P2GS: Physical Prior-guided Gaussian Splatting for Photometrically Consistent Urban Reconstruction

Supplementary Material

A. CARLA Dataset

We build a synthetic dataset in CARLA 0.9.15 [2] to isolate exposure variation while fixing geometry, camera poses, and rendering pipeline. This enables controlled evaluation of view-invariant HDR radiance recovery and cross-camera photometric consistency.

A.1. Generation Protocol

All scenarios are generated deterministically in synchronous mode (fixed tick), with dynamic actors disabled to keep the scene static. A single ego vehicle follows a scripted route in Town01, Town03, and Town04. Weather, time-of-day, and illumination are fixed (e.g., clear sky and constant sun altitude). Sensors are time-synchronized at a constant frame rate (10 Hz), and each sequence is 10 s (100 frames).

A.2. Sensor Mounting and Imaging Parameters

Camera rig. A rigid tri-camera rig is attached to the ego vehicle (CARLA coordinates: x forward, y right, z up). Per-sequence extrinsics are constant and shared across exposure settings. Table 1 lists mount poses (vehicle→camera) as $(\text{loc}[m],\ \text{rot}[^\circ])$.

Table 1. Camera mounting (vehicle → camera). Location [m], rotation [deg].

Camera	Loc (x, y, z)	Rot $(\text{roll}, \text{pitch}, \text{yaw})$
Front (ID 0)	$(1.539,\ 0.025,\ 3.845)$	$(0.696,\ 0.420,\ 0.338)$
Front-left (ID 1)	$(1.494,\ -0.091,\ 3.845)$	$(0.003,\ 1.387,\ -44.205)$
Front-right (ID 2)	$(1.489,\ 0.095,\ 3.846)$	$(0.189,\ 0.111,\ 44.756)$

Image formation. Each camera records 1920×1300 sRGB images at 10 Hz with 60° horizontal FOV. Intrinsics are computed from FOV and image size:

$$f_x = f_y = \frac{W}{2\tan(\text{FOV}/2)}, \quad c_x = \frac{W-1}{2},\ c_y = \frac{H-1}{2},$$

which yields $f_x = f_y \approx 1662.77$, $c_x = 959.5$, $c_y = 649.5$ for W=1920, H=1300, FOV=60°. We export KITTI-style 3×4 projection matrices $\mathbf{P}_k = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ for $k \in \{0, 1, 2\}$.

LiDAR. A roof-mounted LiDAR (vehicle→LiDAR location $(0.0,\ 0.0,\ 1.73)$ m) runs at 10 Hz with 64 channels, 1.3 M points/s, range 100 m, vertical FOV $[-25^\circ,\ 15^\circ]$, and no synthetic noise. LiDAR scans are used only for 3D initialization. For interoperability with KITTI-style tooling, we provide per-frame LiDAR poses and camera projection files; basis conversion to the KITTI conventions (e.g., y-axis flip for LiDAR) is applied when exporting calibration.

Photometric masks. For sky-sensitive analyses, we provide per-frame sky masks derived from semantic segmentation. These masks are binary and aligned to each RGB frame.

A.3. Exposure Control and Photometric Settings

Exposure is controlled solely via ISO under manual exposure mode. Shutter speed and aperture remain fixed, and the camera response / ISP is held constant. We provide two subsets:

- **ISO-Const**: a fixed ISO of **8** for all frames and cameras . Illumination is constant and identical across cameras.
- **ISO-Var**(*std*): per-frame ISO is drawn i.i.d. from $\mathcal{N}(\mu{=}8,\ \sigma \in \{2, 4\})$, integer-rounded and lower-bounded, with the same geometry/poses as ISO-Const. This yields controlled photometric drift while preserving geometry and poses.

RGB images are recorded in sRGB (8-bit); internal linear-radiance values are used only for sanity checks. No denoising or sharpening is applied.

B. Details of the Optimization

The proposed model is trained under a unified loss that jointly optimizes geometric, radiative, and photometric parameters. Since the entire pipeline is differentiable, the SH coefficients $\mathbf{c}_k$, exposure scales e_i, and tone-mapping gammas γ_i are jointly optimized via automatic differentiation, enabling unsupervised separation of scene-specific and camera-specific factors.

The total loss is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \lambda_{\text{exp}}\mathcal{L}_{\text{exp}} + \mathcal{L}_{\text{reg}}, \tag{1}$$

where $\lambda_{\text{exp}} = 0.1$, and $\mathcal{L}_{\text{reg}}$ is the sum of regularization terms.

Photometric reconstruction loss. $\mathcal{L}_{\text{photo}}$ minimizes the error between the observed LDR image I_i and the reconstructed $\hat{I}^i_{\text{LDR}} = T_i(e_i \cdot \hat{I}^i_{\text{linear}})$:

$$\mathcal{L}_{\text{photo}} = (1 - \lambda_{\text{dssim}})\mathcal{L}_{L1} + \lambda_{\text{dssim}}\mathcal{L}_{\text{SSIM}}, \tag{2}$$

where $\mathcal{L}_{L1} = \frac{1}{HW}\sum_u |\hat{I}^i_{\text{LDR}}(u) - I_i(u)|$ and $\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\hat{I}^i_{\text{LDR}}, I_i)$. We set $\lambda_{\text{dssim}} = 0.2$. L1 promotes pixel-level accuracy, while SSIM encourages local structural consistency, allowing optimization of both luminance and perceptual fidelity.

Relative exposure consistency loss. $\mathcal{L}_{\text{exp}}$ enforces linearity of relative exposure in HDR space. For any view pair (i, j) with relative ratio $\alpha_{ij} = e_j/e_i$, the following should hold:

$$\hat{I}^j_{\text{linear}} = \alpha_{ij}\, \hat{I}^i_{\text{linear}}. \tag{3}$$

Thus, the loss is defined as

$$\mathcal{L}_{\text{exp}} = \frac{1}{M} \sum_{(i,j)\in\mathcal{P}} \left\| \alpha_{ij}\hat{I}^i_{\text{linear}} - \hat{I}^j_{\text{linear}} \right\|_1, \tag{4}$$

where $\mathcal{P}$ is the set of sampled view pairs and M is the number of pixels. This term requires no ground-truth exposure and directly enforces HDR-space consistency, decoupling exposure from tone nonlinearity.

Regularization. To prevent global-scale ambiguity in exposure, suppress variance-induced instability, and avoid non-physical tone curves, we define

$$\begin{aligned}\mathcal{L}_{\text{reg}} = \lambda_{\text{escale}}\mathbb{E}_i[(e_i - 1)^2] + \lambda_{\text{evar}}\text{Var}(e_i) \\ +\lambda_\gamma \mathbb{E}_i[(\gamma_i - \gamma_{\text{prior}})^2].\end{aligned} \tag{5}$$

Each term targets a distinct failure mode of unsupervised exposure–tone disentanglement and is weighted independently.

(a) The first term : Global scale normalization. Because $\mathcal{L}_{\text{exp}}$ constrains only relative exposure ratios $\alpha_{ij} = e_j/e_i$, the solution is invariant under the transformation $\{e_i\} \to \{c\, e_i\}$ for any constant $c > 0$. This leaves a degree of freedom that can arbitrarily rescale the HDR radiance $\hat{I}_{\text{linear}}$, impeding identifiability and destabilizing optimization. We remove this ambiguity by softly anchoring the absolute scale of e_i at 1.0. The penalty centers the exposure distribution without overconstraining inter-view differences. In practice, this term improves gradient conditioning by preventing the HDR field from absorbing arbitrary global gains.

(b) The second term : Relative consistency via variance suppression. Even with the global mean fixed, large dispersion of $\{e_i\}$ across views can mimic tone nonlinearity and cause brittle solutions, especially under limited view overlap or photometric noise. We therefore penalize the empirical variance of exposure: This term encourages tight clustering of exposures around a common scale while allowing small, data-driven deviations to remain. Empirically, setting $\lambda_{\text{evar}} > \lambda_{\text{escale}}$ implements a hierarchical constraint: the global level is weakly anchored by ***The first term***, whereas ***The second term*** strongly regularizes inter-view spread, yielding improved numerical stability and temporal smoothness.

(c) The third term : Tone regularization with physical prior. Unconstrained optimization of per-view gamma can exploit tone nonlinearity to explain exposure or radiance mismatch, leading to non-physical tone curves and optimization collapse. We impose a soft prior around the sRGB standard The quadratic penalty preserves gradient flow and allows data-driven deviations when statistically justified, while discouraging extreme, non-physical solutions. This term directly stabilizes the interaction between exposure scaling and tone mapping, which is critical for preserving HDR linearity in the radiance field. The differentiable pipeline enables fully unsupervised separation of scene-specific radiance $\hat{I}_{\text{linear}}$ and camera-specific parameters (e_i, γ_i), effectively removing camera-dependent illumination artifacts inherent in conventional 3DGS.

C. Metrics

In addition to standard image quality metrics (PSNR, SSIM, and LPIPS), we introduce two novel metrics designed to evaluate photometric stability and exposure robustness in 3DGS for autonomous driving scenes: the HDR Inconsistency Score (HIS) and the Standard Deviation of Luminance (Std-Luminance).

C.1. HDR Inconsistency Score

HIS measures the temporal stability of exposure compensation in HDR-enabled reconstructions. It quantifies frame-to-frame luminance variation after exposure correction as follows:

$$\text{HIS} = \frac{1}{T-1}\sum_{t=1}^{T-1} \mathcal{D}_{\text{HDR}}(R_t, R_{t+1}, e_t, e_{t+1}), \tag{6}$$

$$\mathcal{D}_{\text{HDR}} = \left\| \text{OETF}^{-1}(R_t)\cdot e_t - \text{OETF}^{-1}(R_{t+1})\cdot e_{t+1} \right\|_2,$$

where T denotes the total number of frames, and t indexes time. $R_t \in [0,1]^{H\times W\times 3}$ is the rendered image at time t, $e_t \in \mathbb{R}^+$ is the exposure scale at time t, and $\text{OETF}^{-1} : [0,1] \to \mathbb{R}^+$ denotes the inverse Opto-Electronic Transfer Function that maps sRGB to linear RGB space. The operator $\|\cdot\|_2$ represents the L2 norm computed over all pixels. Lower HIS values indicate more stable exposure compensation and higher temporal illumination consistency across frames.

C.2. Standard Deviation of Luminance

Std-Luminance evaluates global brightness consistency across all rendered views after exposure compensation. For

Table 2. Quantitative comparison of ours versus dynamic scene models on the CARLA Dataset. Our method preserves both robustness and reconstruction and NVS quality.

Methods	Train data	Reconstruction SSIM ↑	PSNR ↑	LPIPS ↓	ΔPSNR ↑	Novel view synthesis SSIM ↑	PSNR ↑	LPIPS ↓	ΔPSNR ↑
StreetGS[4]	ISO std2	0.733	20.16	0.532	-2.07	0.512	11.04	0.804	-0.44
	ISO std4	0.699	18.09	0.564		0.496	10.60	0.818	
DeSIRE-GS[3]	ISO std2	0.726	16.94	0.379	-4.69	0.699	16.35	0.401	-3.83
	ISO std4	0.647	12.25	0.484		0.633	12.52	0.485	
OmniRe[1]	ISO std2	0.733	19.84	0.534	-1.73	0.511	11.04	0.806	**-0.41**
	ISO std4	0.699	18.11	0.565		0.498	10.63	0.823	
Ours	ISO std2	**0.851**	**23.76**	**0.241**	**-0.85**	**0.836**	**22.72**	**0.255**	-0.95
	ISO std4	0.847	22.91	0.250		0.831	21.77	0.262	

each view i, pixel luminance $L_i(x, y)$ is computed using the ITU-R BT.601 definition:

$$L_i(x,y) = 0.299 \cdot R_i^{(r)}(x,y) + 0.587 \cdot R_i^{(g)}(x,y) + 0.114 \cdot R_i^{(b)}(x,y), \tag{7}$$

where $(x, y) \in \{1, \ldots, W\} \times \{1, \ldots, H\}$ are pixel coordinates, and $R_i^{(r)}(x, y)$, $R_i^{(g)}(x, y)$, $R_i^{(b)}(x, y) \in [0, 1]$ denote the normalized RGB components of the rendered image R_i. The per-view mean luminance $\bar{L}_i$ is then computed as:

$$\bar{L}_i = \frac{1}{H \cdot W} \sum_{x=1}^{W} \sum_{y=1}^{H} L_i(x,y), \tag{8}$$

where $\bar{L}_i \in [0, 1]$ represents the average scene brightness for view i. Finally, the global brightness consistency across all N views is quantified as:

$$\text{Std-Luminance} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\bar{L}_i - \frac{1}{N} \sum_{j=1}^{N} \bar{L}_j \right)^2}, \tag{9}$$

where $\frac{1}{N} \sum_{j=1}^{N} \bar{L}_j$ denotes the global mean luminance across all views. Lower Std-Luminance values indicate higher inter-view brightness consistency and more reliable exposure normalization.

D. Comparison with Dynamic Scene Models

We compare our approach with dynamic-scene models, including StreetGS[4], DeSIRE-GS[3], and OmniRe[1]. All baselines are trained using their publicly released implementations with hyperparameters strictly following the settings reported in their respective papers. Quantitative results are presented in 2. Across both training conditions (std2 and std4), our method consistently outperforms previous work.

In reconstruction, our approach achieves the highest ΔPSNR among all methods, demonstrating robustness even compared to dynamic-scene models that explicitly track moving objects. For novel-view synthesis (NVS), P2GS achieves a lower ΔPSNR than StreetGS and OmniRe. However, it is important to note that both methods exhibit noticeably degraded NVS quality. In the more challenging std4 setting, our method yields substantial improvements over OmniRe, achieving gains of 62.6% in SSIM, 97.2% in PSNR, and 67.5% in LPIPS.

These results highlight the strong robustness of P2GS under heterogeneous illumination, even when competing methods fail to maintain photometric consistency.

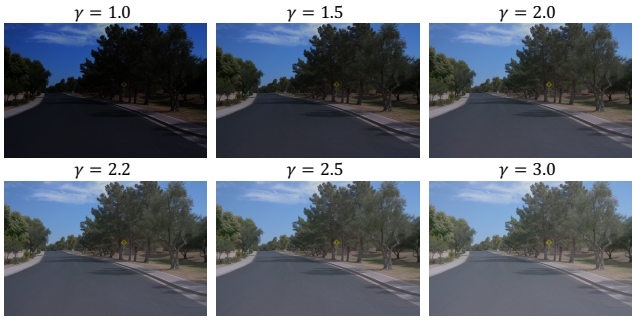

Figure 1. Gamma comparison on the Waymo Open Dataset.

E. Application

Our decoupling of exposure scale and tone-mapping enables rendering with arbitrary combinations of e and γ. The key advantage is that global illumination consistency is preserved while allowing the brightness of the entire scene to be adjusted. This capability is independent of camera-specific characteristics or the time of data acquisition. Such controllability can be partially leveraged to emulate challenging conditions such as strong sunlight or adverse weather offering potential applications for evaluating autonomous driving models.

Importantly, our method supports arbitrary e and γ val-

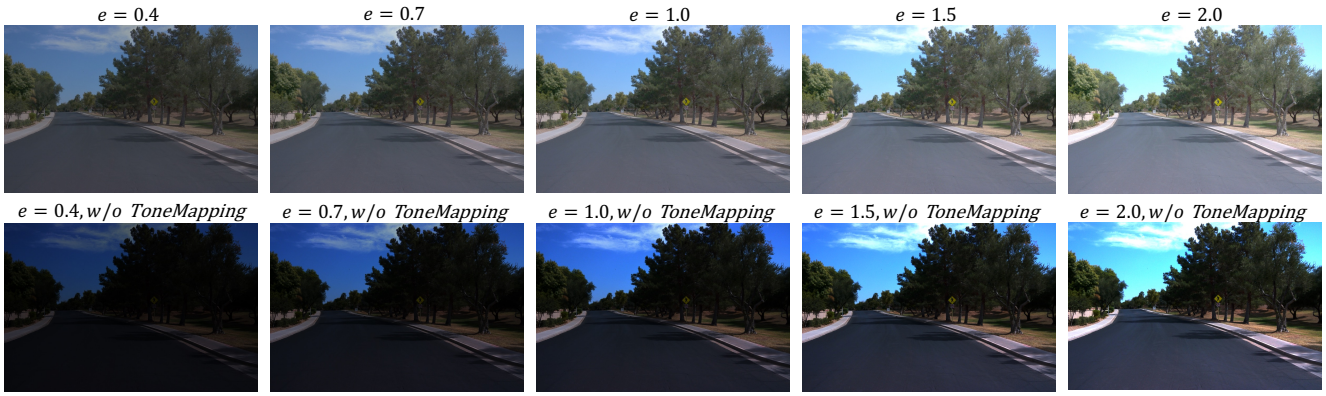

Figure 2. Exposure scale and Tone-Mapping comparison on the Waymo Open Dataset.

ues **without compromising rendering speed**, ensuring scalability. Fig.1 illustrates rendering results under different γ values, while Fig.2 compares renderings produced with varying e values with and without tone-mapping.

References

[1] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In The Thirteenth International Conference on Learning Representations, 2025. 3

[2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning, pages 1–16, 2017. 1

[3] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025. 3

[4] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In ECCV, 2024. 3