{p}$ from 3DGS rendering. $\bar{(\cdot)}$ is the homogeneous representation. $[\mathbf{R}\ \mathbf{t}]$ and $\mathbf{K}$ denotes the camera external and internal parameters. $\mathrm{NORM}(\cdot)$ is the depth normalization. Subsequently, we follow [42] to extract semantic features by 2D foundation model. According to the mapping from the pixel to feature map, we adopt the total squared error between the re-projected pixels as follows:

$$\mathcal{L}_f = \sum_{k\in\mathcal{A}}\sum_{(i,j)\in\mathcal{M}} \|\hat{\mathbf{F}}_i^k(\mathbf{p}_{\mathrm{S}}) - \mathbf{F}_j^k(\dot{\mathbf{p}}_{\mathrm{T}})\|_2^2, \tag{7}$$

where $\mathcal{A}$ denotes the set of hierarchical semantic features' indexes. $\hat{\mathbf{F}}$ denotes hierarchical semantics are from rendering images. $\mathcal{M}$ is the set of re-projected pixels. i and j are the indexes of re-projected pixels in view S and T, respectively. The pixels that project out of bounds will be filtered out. Till here, the inter-view hierarchical semantic feature pseudo-labels will provide extra supervisions beyond the standard sparse photometric GT labels. Then SU-RGS can propagate the gradients to optimize the model parameters of environment light, BRDF, opacity, covariance matrix, and position of all Gaussians. Notably, we freeze the model parameters of 2D foundation model during 3DGS training.

4.4. Transient Object Perception

For a scene under the static illumination, we can realize the inverse rendering and relight the scene through the varying appearances rendering and sparse inverse rendering components. However, in real-world scenes, there will be some uncertain transient objects (e.g., visitors), which inevitably result in some problematic labels for the mentioned modules, causing 3DGS model to converge to a local optimum. Towards this end, we propose a fine-grained transient object perception strategy, which integrates the scene geometry and hierarchical semantics building upon the by-product of sparse inverse rendering component.

The core of the inspiration is to discriminate the relativity of the semantics for the re-projected pixels. On the basis of the inter-view hierarchical semantic pseudo-labels, we can yield the matching pixels (e.g., $\mathbf{p}_{\mathrm{S}}$ and $\dot{\mathbf{p}}_{\mathrm{T}}$) and the mapping from the pixels to feature map. Then we get the re-projected depth $\dot{d}_{\mathrm{T}}$ from pixel $\mathbf{p}_{\mathrm{S}}$ to $\dot{\mathbf{p}}_{\mathrm{T}}$, and the rendering depth $\hat{d}_{\mathrm{T}}$ of pixel $\dot{\mathbf{p}}_{\mathrm{T}}$. Notably, if there is a transient object in view S or T, there will be a significant distance between the depth $\dot{d}_{\mathrm{T}}$ and $\hat{d}_{\mathrm{T}}$. For example, if a transient object only appears in view T, the re-projected depth $\dot{d}_{\mathrm{T}}$ is smaller than the rendering depth $\hat{d}_{\mathrm{T}}$ obviously, because of the obstacle is closer to camera. Based on this phenomenon, we quantify the uncertainty of the transient object using a piecewise function, that integrates scene geometric and semantics to calculate an adaptive transience coefficient determining the contribution of labels' gradients. The details are as follows:

$$\beta = \begin{cases} \xi & , 0 \le E < F \\ \mathrm{EXP}(-E)\cdot\xi, & F \le E \le 1 \end{cases}, \tag{8}$$

where $E = |\dot{d}_{\mathrm{T}} - \hat{d}_{\mathrm{T}}|/\mathrm{MAX}(\dot{d}_{\mathrm{T}}, \hat{d}_{\mathrm{T}})$ is the error of the re-projected and renderer depth. $\mathrm{MAX}(\cdot)$ denotes the maximum value calculation. $F = \mathrm{MIN}(\dot{d}_{\mathrm{T}}/\hat{d}_{\mathrm{T}}, \hat{d}_{\mathrm{T}}/\dot{d}_{\mathrm{T}}) + \epsilon$ represents the adaptive boundary, where $\epsilon \in (-1, 1)$ is a hyper-parameter (e.g., -0.1 in experiments). $\mathrm{EXP}(\cdot)$ is the exp-function. And ξ denotes the similarity coefficient of the hierarchical semantic features, which is calculated as follows:

$$\xi = \frac{1}{N}\sum_{l=1}^{N} \mathrm{SIM}\left(\mathbf{F}_l(\mathbf{p}_{\mathrm{S}}), \mathbf{F}_l(\dot{\mathbf{p}}_{\mathrm{T}})\right), \tag{9}$$

where $\mathrm{SIM}(\cdot)$ is the cosine similarity calculation function. $\mathbf{F}_l$ denotes the l-th semantic feature of the GT image, which can be extracted in Sec. 4.3 by the semantics pseudo-labels.

After that, SU-RGS employs the adaptive transience coefficient to quantify or eliminate the transient object adaptively through controlling the contribution of the gradients from intra-view labels and inter-view pseudo labels. Notably, β will be assigned to both view S and T, that the corelation pixels with transient uncertainties will be attenuated or even rejected from participating in backpropagation.

4.5. Network Training

To narrow the gap of the rendering and GT color, we adopt the standard photometric loss [1], the semantics to-

tal squared error loss (i.e., Eq. (7)), and a D-SSIM term [3] as the training objective function. The details are as follows:

$$\mathcal{L} = \beta(\lambda_1 \sum_{\mathbf{p}\in\mathcal{R}} \|\hat{C}(\mathbf{p}) - C(\mathbf{p})\|_2^2 + \lambda_2 \mathcal{L}_{\text{D-SSIM}} + \lambda_3 \mathcal{L}_f), \tag{10}$$

where $\mathcal{R}$ is the set of pixels in the images. β is the adaptive transient coefficient. Notably, for the varying appearance rendering (*cf.* Sec. 4.2), $\hat{C}(\mathbf{p})$ is the various colors of the appearance, which is rendered by color $\dot{\mathbf{c}}_i$ with Eq. (5). Specifically, with respect to the sparse inverse rendering (*cf.* Sec. 4.3), we employ PBR formulation to replace the classical volume rendering in Eq. (3) for learning scene's geometry, material, and illuminations by Eq. (4).

For SU-RGS, the training is split into two stages with end-to-end. Firstly, we only use the standard volume rendering equation (i.e., Eq. (3)) with the varying appearances rendering and transient object perception (i.e., Eq. (5) and Eq. (8)) to synthesize scene color. Secondly, we employ the physical-based rendering formulation (PBR) (i.e., Eq. (4)) with the transient object perception module (i.e., Eq. (8)) to refine the Gaussians' geometry and learn the material and environment illuminations of the scene. The loss function is set to be the same for two stages. And we calculate both the re-projection from view S to T and view T to S for the sparse inverse rendering and the transient object perception.

5. Experiments

5.1. Experimental Setup

Datasets & Metrics We evaluate all baselines and our presented SU-RGS on the ***NeRF-OSR*** [13] and ***TensoIR Synthetic*** [14], which provide various illuminations scenes. We also verify the transient object perception on ***NeRF On-the-go*** [15] real-world dataset. It comprises a diverse array of casually captured indoor and outdoor sequences, with transit ratios varying between 5% and 30%. Notably, this dataset demonstrates negligible illumination changes across different viewpoints. For evaluation purposes, we follow the authors to set test set that mentioned in the published papers. For the novel view synthesis, we follow [13] to select Peak Signal-to-Noise Ratio (PSNR) [7], Structural Similarity Index Measure (SSIM) [43], and Learned Perceptual Image Patch Similarity (LPIPS) [44] metrics. To verify the efficacy of normal reconstruction, we follow [5] to use Mean Angular Error (MAE) on TensoIR Synthetic dataset.

Baselines We compare to four SOTA relightable radiance field baselines. ***TensoIR*** [14] is an inverse rendering approach based on tensor factorization and neural fields, which jointly achieves radiance field reconstruction and physically-based model estimation. ***R3DG*** [6] is a differentiable point-based rendering framework to achieve photo-realistic scene relighting. ***GSIR*** [5] leverages forward mapping volume rendering to achieve photorealistic novel

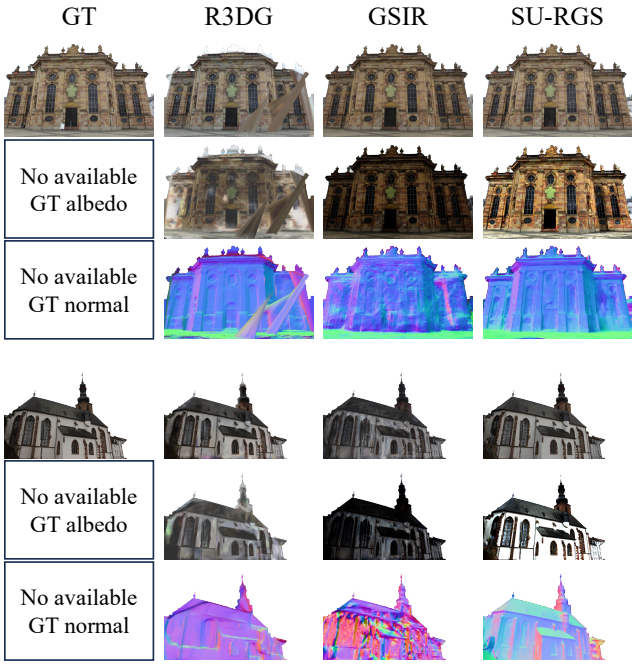

Figure 3. Visualization of relighting novel view, decomposition, and geometry reconstruction in NeRF-OSR dataset.

view synthesis and relighting results. ***NeRF-OSR*** [13] is a relightable NeRF under various illuminations in outdoor scenes, which suffers from dense input views. Notably, we follow NeRF-OSR to modify the first three methods so that each view corresponds to a lighting parameter rather than each scene, for adapting to various illuminations. Specifically, we also compare with ***WildGaussians*** [40] on NeRF On-the-go dataset, which is a representative method for the transient object perception whereas the previous strong baselines without this ability.

Implementation Details We implement our framework based on previous work [3, 5] in Python using the PyTorch and wrote custom CUDA kernels for rasterization. For SU-RGS, we use Adam optimizer [45] for training, and the training process includes the initial stage and decomposition stage. The first stage trains with the standard volume rendering equation for 30K iterations. In the second stage, we employ PBR formulation for 5K iterations. As for all baselines, we defer to the paper and source codes to conduct experiments with the same settings as SU-RGS, which are with 24 input views under unconstrained illuminations and 35K iterations with about 1.5 hour on a single RTX 3090. Especially, we follow NeRF-OSR [13] to modify TensoIR, R3DG, and GSIR for the various illuminations inputting.

5.2. Comparing to SOTA in Real-world Scenes

We evaluate SU-RGS and the SOTA inverse rendering methods by radiance filed with 24, 12, and 6 input views in different metrics on NeRF-OSR dataset.

As illustrated in Table 1, the quantitative results of SU-

Table 1. The quantitative results of all methods for relighting novel view in NeRF-OSR dataset with 24, 12, and 6 inputting views, respectively.

Method	24 Views			12 Views			6 Views		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
TensoIR	16.09	0.514	0.583	13.33	0.466	0.580	11.22	0.398	0.686
R3DG	17.08	0.661	0.396	15.19	0.519	0.461	13.63	0.423	0.613
GSIR	18.59	0.726	0.352	15.45	0.540	0.412	14.54	0.491	0.574
NeRF-OSR	16.16	0.521	0.576	13.45	0.471	0.591	11.14	0.393	0.692
SU-RGS	**19.95**	**0.876**	**0.244**	**18.35**	**0.719**	**0.312**	**16.04**	**0.664**	**0.438**

Table 2. The quantitative results of all methods for relighting novel view and normal reconstruction in TensoIR synthetic dataset with 24, 12, and 6 inputting views, respectively.

Method	24 Views				12 Views				6 Views			
	PSNR ↑	SSIM ↑	LPIPS ↓	MAE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MAE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MAE ↓
TensoIR	28.52	0.863	0.139	8.89	27.07	0.850	0.170	15.26	26.99	0.828	0.257	14.55
R3DG	28.66	0.882	0.136	8.95	27.18	0.870	0.142	14.99	27.05	0.827	0.240	14.12
GSIR	29.29	0.896	0.125	8.15	27.41	0.902	0.134	12.11	27.18	0.835	0.216	13.99
NeRF-OSR	28.32	0.861	0.143	8.31	27.09	0.835	0.172	13.84	26.93	0.824	0.269	14.64
SU-RGS	**33.05**	**0.976**	**0.033**	**5.17**	**29.19**	**0.946**	**0.067**	**7.22**	**28.05**	**0.930**	**0.079**	**12.32**

RGS outperform all baselines in three metrics on all groups by a large margin. Specifically, compared with the strong baseline GSIR, SU-RGS has increased by 1.36, 0.150, and 0.108 with respect to PSNR, SSIM, and LPIPS, respectively. This phenomenon demonstrates that the presented varying appearance modeling and hierarchical semantic pseudo-labels can augment the ability of 3DGS from sparse and unconstrained illuminations. Especially, the transient uncertainty perception strategy further facilitates the outdoor scene optimization. For the qualitative analysis of the scene relighting, decomposition, and geometry reconstruction, SU-RGS outperforms the SOTA relightable inverse rendering 3DGS models significantly under sparse input views, which are shown in Fig. 3. This further illustrates the superiority of the varying appearance and semantics pseudo-labels framework to augment 3DGS for real-world scene inverse rendering. Notably, we omit the quantitative analysis of scene decomposition and geometry reconstruction on NeRF-OSR datasets, because this benchmark is without GT albedo, metallic, roughness, and normal labels.

5.3. Comparing to SOTA in Synthetic Scenes

We also evaluate SU-RGS and baselines as the same settings as Sec. 5.2 on the TensoIR Synthetic dataset.

As illustrated in Table 2, the quantitative results perform a similar trend to Table 1. Specifically, compared with GS-IR, SU-RGS has improved PSNR, SSIM, LPIPS, and MAE with 3.76, 0.080, 0.092, and 2.98, respectively. This phenomenon further verifies the effectiveness of the presented varying appearance modeling and hierarchical semantic pseudo-labels for 3DGS learning from sparse and unconstrained illuminations. For the qualitative analysis of the scene relighting, decomposition, and geometry reconstruction, as shown in Fig. 4, SU-RGS outperforms the SOTA methods consistently. This illustrates the robustness of SU-RGS to perform satisfactory inverse rendering in both indoor and outdoor scenarios.

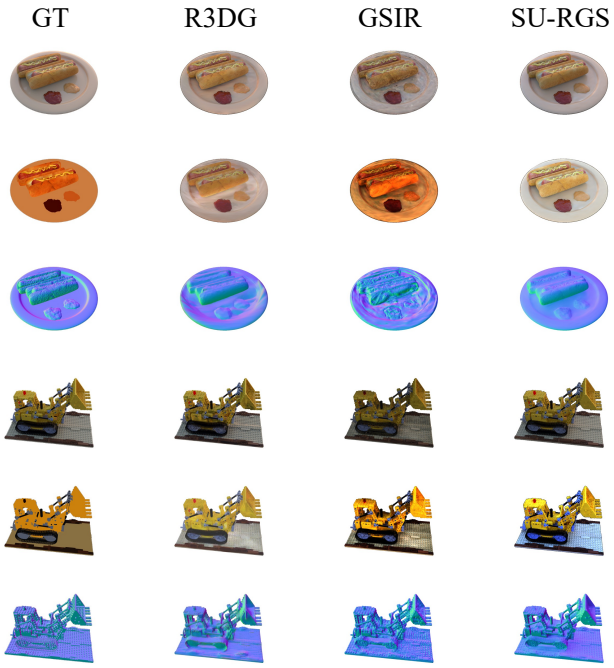

Figure 4. Visualization of relighting, decomposition, and geometry rendering for novel view in TensoIR Synthetic.

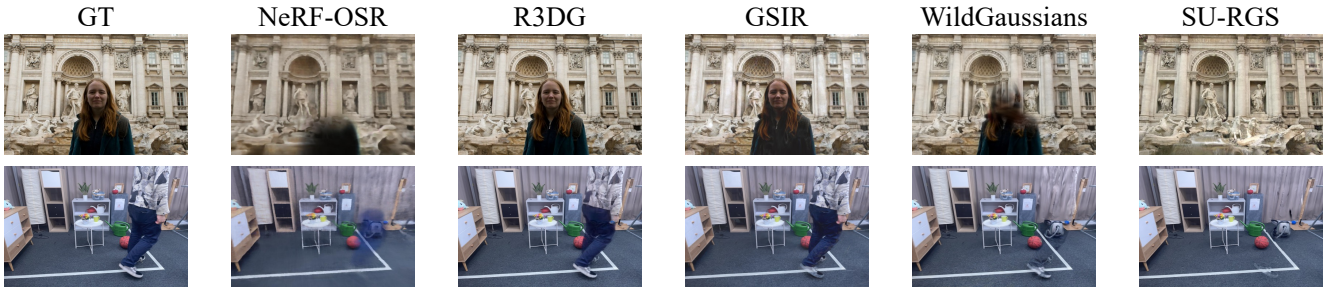

Figure 5. Visualization of transient object perception results on NeRF-OSR and NeRF On-the-go datasets.

Table 3. The effectiveness ablation analysis of the varying appearances rendering.

Variants	AE	BC	PE	RD	PSNR ↑	SSIM ↑	LPIPS ↓
1					17.01	0.521	0.550
2	✓				18.20	0.616	0.486
3	✓	✓			19.09	0.723	0.357
4	✓	✓	✓		19.51	0.793	0.294
5	✓	✓	✓	✓	**19.95**	**0.876**	**0.244**

Table 4. The effectiveness ablation analysis of the sparse inverse rendering.

Variants	S→T	S↔T	w/ H	PSNR ↑	SSIM ↑	LPIPS ↓
1				18.19	0.694	0.370
2	✓			19.11	0.734	0.336
3	✓	✓		19.66	0.827	0.282
4	✓	✓	✓	**19.95**	**0.876**	**0.244**

Table 5. The effectiveness ablation analysis of the transient object perception.

Variants	PSNR ↑	SSIM ↑	LPIPS ↓
w/ Constant	19.50	0.728	0.404
w/ Cosine	19.62	0.793	0.312
w/ Truncation	19.55	0.756	0.340
SU-RGS	**19.95**	**0.876**	**0.244**

5.4. Method analysis

In this section, we conduct a comprehensive analysis of the presented components that are essential to our approach. All methods are evaluated on the NeRF-OSR dataset.

Effect of Varying Appearances Rendering We ablate the vital components of varying appearances rendering (VAR) for SU-RGS in Table 3 to evaluate the effectiveness. AE, BC, PE, and RD denote the appearance embedding, base color, position embedding, and ray direction, respectively. We observe that the results increase consistently from rows 1 to 5. Especially, the last row outperforms the previous variants, which illustrates the presented varying appearances MLP component can model the scene under various illuminations with a passable geometry reconstruction.

Effect of Sparse Inverse Rendering We also evaluate the proposed sparse inverse rendering (SIR) in SU-RGS by ablating each component (see the results in Table 4). Comparing with the first row, the others perform better, because the inter-view semantic pseudo-labels augment the 3DGS optimization from sparse views and unconstrained illuminations. Notably, the third row outperforms the second verifies the dual-direction re-projection can facilitate the effectiveness of pseudo-labels. And the last row (i.e., with hierarchical) achieves the best results, which depicts the hierarchical semantics is necessary to boost the sparse inverse rendering.

Effect of Transient Object Perception We evaluate the effectiveness of transient object perception (TOP) in Table 5. We first set the boundary with a constant value for Eq. (8) (i.e., w/ Constant). Then we replace all transience coefficients with the cosine similarity. Besides, we set the transience coefficient as the truncation (i.e., 0 or 1). It is obvious that SU-RGS outperforms all variants significantly. The phenomenon illustrates the presented adaptive transience perception quantifies the transient object by combining the scene geometry and semantics. Specifically, as illustrated in Fig. 5, the qualitative results also demonstrate the effectiveness SU-RGS. Notably, it only adapts the SIR's by-products without the extra expenditure of rendering, which depicts the superiority of the transient object perception.

6. Conclusion

Current relightable 3DGS methods for scene inverse rendering struggle with dense inputting views and static illuminations. This work presents a relightable 3D Gaussian splatting from only sparse views under unconstrained illuminations (dubbed SU-RGS). We first construct varying appearance rendering module to realize each Gaussian performing varying colors, which provides the satisfactory geometry prior for initializing inverse rendering. Specifically, SU-RGS proposes sparse inverse rendering module to establish inter-view hierarchical semantic pseudo-labels for compensating extra supervisions. Moreover, the presented SU-RGS also elaborates the transient object perception to quantify the scene's uncertainty. We evaluate SU-RGS and state-of-the-art methods in three challenging scenarios. The experiments verify the effectiveness and robustness of SU-RGS. We also plan to extend our strategy to other challenging applications of 3DGS (e.g., relighting with unposed views).

Acknowledgements This work was supported in part by the National Key R&D Program of China under Grants 2023YFF0906200, in part by Natural Science Foundation of China under Grants 62406222, in part by Tianjin University under the Emerging Direction Cultivation Project of Interdisciplinary Center (Intelligent Protection and Utilization of Digital Cultural Heritage).

References

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the 16th European Conference on Computer Vision*, pages 405–421, 2020. 1, 2, 4, 5

[2] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18622–18631, 2022. 1

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42:139:1–139:14, 2023. 1, 2, 3, 6

[4] Joanna Kaleta, Kacper Kania, Tomasz Trzcinski, and Marek Kowalski. Lumigauss: High-fidelity outdoor relighting with 2d gaussian splatting. In *arXiv preprint arXiv: 2408.04474*, pages 1–12, 2024. 1, 3

[5] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 1, 3, 4, 6

[6] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *Proceedings of the 18th European Conference on Computer Vision*, pages 73–89, 2024. 1, 3, 4, 6

[7] Qi Zhang, Chi Huang, Qian Zhang, Nan Li, and Wei Feng. Learning geometry consistent neural radiance fields from sparse and unposed views. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8508–8517, 2024. 2, 5, 6

[8] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *Proceedings of the 18th European Conference on Computer Vision*, pages 145–163, 2024. 2

[9] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. Cor-gs: Sparse-view 3d gaussian splatting via co-regularization. In *Proceedings of the 18th European Conference on Computer Vision*, pages 335–352, 2024. 2

[10] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM Transactions on Graphics*, 384:78:1–78:14, 2019. 2

[11] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William A. P. Smith. Self-supervised outdoor scene relighting. In *Proceedings of the 16th European Conference on Computer Vision*, pages 84–101, 2020. 2

[12] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 2

[13] Viktor Rudnev, Mohamed Elgharib, William A. P. Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *Proceedings of the 17th European Conference on Computer Vision*, pages 615–631, 2022. 2, 6

[14] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 2, 6

[15] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8931–8940, 2024. 2, 6

[16] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19726, 2023. 2

[17] Huiqiang Sun, Xingyi Li, Liao Shen, Xinyi Ye, Ke Xian, and Zhiguo Cao. Dyblurf: Dynamic neural radiance fields from blurry monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7527, 2024.

[18] Jiahao Chen, Yipeng Qin, Lingjie Liu, Jiangbo Lu, and Guanbin Li. Nerf-hugs: Improved neural radiance fields in non-static scenes using heuristics-guided segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19436–19446, 2024. 2

[19] Hanxin Zhu, Tianyu He, Xin Li, Bingchen Li, and Zhibo Chen. Is vanilla mlp in neural radiance field enough for few-shot view synthesis? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20288–20298, 2024. 2

[20] Yingji Zhong, Lanqing Hong, Zhenguo Li, and Dan Xu. Cvt-xrf: Contrastive in-voxel transformer for 3d consistent radiance fields from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21475, 2024.

[21] Minseop Kwak, Jiuhn Song, and Seungryong Kim. Geconerf: Few-shot neural radiance fields via geometric consistency. In *Proceedings of the International Conference on Machine Learning*, pages 18023–18036, 2023. 2

[22] Weiyao Wang, Pierre Gleize, Hao Tang, Xingyu Chen, Kevin J. Liang, and Matt Feiszli. Icon: Incremental confidence for joint pose and radiance field optimization. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5406–5417, 2024. 2

[23] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5721–5731, 2021.

[24] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, and Ameesh Makadia. Lu-nerf: Scene and pose estimation by synchronizing local unposed nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18266–18275, 2023.

[25] Injae Kim, Minhyuk Choi, and Hyunwoo J. Kim. Up-nerf: Unconstrained pose-prior-free neural radiance fields. In *arXiv preprint arXiv: 2311.03784*, pages 1–13, 2023. 5

[26] Jiahui Zhang, Fangneng Zhan, Yingchen Yu, Kunhao Liu, Rongliang Wu, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Pose-free neural radiance fields via implicit pose regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3511–3520, 2023. 2

[27] Wei Feng, Chi Huang, Qi Zhang, Qian Zhang, and Nan Li. Trigs: Tri-consistency 3d gaussian splatting from sparse and unposed views. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1–10, 2025. 2

[28] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 3

[29] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024. 3

[30] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 3

[31] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024. 3

[32] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 3

[33] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21719–21728, 2024. 3

[34] Yihang Chen, Qianyi Wu, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Hac: Hash-grid assisted context for 3d gaussian splatting compression. In *Proceedings of the 18th European Conference on Computer Vision*, pages 422–438, 2024. 3

[35] Wei Feng, Kangrui Ye, Qi Zhang, Qian Zhang, and Nan Li. 2d gaussian splatting for outdoor scene decomposition and relighting. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1–9, 2025. 3

[36] Jia-Mu Sun, Tong Wu, Yong-Liang Yang, Yu-Kun Lai, and Lin Gao. Sol-nerf: Sunlight modeling for outdoor scene decomposition and relighting. In *Proceedings of the ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia*, pages 31:1–31:11, 2023. 3

[37] Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. In *arXiv preprint arXiv: 2406.10373*, pages 1–15, 2024. 3

[38] Tianyuan Zhang, Zhengfei Kuang, Haian Jin, Zexiang Xu, Sai Bi, Hao Tan, He Zhang, Yiwei Hu, Milos Hasan, William T. Freeman, Kai Zhang, and Fujun Luan. Relitlrm: Generative relightable radiance for large reconstruction models. In *arXiv preprint arXiv: 2410.06231*, pages 1–15, 2024. 3

[39] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *Proceedings of the 18th European Conference on Computer Vision*, pages 341–359, 2024. 3

[40] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. In *arXiv preprint arXiv: 2407.08447*, pages 1–15, 2024. 3, 6

[41] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 599–608, 2018. 4

[42] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 5

[43] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 6

[44] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[45] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15, 2015. 6