L^2DGS: Low-Light Dynamic Gaussian Splatting

Ashish Kumar A. N. Rajagopalan
Indian Institute of Technology Madras, India
ashish.k.research@gmail.com, raju@ee.iitm.ac.in

Abstract

Synthesizing novel spatiotemporal views of dynamic scenes is challenging due to object and camera motion, and sparse observations. While recent Neural Radiance Field (NeRF) and Gaussian Splatting (GS) methods enable 4D dynamic scene reconstruction, they predominantly assume well-lit inputs. Existing low-light reconstruction approaches are limited to static scenes and mainly focus on brightness enhancement while overlooking underlying scene structure. Reconstructing well-lit dynamic scenes from low-light inputs is particularly challenging due to motion-induced shadows, occlusions, and disocclusions, making the problem highly ambiguous. We propose L^2DGS (Low-Light Dynamic Gaussian Splatting), a self-supervised 4D GS framework that directly reconstructs well-lit dynamic scenes from low-light videos. The method decomposes the scene into view- and time-dependent illumination and view-time-invariant reflectance components. We introduce an Occlusion-Disocclusion Network (OCD-Net) to model temporal intensity variations and Brightness Attenuation Features (BAFs) with a BAF Enhancement Network (BAFE-Net) to enable geometry- and photometry-aware transformation between well-lit and low-light observations for self-supervision. L^2DGS operates on standard sRGB inputs without requiring camera metadata. Experiments on simulated and proposed real Low-Light Dynamic Video (L^2DyV) datasets demonstrate superior qualitative and quantitative performance over prior methods. The dataset is available at: https://github.com/akumar005/L2DGS.

1. Introduction

Recent advances in NeRFs and 3D Gaussian Splatting (3DGS) have greatly improved novel-view synthesis from well-lit images and videos. Most of these methods assume favorable illumination, whereas real-world scenes often exhibit uncontrolled and challenging lighting. Low-light photography often suffers from reduced visibility, diminished details, color loss, and higher noise levels. Some recent studies [3, 4, 6, 19, 31, 34, 35, 45, 47, 52] have attempted reconstruction of well-lit scenes and novel view synthesis from low-light input, but these are limited to static scenes and cannot model temporal scene dynamics.

The problem of reconstructing dynamic scenes (i.e., a scene having a moving object) and synthesizing spatio-temporal novel well-lit views from low-light inputs remains largely underexplored. The problem becomes quite complex in the presence of a moving object due to several factors: (i) Adjusting camera settings such as exposure can introduce unwanted artifacts, including motion blur. (ii) Object motion representation and learning is quite challenging. (iii) Inherent sparsity (as one particular time instance is observed only once in the case of a single camera). (iv) Motion and view-dependent brightness/darkness in the scene. (v) Object motion introduces textural ambiguity in the background. As the object moves, the position of its shadow on the background (cast shadow) may shift, making it difficult to determine whether a given region is inherently dark or appears dark due to shadowing. (vi) Additionally, object motion and viewpoint variations can cause shadows onto the object itself (self-shadow). Moreover, the correlation between object motion and shadow dynamics depends on the specific 3D scene and viewpoint.

Methods such as [2, 12, 13, 15, 22, 24, 25, 37, 40, 41, 49, 51, 53] address low-light enhancement but cannot synthesize spatiotemporal novel views. Performing the dual task of low-light enhancement and dynamic novel view synthesis as a single, end-to-end task can ensure efficiency, consistency, and robustness. Instead of treating them as two separate processes, we posit that an integrated pipeline offers several advantages. (i) A joint model can learn a unified geometric and photometric feature space, ensuring that enhancements are performed in a manner that naturally extends to view-synthesis. (ii) Enhancing low-light dynamic scenes and synthesizing new views in space and time requires understanding scene illumination and reflectance, scene geometry and depth, and object motion, which a single model can jointly learn for consistent lighting and view generation. (iii) An end-to-end model can learn how to jointly enhance and synthesize views for moving objects, ensuring motion consistency across frames.

Dynamic scene reconstruction from inputs degraded with

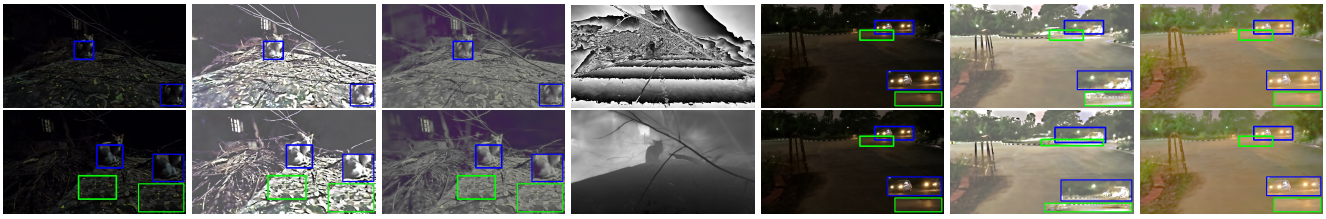

Figure 1. We depict two challenging dynamic scenarios in low-light. The first case involves self-occlusion/disocclusion caused by dynamic object motion, highlighted in blue in column 1 (low-light test view), column 2 (output from [4]), and column 3 (our output) across two time instances (i.e., each row). While the recent method [4] fails to reconstruct the scene and simply enhances the brightness, leading to saturation effects (in green bounding box), our method synthesizes the scene realistically. Column 4 shows the depth map (row 1 (output of [4]) and row 2 (our output)) to further validate the scene-aware reconstruction capabilities of L^2DGS. In the second case (Columns 5, 6, and 7) serve to illustrate the importance of modeling temporal illumination variations. L^2DGS accurately reconstructs both the cars (blue bounding boxes) and captures temporal lighting changes (green bounding boxes).

low-light artifacts is an underexplored area. In this work, we propose Low-Light Dynamic Gaussian Splatting (L^2DGS), a novel framework for directly reconstructing well-lit dynamic scenes from low-light sRGB inputs. The proposed method operates directly on in-the-wild captured inputs and does not rely on motion masks, camera metadata, or any form of explicit supervision from well-lit references. Built upon the recently introduced 3D-GS [14], L^2DGS inherits its strengths in high-fidelity rendering, real-time performance, and computational efficiency [11, 14, 33]. Unlike prior GS-based methods [11, 14, 20, 33, 39], which assign view-dependent color to each Gaussian, we instead associate each Gaussian with two distinct attributes: illumination and reflectance. We condition illumination on view and time, and treat reflectance as the scene intrinsic. The final color is the product of both attributes. As a result, the color associated with each Gaussian is a function of both time and viewpoint, enabling more accurate modeling of dynamic scenes, which most GS and NeRF-based approaches do not explicitly capture (see Fig. 1). We also propose a systematic methodology to transform latent well-lit scenes to low-light observations for enabling self-supervision. Our key contributions are as follows:

- The proposed L^2DGS directly reconstructs latent well-lit dynamic scenes from low-light inputs in a fully self-supervised manner. L^2DGS explicitly models scene color as both view and time-dependent, enabling robust disambiguation of color variations caused by object motion and lighting inconsistencies in dynamic situations.
- We propose an Occlusion-Disocclusion Network (OCD-Net) to model time-dependent intensity which arises due to illumination variations, often caused by object motion.
- To facilitate robust reconstruction under low illumination, we introduce a Signal Amplification Regularizer (SAR), which enhances signal strength and improves the fidelity of the reconstructed scene.
- We introduce Brightness Attenuation Features (BAFs), which are intrinsic attributes associated with each Gaussian, and a BAF enhancement module (BAFE-Net). Complemented by a Scene-Adaptive Structural Regularizer (SASR), these components together facilitate a scene-aware and structurally consistent transformation from well-lit to low-light domain to enable effective self-supervision.
- We introduce the first Low-Light Dynamic Video (L^2DyV) dataset of real low-light dynamic scenes and show through extensive synthetic and real-data experiments that L^2DGS consistently outperforms the state-of-the-art.

2. Related Works

Since L^2DGS can reconstruct 3D scenes and synthesize novel views, we review recent works in the NeRF and Gaussian Splatting (GS) domains, which share similar 3D reconstruction and view synthesis capabilities.

NeRF [26] and GS [14] have emerged as powerful techniques for reconstructing 3D scenes from monocular and multiview images or videos while enabling photorealistic novel view synthesis. NeRF represents scene geometry and appearance by modeling the color of a point as a view-dependent radiance field, leveraging volumetric rendering principles. In contrast, GS formulates the scene as a collection of 3D Gaussians, where the color associated with each Gaussian is explicitly modeled as view-dependent. Extensive research on NeRF-based [7, 9, 16, 46, 54] and GS-based [23, 30, 42, 48] methods has primarily focused on reconstructing scenes from well-lit inputs with clear scene visibility. Several NeRF-based approaches [1, 17, 18, 21, 38] and GS-based methods [5, 20, 28, 39, 44] have been proposed to extend these models to dynamic scenes, but they require a clear, well-lit scene.

Among works that deal with low-light static scenes, Lighting up NeRF [34] decomposes a static scene into view-

dependent and view-independent components and enhances the view-dependent component. Aleth-NeRF [3] models the formation of low-light scenes due to the presence of occluders and learns the concealing fields. Lush-NeRF [31] addresses the issue of low-light enhancement in the presence of blur and noise but uses an off-the-shelf method for low-light enhancement. NeRF-in-Dark [27] extends NeRF to operate in dark environments; however, it requires RAW images as input, which typically contain more information and processing flexibility. Lo-Gaussian [47] decomposes a scene into low illumination and normal light scene. DarkGS [50] proposes a method tailored for robotic applications, addressing illumination variability through color calibration. Gaussian-in-Dark [45] relies on the exposure and camera metadata as inputs for learning a well-lit scene. LL-Gaussian [32] uses a well-lit frame, estimated from a state-of-the-art network, as supervision. Luminance-GS [4] proposes view-dependent color mapping and curve adjustment. Bright-NeRF [36] uses RAW images. [43] proposes exposure correction inspired by light scattering and absorption theory. Lita-GS [52] extracts an illumination-invariant physical prior, which is used in the subsequent stage for lighting agnostic rendering. LL-GS [35] introduced a decomposable Gaussian representation for targeted color enhancement and a direction-based optimization strategy to maintain multi-view consistency. [19] uses a dual branch architecture where the first branch estimates the well-lit scene and the second branch estimates the illumination transition to address the low-light scene reconstruction for a static scene.

The most closely related work to ours is Lighting up NeRF [34], but key differences exist: (i) [34] is restricted to static scenes and cannot handle motion. (ii) Our approach jointly transforms illumination and reflectance in a scene-aware manner, preserving both color and brightness. (iii) The proposed Scene-Adaptive Structural Regularizer (SASR) enforces structural consistency. (iv) We model illumination as time- and view-dependent to capture dynamic motion. Unlike [34], our method is explicitly 3D scene-aware, and offers faster training with real-time inference capability.

3. Methodology

The proposed method directly reconstructs a well-lit dynamic scene in a fully self-supervised manner, given N low-light observations $\left\{I_d^t \in \mathbb{R}^{H\times W\times 3}\right\}|_{t=1}^{N}$, the corresponding camera poses ($R_t \in \mathbb{R}^{3\times 3}, \underline{T_t} \in \mathbb{R}^3, K \in \mathbb{R}^{3\times 3}$), and a sparse point cloud $\left\{\underline{\mu_i}^c \in \mathbb{R}^3\right\}|_{i=1}^{M}$ as inputs. R_t and $\underline{T_t}$ are camera rotation matrix and translation vector at time t. K is the camera intrinsic matrix. H (height) and W (width) represent the resolution of the input frames. We assume the 3D dynamic scene is composed of Gaussian primitives. In particular, we begin with a set of canonical Gaussians $G_i^c\left(\underline{\mu_i}^c, \Sigma_i^c\right)$, centered at $\underline{\mu_i^c}$ with covariance $\Sigma_i^c \in \mathbb{R}^{3\times 3}$. With each $G_i^c\left(\underline{\mu_i}^c, \Sigma_i^c\right)$ we associate view-dependent illumination $l_i^c(\underline{v}) \in \mathbb{R}^+$, view-independent reflectance $\underline{r_i} \in \mathbb{R}^3$, and opacity $o_i \in \mathbb{R}^+$. $\underline{v} \in \mathbb{R}^3$ is a view. The view-dependency is encoded using Spherical Harmonics (SH). At a query time t, we transform $G_i^c\left(\underline{\mu_i}^c, \Sigma_i^c\right)$ to a new state $G_i^t\left(\underline{\mu_i}^t, \Sigma_i^t\right)$ using Hexplane [1] and Occlusion-Disocclusion Network (OCD-Net) that we propose. Finally, we rasterize $G_i^t\left(\underline{\mu_i}^t, \Sigma_i^t\right)$ and estimate the illumination map $L_w(\underline{v}, t) \in \mathbb{R}^{H\times W}$, and reflectance map $R_w \in \mathbb{R}^{H\times W\times 3}$, product of which is the well-lit image $I_w(\underline{v}, t) \in \mathbb{R}^{H\times W\times 3}$. The rasterizer maps the 3D Gaussians onto the 2D image plane, orders them by depth to preserve proper visibility, and applies alpha blending to combine the splats into the final rendered image. During training, we transform $I_w(\underline{v}, t)$ to the low-light to enforce self-supervision. To achieve this, we further introduce two additional features $b_{1i} \in \mathbb{R}^+$ and $b_{2i} \in \mathbb{R}^+$ which we name as Brightness Attenuation Features (BAFs), to each Gaussian notationally, $G_i^c := \{\underline{\mu_i^c}, \Sigma_i^c, l_i^c, \underline{r_i}, o_i, b_{1i}, b_{2i}\}$ and $G_i^t := \{\underline{\mu_i^t}, \Sigma_i^t, l_i^t, \underline{r_i}, o_i, b_{1i}, b_{2i}\}$. This transformation is further aided by SASR, and SAR to maximize the signal strength. During inference, a queried camera pose and time instant enable the rendering of both well-lit and low-light images, ensuring spatial and temporally consistent scene rendering. An overview of the proposed framework is illustrated in Fig. 2. Next, we explain the transformation of $G_i^c\left(\underline{\mu_i}^c, \Sigma_i^c\right)$ to a state $G_i^t\left(\underline{\mu_i}^t, \Sigma_i^t\right)$ at time t along with well-lit scene estimation and self-supervision strategy.

3.1. 3D Dynamic GS Preliminaries

Following 3DGS [14], we decompose $\Sigma_i^c = Q_i S_i S_i^T Q_i^T$ where $Q_i \in \mathbb{R}^{3\times 3}$ is a rotation matrix and $S_i \in \mathbb{R}^{3\times 3}$ is a diagonal scaling matrix so as to constrain Σ_i^c to be positive semidefinite. Since the scene consists of dynamic objects (i.e., moving objects), we model the Gaussian state at time t using motion field. Motion field refers to the apparent motion of a Gaussian at any time t. The Hexplane [1] inspires our representation of the 3D motion field, which consists of 6 feature planes: XY, Zt, YZ, Xt, ZX, and Yt. In particular, given a G_i^c with $\underline{\mu_i}^c = [X, Y, Z]^T$ and a query time t, we extract and aggregate features from the motion field as

$$f(\underline{\mu_i}^t) = g\big(F(X,Y)\cdot F(Z,t) + F(Y,Z)\cdot F(X,t) + F(Z,X)\cdot F(Y,t)\big) \quad (1)$$

$F(\cdot)$ denotes the feature in the specified plane and $g(\cdot)$ is a feature aggregation function. The feature $f(\underline{\mu_i}^t)$ is then passed through three separate MLP-heads namely μ-Head, Q-Head, and S-Head respectively as shown in Fig. 2, similar to 4DGS [39], to estimate $\underline{\mu_i^t}$, Q_i^t, and S_i^t ($\Sigma_i^t = Q_i^t S_i^t (Q_i^t S_i^t)^T$). The 3D Gaussian when projected onto the image plane results in a 2D Gaussian with mean $\underline{\mu_t}^{2D} \in \mathbb{R}^2$ and covariance $\Sigma_t^{2D} \in \mathbb{R}^{2\times 2}$. The view-dependent illumi-

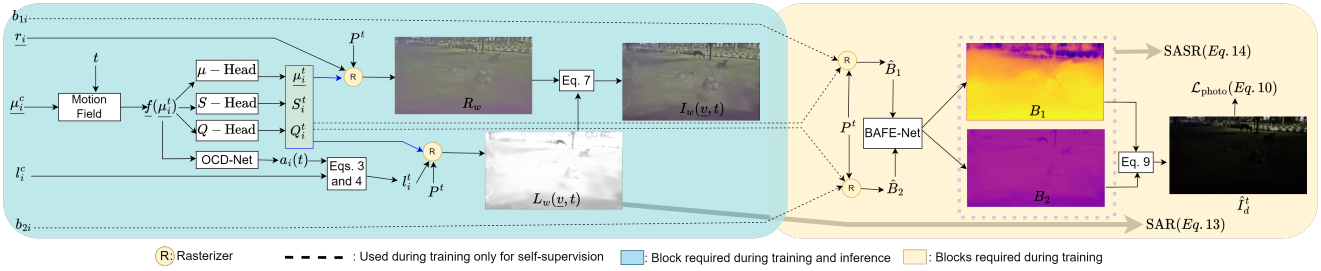

Figure 2. **The proposed framework.** We start with the canonical Gaussian G_i^c and estimate its state at time t as G_i^t. We decompose the estimated well-lit image I_w into a (view-time)-independent component R_w, and a (view-time)-dependent component L_w. Our BAFE-Net estimates the domain transformation quantities B_1 and B_2, which are solely used during training and discarded at inference. The example shown corresponds to the L^2DGS result for one of the scenes from L^2DyV.

nation l_i^c is given by

$$l_i^c = \sum_{j=1}^{(k+1)^2} b_j \cdot B_j(\underline{v}) \tag{2}$$

where $B_j(\underline{v})$ is the SH basis, and $b_j \in \mathbb{R}$ is the basis mixing coefficient. k is the number of SH bands.

3.2. OCD-Net and Well-lit Scene Estimation

The observed intensity can vary due to external factors such as object motion. We attribute the intensity variations caused by these external factors to the time-dependent transformation of l_i^c, which we denote as l_i^t.

$$l_i^t = \sum_{j=1}^{(k+1)^2} a_j(t) \cdot b_j \cdot B_j(\underline{v}) \tag{3}$$

where $a_j(t) \in \mathbb{R}^+$ is the scale factor governing time-dependent transformation such as occlusion or disocclusion.

OCD-Net: Mathematically, time dependency arises when G_i^t becomes occluded or disoccluded, either due to its own motion or the motion of another Gaussian G_m^t $(m \neq i)$ present in the 3D space. Hence, the need for OCD-Net, which we propose, arises to estimate the time-dependent scale factor $a_j(t)$.

$$a_j(t) = \mathcal{F}_2\big(\mathcal{F}_1(G_i^t, G_m^t; t)\big) = \text{OCD-Net}(f(\underline{\mu}_i^t)). \tag{4}$$

In the above Eq., $\mathcal{F}_2$ estimates the effective gain or loss of illumination, and $\mathcal{F}_1$ relates all the Gaussians (G_m^t) in the scene, that affect the overall illumination of G_i^t. OCD-Net models the combined effect of $\mathcal{F}_1$ and $\mathcal{F}_2$. The input to OCD-Net is $f(\underline{\mu}_i^t)$ (Eq. 1), which encodes the spatial and temporal scene dynamics obtained from the motion field, and the network maps this representation to $a_j(t)$, implicitly learning $\mathcal{F}_1$ and $\mathcal{F}_2$. We provide architectural details of OCD-Net in suppl.

Finally, we estimate $L_w(\underline{v}, t) \in \mathbb{R}^{H \times W}$ by applying alpha blending to l_i^t using a rasterizer.

$$[L_w(\underline{v}, t)]_{\underline{p}} = \sum_{i \in \mathcal{N}(\underline{p})} T_i \, \alpha_i \, l_i^t,$$
$$\text{where } \alpha_i = o_i \, \exp\!\big(-\tfrac{1}{2}\,(\underline{p} - \underline{\mu_t}^{2D})^T (\Sigma_t^{2D})^{-1} (\underline{p} - \underline{\mu_t}^{2D})\big)\,,$$
$$\text{and } T_i = \prod_{j=1}^{i-1} (1 - \alpha_j). \tag{5}$$

$\mathcal{N}(\underline{p})$ is the set of G_i^ts that affect the intensity of pixel $\underline{p} \in \mathbb{R}^2$. $[L_w(\underline{v}, t)]_{\underline{p}}$ is the value of $L_w(\underline{v}, t)$ at pixel $\underline{p}$. Similarly, we estimate the reflectance map $R_w \in \mathbb{R}^{H \times W \times 3}$ as

$$[R_w]_{\underline{p}} = \sum_{i \in \mathcal{N}(\underline{p})} T_i \, \alpha_i \, \underline{r_i} \tag{6}$$

Finally, the well-lit image is estimated as

$$I_w(\underline{v}, t) = L_w(\underline{v}, t) \circ R_w \tag{7}$$

where $\circ$ denotes Hadamard product.

Note that our approach is unlike Retinex theory-based methods which do not consider the temporal dimension and decompose an image into static view-dependent and view-independent components. Also most of the prior GS-based methods estimate a pixel intensity solely as a function of view. However, our method extends the dependence on time as well to handle the observed color variation due to external factors such as object motion. The time dependency in the color is effectively modeled by scaling the view-dependent l_i^c with the time-dependent $a_j(t)$ (Eqs. 2 - 4), which is used to render a well-lit image $I_w(\underline{v}, t)$ (Eq. 7).

3.3. Well-Lit to Low-Light Domain Transformation for Self-Supervision

BAFs. The enhancement process should respect the underlying scene geometry. Transformation of a well-lit scene

into a low-light scene should preserve the scene's structural and dynamic properties. During training, we transform $I_w(\underline{v}, t)$ to the observed low-light counterpart I_d^t using BAFs for self-supervision. To achieve this, we introduce two additional per-Gaussian features, b_{1i} and b_{2i} which are randomly initialized and are jointly optimized along with $\underline{\mu_i^c}, \Sigma_i^c, l_i^c, \underline{r_i}$ and o_i. To estimate BAFs in the image plane, we rasterize b_{1i} and b_{2i} to obtain $\hat{B}_1 \in \mathbb{R}^{H\times W}$ and $\hat{B}_2 \in \mathbb{R}^{H\times W}$ respectively, as follows.

$$[\hat{B}_1]_{\underline{p}} = \sum_{i\in\mathcal{N}(\underline{p})} T_i\, \alpha_i\, b_{1i} \qquad [\hat{B}_2]_{\underline{p}} = \sum_{i\in\mathcal{N}(\underline{p})} T_i\, \alpha_i\, b_{2i} \quad (8)$$

BAFE-Net. We found $\hat{B}_1$ and $\hat{B}_2$ to be insufficient in transforming the estimated well-lit image to a low-light image for supervision. Hence, we propose a BAF enhancement convolutional neural network (BAFE-Net) to yield the final 2D BAFs, $B_1 \in \mathbb{R}^{H\times W}$ and $B_2 \in \mathbb{R}^{H\times W}$. BAFE-Net takes the concatenated $\hat{B}_1$ and $\hat{B}_2$ as input, with shape $H \times W \times 2$. Note that BAFE-Net is jointly optimized along with all the Gaussian parameters. We provide details of BAFE-Net architecture in suppl. We estimate the low-light image as

$$\hat{I}_d^t = L_w(\underline{v}, t)^{\circ B_2} \circ R_w^{\circ B_1} \quad (9)$$

Here, $a^{\circ b}$ denotes element-wise exponentiation. Through SASR, B_1 and B_2 are regularized to align with the gradient of R_w and L_w, respectively, facilitating scene-aware domain transformation. The intuition to use the exponential in Eq. 9 is inspired by the gamma correction which is also differentiable. Importantly, our method is effective for both dynamic foreground objects and static backgrounds and can handle any sudden changes in observed intensity arising from object motion. Our approach enhances the robustness and perceptual quality of dynamic scene reconstruction under low-light conditions.

Adopting such a forward-modeling strategy, where the low-light image is synthesized from its well-lit counterpart, provides greater control over the constituent scene components, enabling selective enhancement to achieve a higher Signal to Noise Ratio (SNR). In contrast, an inverse reconstruction approach of estimating the well-lit scene by enhancing low-light scene components can inevitably amplify noise and introduce artifacts, further degrading reconstruction fidelity.

Overall, the proposed design confers (i) A significant advantage during inference (BAFs and BAFE-Net are not needed at test time), enabling real-time execution at more than 30 frames per second. (ii) Our approach enables enhanced signal strength while effectively suppressing signal-dependent noise in the estimated well-lit image. (iii) We show that incorporating BAFs provides a stronger constraint on the optimization process, effectively reducing color distortions.

It is evident that the estimation of a well-lit scene from low-light inputs is an ill-posed problem. Multiple combinations of L_w, R_w, B_1, and B_2 can lead to the same solution. This necessitates the need to incorporate constraints as discussed next.

3.4. Loss Function and Regularizers

Data fidelity loss: This enforces the reconstructed low-light image ($\hat{I}_d^t$) to be consistent with the input low-light image (I_d^t) and is given by

$$\mathcal{L}_{\text{photo}} = ||I_{\text{d}}^t - \hat{I}_{\text{d}}^t||_1 + (1 - SSIM((I_{\text{d}}^t + \epsilon)^\eta, (\hat{I}_{\text{d}}^t + \epsilon)^\eta)) \quad (10)$$

$||\cdot||_1$ represents L_1 norm. η is decayed exponentially from 0.95 to 0.5 and $\epsilon = e^{-5}$ for training stability.

Exposure regularizer: Zero-DCE [10] used this regularizer on the enhanced image, but we found it to offer better performance when applied to R_w in our framework. It is given as

$$\mathcal{L}_{\text{exp}} = \frac{1}{M}\sum_{k=1}^{M} |\overline{R}_w^k - e|, \quad (11)$$

Superscript $\bar{\cdot}$ denotes the average intensity value. The goal of this regularizer is to mitigate under or overexposed regions by controlling the deviation of the average value of R_w (that is, $\overline{R}_w^k \in \mathbb{R}^+$) in M non-overlapping local regions from a well-desired exposure level $e = 0.6$ in 16×16 window.

Edge-aware depth smoothness loss: It is important to enhance scene brightness while maintaining the underlying geometry. This regularizer constrains depth to follow a smooth transition based on the photometric characteristic of the estimated image.

$$\mathcal{L}_D = |\partial_x D|\, e^{-|\partial_x I_w|} + |\partial_y D|\, e^{-|\partial_y I_w|} \quad (12)$$

where $D \in \mathbb{R}^{H\times W}$ is the estimated depth map.

Signal Amplification Regularizer (SAR): In low-light imaging, photon scarcity leads to increased sensor noise, resulting in degraded signal quality. Consequently, direct estimation of $L_w(\underline{v}, t)$ and R_w can introduce noise artifacts, yielding a suboptimal well-lit reconstruction. To mitigate this issue, we aim to amplify the signal and minimize noise. Specifically, we enforce the maximization of $L_w(\underline{v}, t)$ via the following objective as

$$\mathcal{L}_{\text{L}} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} |1 - [L_w(\underline{v}, t)]_{i,j}| \quad (13)$$

$[L_w(\underline{v}, t)]_{i,j}$ is the value of L_w at location (i, j). We maximize $L_w(\underline{v}, t)$ as it controls brightness and varies with time. We set the value to 1 to keep $L_w(\underline{v}, t)$ as high as possible, which is essential for generating a well-lit scene as

$R_w \in [0, 1]$. However, maximizing R_w, analogous to Eqn. 13 has negligible impact on both qualitative and quantitative outcomes, as demonstrated in our ablation studies. A potential explanation is that R_w does not depend on view and time, and hence the noise is effectively averaged out when aggregating R_w across multiple viewpoints.

Scene-Adaptive Structural Regularizer (SASR): Transformation of a well-lit image into a low-light image should respect the scene's photometric properties and underlying structure, motivating a regularizer that enforces scene-aware transformation. Since B_1 and B_2 operate on R_w and L_w, respectively (Eq. 9), we heuristically encourage their gradients to be similar so that they can effectively capture the underlying structural consistency between R_w and L_w, and transform them accordingly. The BAF values must align with the observed data and adapt temporally to account for scene dynamics. We propose the following regularizers:

$$\mathcal{L}_{\text{B1}} = ||\beta_1 \nabla B_1 - \nabla R_w||_1 \quad \text{and} \quad \mathcal{L}_{\text{B2}} = ||\beta_2 \nabla B_2 - \nabla L_w||_1 \tag{14}$$

where ∇ denotes the gradient operator. The final objective function is given by

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_1 \mathcal{L}_{\text{exp}} + \lambda_2 \mathcal{L}_{\text{L}} + \lambda_3 \mathcal{L}_{\text{B1}} + \lambda_4 \mathcal{L}_{\text{B2}} + \lambda_5 \mathcal{L}_{\text{D}} \tag{15}$$

4. Experiments

Training. We adopt 4DGS [39] as our baseline and build our code on it. The resolution of motion field along the X, Y, and Z axes is 64, and the resolution in t axis is $\frac{N}{2}$. We follow a similar training strategy as [39], allocating first 3000 iterations for coarser training and the next 20000 iterations for finer training. During the coarse stage, the motion field and the OCD-Net are inactive. During the finer stage, all components of L^2DGS are active. Scene decomposition is performed during the coarse as well as fine stages, and the proposed method is end-to-end and fully self-supervised. We use a batch size of 2. The initial learning rate (LR) for BAFE-Net is set to 0.0016 and decays exponentially to a final value of 0.00016 using a decay factor of 0.01. Similarly, the LR for OCD-Net decays exponentially from 0.00016 to 0.000016 with the same factor. The learning rates for BAFs and reflectance are fixed at 0.00001, while the LR for illumination is set to 0.0025. All other components follow the learning rate schedule described in [39]. Optimization is carried out using Adam. We initialize reflectance and BAFs randomly. In all our experiments, we set β_1 and β_2 to 0.5. The regularization weights λ_1, λ_2, λ_3, λ_4, and λ_5 are chosen as 0.01, 0.05, 1.0, 1.0, and 0.001, respectively, after extensive experimentation. We discuss details of OCD-Net and BAFE-Net in the supplementary. All the experiments are conducted on an NVIDIA RTX 3090 GPU. The training time is ≈ 90 minutes for frames with resolution 450×800.

Datasets. To the best of our knowledge, there is currently no publicly available low-light video dataset featuring significant motion from both the camera and objects in the scene. To quantitatively evaluate our approach, we generate low-light frames using the method from Led-Net [53], applied to existing well-lit dynamic video datasets commonly used in NeRF and Gaussian Splatting methods, such as the iPhone dataset [8] and the HyperNeRF dataset [29]. [53] works by first converting a well-lit RGB image to LAB space. A saturation and exposure map are derived from the lightness channel, L_{well}, of a well-lit image which are then input to a low-light generation network to synthesize a low-light lightness L_{low}. A scale factor $\frac{L_{\text{low}}}{L_{\text{well}}}$ is computed and applied to the well-lit image to produce the final low-light output. Note that [53] is proposed for images and fails for some cases where the generated images do not look like a low-light image and have visible artifacts such as color distortions and flickering. Such unrealistic synthesized data was not considered in our experiments. We use 7 such synthetic videos. Importantly, ours as well as the baseline methods are independent of the low-light data synthesis process and do not incorporate any cues from the synthesis procedure. To evaluate performance on real data, we captured 12 challenging videos which we name as Low-Light Dynamic Videos (L^2DyV) dataset using handheld GoPro Hero10, POCO X3, and Samsung Galaxy M35 smartphones. Each video consists of 100 to 300 frames. While capturing L^2DyV, we ensured diverse camera trajectories, multi-class dynamic objects, diverse visibility across videos, and also within the same scene. These videos do not have ground truth. The resolution of the videos is 450×800 for all experiments. Details of L^2DyV are given in suppl. For a fair comparison, we trained all baselines for each of the synthetic and real videos.

4.1. Results and Comparisons

We show qualitative results in Fig. 3 and present quantitative metrics on synthetic data in Table 1. In the supple., we include additional results and comparisons with two more methods. We have also included the optimization time for all the baselines in the supple. We conduct both qualitative and quantitative comparisons against state-of-the-art low-light NeRF-based methods [3, 34] and low-light GS-based methods [4, 47, 50, 52]. For the synthetic datasets, we exclude Dark-GS [50] due to its reliance on camera metadata, although we do evaluate it on L^2DyV. Lo-Gaussian [47] failed to reconstruct the scene, producing entirely black images for all the scenarios and is hence excluded from our main comparisons. The inference code of Lita-GS [52] is publicly not available, hence we implemented it ourselves for purpose of evaluation. Our method reconstructs the 'Apple' though it is occluded and disoccluded by hand (Fig. 3, row 1). On real data, L^2DGS is able to reconstruct the scene while preserving well-lit areas (Fig. 3, row 2), can

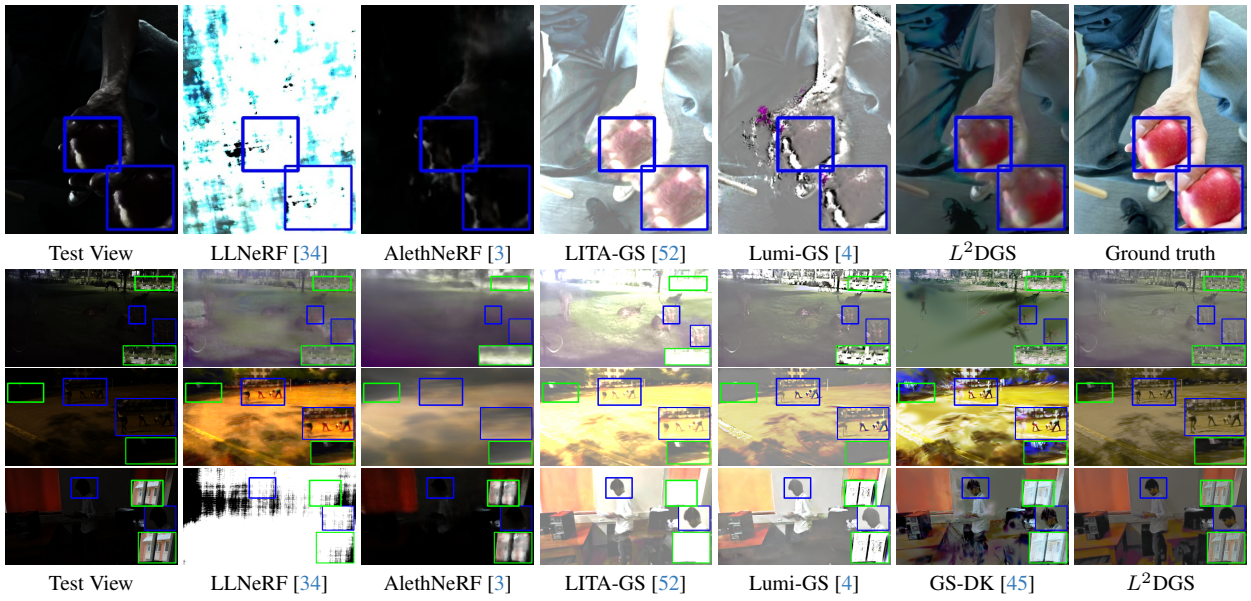

Figure 3. **Comparisons with NeRF and GS-based methods.** Synthetic scene: Apple (row 1), and L^2DyV (real-data): Deer (row 2), Skating (row 3), File (row 4). Our method effectively preserves scene structure and synthesizes well-lit views in a scene-illumination-and-dynamics-aware manner in both static and dynamic regions. Additionally, it preserves shadows. We have additionally included qualitative comparisons with 4DGS [39] in suppl.

Method	Mochi									Handwavy								
	Dynamic			Static			Overall			Dynamic			Static			Overall		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AlethNeRF [3]	14.32	0.50	0.17	15.53	0.66	0.23	14.95	0.59	0.40	11.93	0.38	0.10	12.54	0.39	0.51	12.41	0.39	0.61
LLNeRF [34]	13.60	0.60	0.08	13.37	0.67	0.10	13.45	0.64	0.18	11.33	0.42	0.05	**13.15**	0.54	0.24	**12.78**	0.52	0.29
LITA-GS [52]	10.37	0.41	0.13	10.71	0.64	0.17	10.51	0.54	0.31	9.79	0.40	0.07	9.74	0.36	0.35	9.73	0.36	0.42
Luminance-GS [4]	14.06	0.59	0.14	14.87	0.72	0.18	14.51	0.66	0.32	12.13	0.50	0.07	11.97	0.56	0.33	11.99	0.55	0.39
4DGS [39]	7.03	0.15	0.15	6.41	0.29	0.20	6.65	0.23	0.35	5.86	0.04	0.09	5.75	0.06	0.44	5.77	0.06	0.53
L^2DGS	**21.00**	**0.78**	**0.07**	**16.13**	**0.77**	**0.09**	**17.61**	**0.77**	**0.16**	**12.21**	**0.51**	**0.04**	12.21	**0.62**	**0.22**	11.72	**0.58**	**0.26**
Method	**Apple**									**Creeper**								
	Dynamic			Static			Overall			Dynamic			Static			Overall		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
AlethNeRF [3]	4.89	0.04	0.16	6.30	0.04	0.47	5.89	0.04	0.63	6.85	0.02	0.43	5.07	0.02	0.23	6.15	0.02	0.66
LLNeRF [34]	6.43	0.27	0.16	5.19	0.23	0.47	5.47	0.24	0.63	6.02	0.07	0.50	6.36	0.09	0.26	6.14	0.08	0.77
LITA-GS [52]	10.22	0.24	0.12	9.37	0.25	0.38	9.48	0.25	0.51	11.76	0.49	0.03	11.37	0.56	0.19	11.44	0.55	0.23
Luminance-GS [4]	12.22	0.32	0.11	12.62	0.47	0.33	12.50	0.43	0.44	11.48	0.30	0.16	**14.50**	0.48	0.08	12.30	0.36	0.24
4DGS [39]	5.06	0.07	0.13	6.35	0.06	0.39	5.98	0.06	0.53	6.98	0.04	0.36	5.22	0.05	0.19	6.29	0.04	0.55
L^2DGS	**13.26**	**0.58**	**0.06**	**13.67**	**0.65**	**0.18**	**13.31**	**0.63**	**0.24**	**15.49**	**0.55**	**0.12**	12.01	**0.60**	**0.07**	**13.96**	**0.57**	**0.19**

Table 1. Comparisons across synthetic simulated low-light scenes. L^2DGS consistently outperforms all competing methods by a significant margin in dynamic regions. Metric values are computed on well-lit images.

handle shadows due to object motion, and can better handle background visibility (Fig. 3, row 3). In the indoor scene (Fig. 3, row 4), our method recovers human face without introducing artifacts in the well-lit window area. These results demonstrate that L^2DGS yields reliable results in diverse situations, and consistently outperforms competing NeRF and GS-based methods. Notably, ours is the only method that simultaneously learns to reconstruct the scene and enhance low-light inputs. In contrast, other methods fail to preserve scene content and structural integrity and also suffer from saturation effects. The depth maps for each method are provided in supple. Please refer to the supple and its accompanying video for additional qualitative results.

User study on real scene synthesis: To further assess perceptual quality, we conducted two complementary user studies. (i) Binary study: Participants were asked to make a binary choice from different methods and select a single preferred result. This binary choice design encourages clear and decisive judgment. (ii) Rating study: Participants rated the output of each method on a scale of 1 (poor) to 5 (best) on each scene. In total, 62 individuals participated across the two studies, including high school students, undergraduates, postgraduates, and industry professionals. Many of them did not have a background in computer vision, ensuring that the evaluation reflects general human perception. The studies covered outputs from six methods across 12 real-world scenes. Note that the participant groups were not identical across the two studies. Evaluation was conducted

using five criteria. (i) Natural-looking brightness and lighting. (ii) Foreground reconstruction. (iii) Preservation of scene structure. (iv) Faithful and natural color. (v) Minimal reconstruction artifacts. The mean scores for each method are given in Table 2. Our method received the highest average ratings (across all criteria), highlighting its ability to synthesize images that align with human preferences.

	Criteria (i)	Criteria (ii)	Criteria (iii)	Criteria (iv)	Criteria (v)
AlethNeRF [3]	0.00 / 1.56	0.01 / 1.36	0.10 / 1.58	0.10 / 1.26	0.01 / 1.34
LLNeRF [34]	23.14 / 3.05	2.01 / 2.97	1.12 / 2.91	25.80 / 2.84	6.03 / 2.83
LITA-GS [52]	1.64 / 1.59	0.18 / 2.32	2.95 / 2.54	1.74 / 1.98	6.33 / 2.36
GS-DK [45]	0.00 / 2.09	2.23 / 2.21	0.29 / 2.03	0.001 / 2.13	1.11 / 2.02
Luminance-GS [4]	0.02 / 2.39	2.01 / 2.81	1.10 / 2.36	0.001 / 2.67	6.11 / 2.55
L^2DGS	**75.20 / 4.05**	**93.57 / 4.33**	**94.34 / 4.08**	**72.36 / 4.36**	**80.41 / 4.28**

Table 2. User study on L^2DyV (values in % / Average ratings (scale: 1-5). In both studies, L^2DGS consistency performs better, showing its ability to align with the human perception.

5. Ablations

In this section, we study the effect of different components of our framework.

Without SSIM. We remove SSIM loss from $\mathcal{L}_{\text{photo}}$ and train the network. Qualitative analysis (in supple) as well as quantitative analysis (Table 3) suggest that SSIM helps to enhance scene visibility and sharpness. **Without $\mathcal{L}_{\text{exp}}$.** This loss enforces brightness consistency in a local neighborhood. In the absence of $\mathcal{L}_{\text{exp}}$, the estimated output is again low-lit. Hence, the quantitative value suffers. **Without $\mathcal{L}_{\mathbf{L}}$.** In the absence of $\mathcal{L}_L$, the estimated scene suffers from low visibility. **Without $\mathcal{L}_{\mathbf{B}}$.** Though without $\mathcal{L}_B$, the scene brightness is enhanced, it does not follow the scene dynamics and causes smearing of edges. **Without BAFE-Net.** BAFE-Net enhances BAFs in the 2D space by leveraging local neighborhood information, thereby facilitating a more effective transformation of the scene domain. In its absence, the result is low-light. **Without OCD-Net**. The estimated well-lit image exhibits degradation in both qualitative appearance and quantitative measures. **Without $\mathcal{L}_{\mathbf{D}}$.** The goal of $\mathcal{L}_D$ is to aid in the depth learning, which should be consistent with the scene photometry. Its absence can lead to blurry outputs. **R maximization ($\mathcal{L}+0.05|1-R_w|$).** It has insignificant effect on the final output. **Maximization of R_w instead of L_w in Eq. 13.** This leads to very poorly lit images. **Ablation on B_1 and B_2 (Eq. 9).** We conduct two independent ablations, setting $B_1 = 1$ and $B_2 = 1$ separately. Both configurations lead to color distortions. In summary, the estimation of the well-lit image $I_w(\underline{v}, t)$ depends on a single-channel view and time dependent illumination map $L_w(\underline{v}, t)$ and a three-channel intrinsic reflectance map R_w. The loss term $\mathcal{L}_L$ encourages $L_w(\underline{v}, t)$ to maximize illumination strength, while $\mathcal{L}_{\text{exp}}$ constrains R_w to maintain an average intensity value e within a local neighborhood, allowing R_w the flexibility to adjust individual pixel values locally. The proposed SASR module enables scene-aware transformation between well-lit and low-light domains. Furthermore, the SSIM component in $\mathcal{L}_{\text{photo}}$ helps preserve structural similarity between the ground truth I_d^t and the estimated $\hat{I}_d^t$ across multiple intensity levels induced by η. Together, these components work synergistically to guide I_w toward its well-lit appearance while maintaining scene awareness.

Experimental setting	Dynamic			Static			Overall		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
W/o SSIM	14.12	0.41	0.20	12.84	0.52	0.29	13.20	0.49	0.49
W/o $\mathcal{L}_{exp}$	8.36	0.23	0.11	7.30	0.30	0.19	7.65	0.27	0.31
W/o $\mathcal{L}_D$	16.26	0.56	0.11	13.57	0.61	0.14	14.27	0.60	0.25
W/o $\mathcal{L}_B$	14.40	0.53	0.10	14.71	0.68	0.16	14.56	0.62	0.26
W/o $\mathcal{L}_L$	7.82	0.18	0.14	6.57	0.23	0.22	6.99	0.22	0.37
W/o BAFE-Net	6.73	0.13	0.26	5.87	0.21	0.42	6.14	0.19	0.68
W/o OCD-Net	15.06	0.58	0.07	12.73	0.70	0.12	13.36	0.65	0.19
$\mathcal{L} + 0.05\|1-R_w\|$	16.73	0.63	0.07	13.67	0.70	0.12	14.57	0.68	0.19
Maximize R_w (Eq. 13)	7.74	0.17	0.18	6.45	0.22	0.23	6.89	0.21	0.37
$B_1 = 1$	14.23	0.57	0.07	11.66	0.65	0.13	12.46	0.62	0.20
$B_2 = 1$	12.59	0.41	0.11	12.95	0.57	0.16	12.54	0.51	0.27
With $\mathcal{L}$	**16.73**	**0.64**	**0.07**	**13.86**	**0.71**	**0.12**	**14.68**	**0.69**	**0.18**

Table 3. Average values across Sriracha, Creeper, Mochi, and Apple videos. Overall $\mathcal{L}$ delivers the best performance. Metric values are computed on well-lit images.

Note that we also evaluate the estimated $\hat{I}_d^t$ against the ground truth I_d^t. The average PSNR for dynamic/static/overall regions is 31.28/32.96/31.86, with SSIM of 0.92/0.96/0.94, confirming the effectiveness of the well-lit to low-light modeling. **Limitations:** Handling extreme low-light scenes remains a limitation of our approach. Our method requires camera poses and a 3D point cloud as inputs. However, in extremely dark conditions, accurate estimation of these inputs is challenging, as even SOTA methods struggle to recover reliable scene information.

6. Conclusions

We proposed an end-to-end method L^2DGS for reconstructing well-lit dynamic scenes from challenging low-light video sequences with known camera poses. The real Low-Light Dynamic Videos, L^2DyV dataset introduced in our experiments spans a wide range of complexity commonly encountered in low-light environments. These videos exhibit low overall visibility, with some regions appearing relatively well-lit. Despite the non-uniform visibility, our method effectively reconstructs well-lit scenes without overexposing the brighter regions. Our approach introduces several key components: SAR, SASR, OCD-Net, BAFs, and a time-view-dependent intensity decomposition. On synthetic data, L^2DGS performs the best. On L^2DyV, we present qualitative results and user studies, and show that our method outperforms all competing methods. Notably, L^2DGS operates without requiring any form of supervision, such as well-lit ground truth images, motion masks, or camera metadata.

Acknowledgement: Support from Institute of Eminence (IoE) project No. SB22231269EEETWO005001 for Research Centre in Computer Vision and IoE project No. SP22231231CPETWOSSAHOC for Center of Excellence in Sports Science and Analytics is gratefully acknowledged.

References

[1] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2, 3

[2] Shivam Chhirolya, Sameer Malik, and Rajiv Soundararajan. Low light video enhancement by learning on static videos with cross-frame attention. *arXiv preprint arXiv:2210.04290*, 2022. 1

[3] Ziteng Cui, Lin Gu, Xiao Sun, Xianzheng Ma, Yu Qiao, and Tatsuya Harada. Aleth-nerf: Illumination adaptive nerf with concealing field assumption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1435–1444, 2024. 1, 3, 6, 7, 8

[4] Ziteng Cui, Xuangeng Chu, and Tatsuya Harada. Luminance-gs: Adapting 3d gaussian splatting to challenging lighting conditions with view-adaptive curve adjustment. In *CVPR*, 2025. 1, 2, 3, 6, 7, 8

[5] Pinxuan Dai, Peiquan Zhang, Zheng Dong, Ke Xu, Yifan Peng, Dandan Ding, Yujun Shen, Yin Yang, Xinguo Liu, Rynson WH Lau, et al. 4d gaussian videos with motion layering. *ACM Transactions on Graphics (TOG)*, 44(4):1–14, 2025. 2

[6] Kang Du, Zhihao Liang, Yulin Shen, and Zeyu Wang. Gs-id: Illumination decomposition on gaussian splatting via adaptive light aggregation and diffusion-guided material priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26220–26229, 2025. 1

[7] Chen Gao, Yipeng Wang, Changil Kim, Jia-Bin Huang, and Johannes Kopf. Planar reflection-aware neural radiance fields. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 2

[8] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 6

[9] Yiming Gao, Yan-Pei Cao, and Ying Shan. Surfelnerf: Neural surfel radiance fields for online photorealistic reconstruction of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 108–118, 2023. 2

[10] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[11] Siming He, Zach Osman, and Pratik Chaudhari. From nerfs to gaussian splats, and back. *arXiv preprint arXiv:2405.09717*, 2024. 2

[12] Shuai Jin, Yuhua Qian, Feijiang Li, Guoqing Liu, and Xinyan Liang. Pasd: A pixel-adaptive swarm dynamics approach for unsupervised low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9070–9079, 2025. 1

[13] Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 1

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3

[15] Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. Efficient diffusion as low light enhancer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 21277–21286, 2025. 1

[16] Yixing Lao, Xiaogang Xu, zhipeng cai, Xihui Liu, and Hengshuang Zhao. Corresnerf: Image correspondence priors for neural radiance fields. In *Advances in Neural Information Processing Systems*, pages 40504–40520. Curran Associates, Inc., 2023. 2

[17] Lingzhi Li, Zhen Shen, Li Shen, Ping Tan, et al. Streaming radiance fields for 3d video synthesis. In *Advances in Neural Information Processing Systems*, 2022. 2

[18] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[19] Ze Li, Feng Zhang, Xiatian Zhu, Meng Zhang, Yanghong Zhou, and P. Y. Mok. Robust low-light scene restoration via illumination transition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6188–6197, 2025. 1, 3

[20] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21136–21145, 2024. 2

[21] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7664–7674, 2024. 2

[22] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10561–10570, 2021. 1

[23] Tao Lu, Ankit Dhiman, R Srinath, Emre Arslan, Angela Xing, Yuanbo Xiangli, R Venkatesh Babu, and Srinath Sridhar. Turbo-gs: Accelerating 3d gaussian fitting for high-quality radiance fields. *arXiv preprint arXiv:2412.13547*, 2024. 2

[24] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, page 4. Northumbria University, 2018. 1

[25] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5637–5646, 2022. 1

[26] 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 *ECCV*, 2020. 2

[27] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16190–16199, 2022. 3

[28] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024. 2

[29] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 6

[30] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20051–20060, 2024. 2

[31] Zefan Qu, Ke Xu, Gerhard Hancke, and Rynson Lau. Lushnerf: Lighting up and sharpening nerfs for low-light scenes. In *Advances in Neural Information Processing Systems*, pages 109871–109893. Curran Associates, Inc., 2024. 1, 3

[32] Hao Sun, Fenggen Yu, Huiyao Xu, Tao Zhang, and Changqing Zou. Ll-gaussian: Low-light scene reconstruction and enhancement via gaussian splatting for novel view synthesis, 2025. 3

[33] F Tosi, Y Zhang, Z Gong, E Sandström, S Mattoccia, MR Oswald, and M Poggi. How nerfs and 3d gaussian splatting are reshaping slam: A survey. arxiv 2024. *arXiv preprint arXiv:2402.13255*. 2

[34] Haoyuan Wang, Xiaogang Xu, Ke Xu, and Rynson W.H. Lau. Lighting up nerf via unsupervised decomposition and enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12632–12641, 2023. 1, 2, 3, 6, 7, 8

[35] Haoran Wang, Jingwei Huang, Lu Yang, Tianchen Deng, Gaojing Zhang, and Mingrui Li. Llgs: Unsupervised gaussian splatting for image enhancement and reconstruction in pure dark environment. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13957–13963, 2025. 1, 3

[36] Min Wang, Xin Huang, Guoqing Zhou, Qifeng Guo, and Qing Wang. Bright-nerf: Brightening neural radiance field with color restoration from low-light raw images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8): 7817–7825, 2025. 3

[37] Sen Wang, Shao Zeng, Tianjun Gu, Zhizhong Zhang, Ruixin Zhang, Shouhong Ding, Jingyun Zhang, Jun Wang, Xin Tan, Yuan Xie, and Lizhuang Ma. From enhancement to understanding: Build a generalized bridge for low-light vision via semantically consistent unsupervised fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13804–13814, 2025. 1

[38] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[39] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2, 3, 6, 7

[40] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2022. 1

[41] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2281–2290, 2020. 1

[42] Kai Xu, Tze Ho Elden Tse, Jizong Peng, and Angela Yao. Das3r: Dynamics-aware gaussian splatting for static scene reconstruction. *arXiv preprint arXiv:2412.19584*, 2024. 2

[43] Kai Xu, Mingwen Shao, Yuanjian Qiao, and Yan Wang. Physical-aware neural radiance fields for efficient exposure correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):8906–8914, 2025. 3

[44] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[45] Sheng Ye, Zhen-Hui Dong, Yubin Hu, Yu-Hui Wen, and Yong-Jin Liu. Gaussian in the dark: Real-time view synthesis from inconsistent dark images using gaussian splatting. In *Computer Graphics Forum*, page e15213. Wiley Online Library, 2024. 1, 3, 7, 8

[46] Ze-Xin Yin, Peng-Yi Jiao, Jiaxiong Qiu, Ming-Ming Cheng, and Bo Ren. Ms-nerf: Multi-space neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 2

[47] Jingjiao You, Yuanyang Zhang, Tianchen Zhou, Yecheng Zhao, and Li Yao. Lo-gaussian: Gaussian splatting for low-light and overexposure scenes through simulated filter. *Eurographics Association: Eindhoven, The Netherlands*, 2024. 1, 3, 6

[48] 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 (CVPR)*, pages 19447–19456, 2024. 2

[49] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4967–4976, 2021. 1

[50] Tianyi Zhang, Kaining Huang, Weiming Zhi, and Matthew Johnson-Roberson. Darkgs: Learning neural illumination and 3d gaussians relighting for robotic exploration in the dark. *2024 International Conference on Intelligent Robots and Systems (IROS)*, 2024. 3, 6

[51] Chuanjun Zheng, Daming Shi, and Wentian Shi. Adaptive unfolding total variation network for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4439–4448, 2021. 1

[52] Han Zhou, Wei Dong, and Jun Chen. Lita-gs: Illumination-agnostic novel view synthesis via reference-free 3d gaussian splatting and physical priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 21580–21589, 2025. 1, 3, 6, 7, 8

[53] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pages 573–589. Springer, 2022. 1, 6

[54] Lin Zhu, Kangmin Jia, Yifan Zhao, Yunshan Qi, Lizhi Wang, and Hua Huang. Spikenerf: Learning neural radiance fields from continuous spike stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6285–6295, 2024. 2