# Generalizable Sparse-View 3D Reconstruction from Unconstrained Images

## Supplementary Material

In this supplementary material, we elaborate our dataset details (Sec. A) and our model detailts (Sec. B). We also strongly encourage to view our supplementary HTML for full video results.

## A. Dataset Details

### A.1. Training Dataset

For training GenWildSplat, we constructed a large-scale synthetic dataset derived from the DL3DV [6] dataset. Initially, we subsampled 2,000 scenes from DL3DV, focusing specifically on outdoor environments, which resulted in approximately 700 scenes suitable for our goal. These scenes were then processed using our synthetic data generation pipeline to enhance appearance diversity and robustness. In particular, we employed DiffusionRenderer [4] in a classifier-free guidance setting to randomly relight each image, thereby producing a wide range of lighting conditions. To optimize computational efficiency, we executed the inverse rendering step for only one iteration, as preliminary experiments indicated negligible differences compared to performing 15 iterations. The forward rendering step, however, was carried out for 15 iterations to ensure high-fidelity reconstructions of relit appearances. Due to the substantial computational cost of this process, approximately 30–45 minutes per scene, the relighting procedure was applied offline and limited to the 700 selected outdoor scenes.

We incorporated synthetic occlusions during training to mimic the in-the-wild images. Using a pretrained segmentation model [1] on the COCO [5] dataset, we created a comprehensive bank of objects that could serve as occluders. During training, we randomly sampled between 2 to 10 objects from this bank and positioned them in the lower half of the image, mimicing the empirical distribution of occlusions in real-world scenes. Occlusions were added on-the-fly, with corresponding occlusion masks generated in real time and used to supervise the model. This combination of relighting and occlusion augmentation enabled GenWildSplat to learn robust appearance and geometry representations under sparse inputs, diverse illumination, and realistic occlusions, ensuring strong generalization to unseen outdoor scenes.

### A.2. Evaluation Dataset

For evaluation, we carefully curated a set of testing scenes from the MegaScenes dataset to reflect realistic sparsity and complexity. Specifically, we selected scenes containing fewer than 20–25 images, deliberately ensuring that the dataset mimics the sparse viewpoint coverage and diverse lighting conditions encountered in real-world captures. Unlike prior works such as MS-GS [2] and SparseGS-W [3], which simulate sparsity by artificially discarding images from densely captured scenes like PhotoTourism [7], our selection prioritizes authenticity. By using scenes that are naturally sparse, we ensure that the evaluation closely represents practical scenarios where acquiring dense multi-view captures is infeasible.

Furthermore, the chosen scenes exhibit a range of illumination variations, including both subtle and extreme lighting changes, as well as moderate to high levels of transient occlusions. These characteristics create a challenging environment for novel-view synthesis and relighting tasks, providing a rigorous benchmark for assessing the performance and generalization capability of GenWildSplat and in total, we curated 20 scenes.

## B. Additional Architecture Details

### B.1. DPT Backbone

We adopt a Dense Prediction Transformer (DPT) backbone for predicting depth, camera parameters, and Gaussian scene representations. The DPT encoder generates multi-resolution feature maps that feed into three task-specific heads: (i) a depth head producing a dense depth map via convolutional fusion; (ii) a camera head estimating global pose and intrinsics using pooled high-level features followed by an MLP; and (iii) a Gaussian head that outputs per-Gaussian parameters (mean positions, anisotropic covariances, and feature vectors). This separation enables accurate spatial predictions while capturing global camera information in a compact latent representation.

### B.2. Light Encoder

The light encoder uses only the encoder portion of a U-Net style autoencoder built from residual convolutional blocks. Each block contains two convolutional layers, group normalization, and a nonlinear activation. The encoder has six resolution levels with block counts $[1, 2, 2, 4, 4, 4]$, starting at $256 \times 256$ resolution and halving at each level. Latent channel widths are $[32, 64, 128, 128, 256, 512]$. Extrinsic lighting features are extracted from the bottleneck using multiple MLP layers followed by spatial averaging, producing a 16-dimensional vector that captures low-frequency, global lighting. No intrinsic features are used.

### B.3. Segmentation

Occlusion masks are generated using the YOLOv8x.seg model trained on COCO classes. The selected COCO objects: person, bicycle, car, motorcycle, bus, train, truck, boat, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, suitcase, chair, keyboard, book. These objects are used to create an occluder bank for online augmentation during training.

### B.4. Appearance Adapter

The Appearance Adapter is a five-layer MLP mapping the concatenated 16-dimensional extrinsic light code and 75-dimensional per-Gaussian conditioning vector to 75 spherical-harmonic (SH) coefficients. Hidden layer sizes are $[256, 512, 512, 256]$, with nonlinear activations after each layer and a linear output layer. This structure allows smooth low-frequency appearance modulation while retaining sufficient capacity to predict per-Gaussian SH lighting coefficients for rendering.

## References


[1] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 1

[2] Deming Li, Kaiwen Jiang, Yutao Tang, Ravi Ramamoorthi, Rama Chellappa, and Cheng Peng. Ms-gs: Multi-appearance sparse-view 3d gaussian splatting in the wild. *arXiv preprint arXiv:2509.15548*, 2025. 1

[3] Yiqing Li, Xuan Wang, Jiawei Wu, Yikun Ma, and Zhi Jin. Sparsegs-w: Sparse-view 3d gaussian splatting in the wild with generative priors. *arXiv preprint arXiv:2503.19452*, 2025. 1

[4] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, 2025. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[6] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 1

[7] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*. 2006. 1