WildRelight: A Real-World Dataset and Benchmark for Single-Image Relighting

Lezhong Wang Mehmet Onurcan Kaya Siavash Arjomand Bigdeli Jeppe Revall Frisvad
Technical University of Denmark
{lewa, monka, sarbi, jerf}@dtu.dk, lezhong.wang@inria.fr

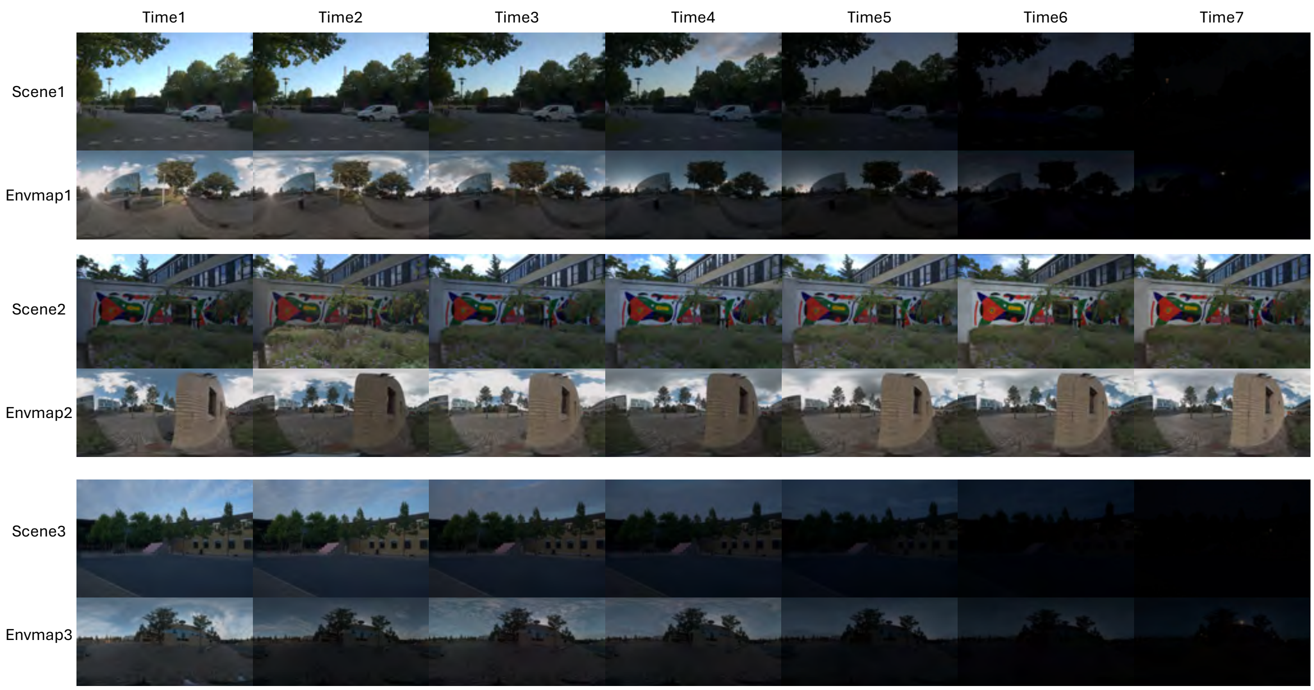

Figure 1. Example image and illumination pairs from our *WildRelight* dataset

Abstract

Recent single-image relighting methods, powered by advanced generative models, have achieved impressive photorealism on synthetic benchmarks. However, their effectiveness in the complex visual landscape of the real world remains largely unverified. A critical gap exists, as current real-world datasets are typically designed for multi-view reconstruction and fail to address the unique challenges of single-image relighting. To bridge this synthetic-to-real gap, we introduce ***WildRelight****, the first in-the-wild dataset specifically created for evaluating and training single-image relighting models. WildRelight features a diverse collection of high resolution outdoor scenes, each paired with a spatially aligned, high-dynamic-range environment map that serves as the ground truth illumination. Our meticulous capture process, using a custom rig to precisely co-locate the nodal points of a primary camera and a 360° camera, ensures the physical accuracy of the lighting data. Using our dataset, we conduct the first comprehensive benchmark of state of the art methods. Our results quantify a significant domain gap, revealing poor zero shot performance for models trained only on synthetic data. Furthermore, we demonstrate that finetuning on WildRelight substantially improves model performance, validating its effectiveness for domain adaptation. The dataset and benchmark will be made publicly available to foster the next wave of research in robust, physically-grounded relighting.*

1. Introduction

Manipulating the illumination within a single photograph is a long standing goal in computer vision and graphics, with profound applications in computational photography, filmmaking, and augmented reality [2, 7, 13, 19, 31]. Recently, the field has seen a dramatic leap forward, propelled by the expressive power of deep generative models [16, 17, 21]. State of the art methods can now decompose a single image into its intrinsic components (albedo, geometry, illumination) and rerender the scene under novel lighting with stunning photorealism [8, 21, 36].

Despite this remarkable progress, a critical question remains unanswered: how well do these models perform outside the sanitized confines of synthetic data? The training and, more importantly, the quantitative evaluation of most inverse rendering models are predominantly conducted on synthetic datasets [15, 16, 36, 38]. While invaluable for development, synthetic data often fails to capture the intricate complexities of real-world light transport, such as subtle atmospheric scattering, complex indirect illumination, and the rich, non-ideal material properties of natural surfaces. This creates a significant domain gap, where a model's impressive performance on synthetic benchmarks may not translate to practical, real-world applications.

To address this critical need, we introduce **WildRelight**. We position WildRelight as a crucial supplement to the existing landscape of invaluable datasets. Our work specifically targets the well-documented and persistent gap in single-image, in-the-wild evaluation. This domain, characterized by the complex and dynamic nature of natural illumination, presents unique acquisition challenges and is fundamentally underrepresented. Therefore, our objective is not to replicate the sheer scale of synthetic benchmarks, but rather to provide a meticulously curated, high-fidelity dataset that offers a physically-accurate benchmark for these specific real-world challenges. We systematically captured a diverse array of outdoor scenes at various times of the day, from the golden hour to harsh midday sun, to encompass a wide spectrum of natural illumination (see examples in Fig. 1). Our core technical contribution is a meticulous data acquisition pipeline. For each high resolution image captured with a primary camera (Sony A7), we simultaneously recorded a full 360° High Dynamic Range (HDR) environment map with a panoramic camera (Insta360 Pro 2). Crucially, we designed and utilized a custom built rig to precisely colocate the nodal points of both cameras. This spatial alignment is vital, as it ensures the captured HDR environment map is a accurate ground truth representation of the incident illumination for its corresponding single-view image.

We provide a detailed comparison of WildRelight against existing real-world multi-illumination datasets in Table 1. These datasets can be broadly categorized. First, controlled laboratory datasets, such as OpenIllumination [17] and ReNe [31], use light stages or robotics to capture high fidelity object data, often with precise One-Light-at-a-Time (OLAT) sources and perfect viewpoint alignment. While invaluable for BRDF estimation, they lack the geometric complexity and rich, full spectrum illumination of "in-the-wild" scenes. Second, multi-view "in-the-wild" datasets, like NeRF-OSR [27] and Objects With Lighting [32], capture real natural lighting but are designed for multi-view 3D reconstruction. Their primary data consists of moving camera trajectories, meaning they lack the static, cross-illumination viewpoint alignment necessary for evaluating single-image relighting. Finally, other single-view datasets, such as Murmann et al. [24] or LSMI [1], are typically captured indoors with simple spotlights, lack HDR, do not provide GT environment maps, and often lack strict viewpoint alignment. In contrast, WildRelight fills a critical, unaddressed gap by being the first dataset to combine all the necessary features for our task: (1) a single-view, static camera setup ensuring strict viewpoint alignment, (2) diverse, "in-the-wild" outdoor scenes, and (3) high-fidelity, spatially aligned HDR environment maps for every image.

2. Related Work

Single-Image Inverse Rendering. Inverse rendering, the ill-posed problem of decomposing a single image into its intrinsic components like albedo, geometry, and illumination [13, 14, 19, 22, 28, 35, 37, 39, 43], has seen a recent resurgence. Driven by advances in deep generative models, particularly latent diffusion models [16, 21, 36], modern methods are demonstrating remarkable, and often photorealistic, capabilities in single-image relighting. These models learn strong priors about the physical world, enabling them to plausibly rerender a scene under novel lighting. However, this progress is hampered by a significant evaluation gap. Lacking a real-world benchmark with ground truth (GT) illumination, these methods are developed and evaluated almost exclusively on synthetic datasets [13, 16, 18, 36, 40]. While synthetic data provides perfect GT, it inherently fails to capture the full complexity of real world phenomena such as the subtle interplay of atmospheric scattering, high frequency shadows from complex foliage, and the rich spectral properties of natural materials. This "sim-to-real" gap means that a model's performance on synthetic data is a poor predictor of its efficacy "in-the-wild". Our work is the first to directly address this critical need by providing a large scale, real-world benchmark specifically for single-image relighting.

Controlled Laboratory Datasets. To acquire GT data from real-world objects, one dominant line of work relies on highly controlled laboratory environments. This ranges from classical photometric stereo setups with sparse, calibrated lights [17, 25, 29, 30, 42] to modern light stages. These sophisticated systems, such as OpenIllumination [17], OpenSubstance [25], and RelightMyNeRF [31], utilize hundreds of controllable LEDs, multiple synchronized cameras, or precise robotics to capture an object's response to illumination. They provide extremely high fidelity data, often including

Table 1. Comparison with existing real-world, multi-illumination datasets. Our dataset is the first to provide a large-scale, single-view benchmark for in-the-wild outdoor scenes, featuring HDR GT illumination. Critically, it ensures strict Viewpoint Alignment across all captured illuminations, enabling direct, pixel-aligned evaluation of relighting methods.

Dataset	**Type**	**Illumination Format**	**Light Source**	**# Illuminations**	**Purpose**	**HDR**	**Alignment**	**# Scenes**
NeRF-OSR [27]	Outdoor	Environment map	Natural	5 per scene	Multi-view	✗	✗	9
OpenIllumination [17]	Objects	OLAT + pattern	Light stage	142+13	Multi-view	✓	✓	64
Murmann et al. [24]	Indoor	Chrome ball	Spotlight	25 per scene	Single-view	✓	✗	1000
Objects With Lighting [32]	Objects	Environment map	Natural	4 per scene	Multi-view	✗	✗	8
LSMI [1]	Indoor	Not Available	Spotlight	1–3 per scene	Single-view	✗	✗	2700
ReNe [31]	Objects	OLAT	Light stage	40 OLATs	Multi-view	✗	✓	20
Ours	**Outdoor**	**Environment map**	**Natural**	**5–7 per scene**	**Single-view**	✓	✓	**30**

precise 3D geometry from scanners and OLAT measurements, which serve as the definitive GT for material BRDF acquisition. The fundamental limitation of these datasets, however, is twofold: (1) Scope: They are constrained to small scale, isolated objects that can fit inside the capture apparatus, making them unsuitable for studying large scale scenes, architecture, or landscapes. (2) Illumination: The illumination, while dense, is composed of discrete, artificial LEDs, which cannot fully replicate the continuous, full spectrum, and High Dynamic Range (HDR) nature of global illumination in an outdoor environment (i.e., the sun, sky, and indirect bounces from the entire surroundings). WildRelight, in contrast, sacrifices per light decomposition to capture the full complexity of natural, "in-the-wild" illumination for large scale scenes.

Neural Inverse Rendering and Relighting. The advent of NeRF [23] has revolutionized 3D reconstruction and sparked a new wave of neural inverse rendering methods. These approaches [3, 37, 38, 41] extend the NeRF framework to decompose a scene into its intrinsic properties from multiple views. They successfully disentangle geometry, materials, and illumination, allowing for high quality novel view synthesis and relighting. However, these methods either require controlled laboratory capture with known lighting [37] or attempt to estimate a single, static illumination (e.g., an environment map) from the multi-view images themselves [38]. While impressive, they do not address the challenge of relighting under diverse, measured real-world illumination conditions.

Multi-View In-the-Wild Datasets. To move neural inverse rendering outdoors, recent datasets have successfully captured large scale scenes under time-varying natural light. The NeRF-OSR [27] dataset captured time lapse videos of buildings to train a NeRF model that can relight the scene by interpolating the learned illuminations. Similarly, the Objects With Lighting (OWL) [32] and Stanford-ORB [12] datasets capture objects from multiple viewpoints under several distinct natural lighting conditions, providing GT envmaps for each. The fundamental design goal of these datasets is to provide *multi-view* data for 3D reconstruction to build a relightable 3D model of the scene. Consequently, their data structure and benchmarks are built around training and testing multi-view reconstruction algorithms (e.g., NeRF, 3DGS) [11, 31, 38]. They are unsuitable for the distinct and equally challenging task of *single-image* relighting, which assumes only one input photograph and no multi-view correspondence.

Classical and Single-View Datasets. The final category of related work consists of datasets that, like ours, are captured from a fixed, single viewpoint. However, these datasets were designed for different tasks and have critical limitations. Classical photometric stereo datasets [29] capture objects under a sparse set of discrete point lights, typically in a darkroom, which is far removed from real world illumination. Other datasets designed for image processing tasks, such as the flash/no-flash dataset [1, 24], provide only two simple, low dynamic range illumination conditions. While these datasets are valuable for their intended purpose, none provide the necessary data to evaluate modern, physically-based relighting algorithms: a large scale collection of high resolution, HDR images of complex scenes, with each image rigorously paired with a spatially aligned, HDR GT envmap. WildRelight is the first dataset to fill this crucial void, bridging the gap between single-image generative models and the physical reality of our world.

3. Dataset Collection and Curation

Following the principles of rigorous and reproducible data acquisition, we introduce the WildRelight dataset, specifically designed for single-image relighting under real-world, dynamic illumination. Our collection protocol is meticulously crafted to ensure high fidelity in both scene capture and environmental lighting measurement.

3.1. Dataset Overview

The WildRelight dataset contains 30 distinct scenes (see Fig. 2 for examples). To capture the full spectrum of natural

Figure 2. We selected some example scenes from the dataset. We meticulously curated the dataset to represent a diverse range of challenging scenarios. The collected scenes encompass complex environmental conditions including tree structures, transparent glass surfaces, and reflective glass materials. For each scene, temporal variations are captured through photographs taken at different time intervals, accompanied by corresponding environmental maps that provide spatial lighting information.

lighting changes, each scene was recorded from a fixed camera position under 5 to 7 different illumination conditions. This approach provides a challenging and realistic benchmark for evaluating single-image relighting algorithms. The core of our data collection strategy is to sample lighting at different times of day, capturing both the subtle shifts of afternoon light and the rapid, dramatic changes during sunset. This temporal variation provides a unique and challenging benchmark for evaluating the robustness of relighting algorithms to the complex, continuous changes of real-world illumination, a domain largely unexplored by existing datasets.

3.2. Data Acquisition Protocol

Our data acquisition relies on a dual-camera system: a Sony A7 is used to capture the high resolution scene, while an Insta360 Pro 2 simultaneously records the full 360-degree environmental illumination map (envmap), see Fig. 3.

Spatio-Temporal Alignment. A critical aspect for ensuring the precise correspondence between a captured image and its lighting environment is the spatial alignment of the two cameras. We guarantee this by co-locating the optical center of the Insta360 Pro 2 with the nodal point (entrance pupil) of the Sony A7 lens. This setup ensures that the captured envmap accurately represents the complete incident light field at the exact vantage point of the scene camera. To minimize temporal discrepancies caused by changing natural light, we streamlined our capture process. For each data point, we first captured the envmap with the Insta360 Pro 2, followed immediately by the scene capture with the Sony A7. This swap was typically accomplished in under one minute, and in rare cases, up to two minutes to avoid transient scene elements like pedestrians.

Nodal Point Alignment. Achieving this precise co-location required us to first empirically determine the "no-parallax point" (entrance pupil) of the specific Sony A7 lens and focal length used. We employed a standard panoramic photography methodology: camera was mounted on a specialized panoramic head, and its forward-backward position was iteratively adjusted. The correct position was identified when rotating (panning) the camera caused zero observable parallax shift between two aligned, depth separated vertical objects (e.g., a near lamppost and a distant utility pole). Once this pivot point was locked, the Insta360 Pro 2 was mounted on a vertical rig, aligning its optical center with this exact vertical axis. A detailed guide of this calibration procedure is provided in the supplementary material.

Temporal Sampling Strategy. Recognizing that natural light evolves at a non-uniform rate, we adopted a variable temporal sampling strategy. During the afternoon (11:00 - 17:00), when illumination changes are gradual, data was

Figure 3. Capture settings. Both the Insta360 Pro2 and Sony A7M2 are mounted on a rail system to enable front-to-back adjustments for precise alignment with the nodal point.

collected every 45 to 60 minutes. In contrast, during the pre-sunset period, when light intensity and chromaticity shift dramatically, we increased the capture frequency to every 10 to 15 minutes.

Camera Parameters. To maintain consistency, we used fixed camera settings where possible. For the Sony A7, we set the ISO to 100, aperture to f/4, and white balance to 5000K. A 40mm focal length was chosen for its natural field of view and minimal distortion, making it ideal for a general purpose dataset. The shutter speed was typically set to 1/500s in the afternoon and 1/100s in the evening. For scenes with extremely high dynamic range, we captured bracketed exposures by adjusting the shutter speed. The Insta360 Pro 2 was configured with matching ISO (100) and white balance (5000K), and its shutter speed was synchronized with the Sony A7. All images were captured in the 16 bit linear RAW format.

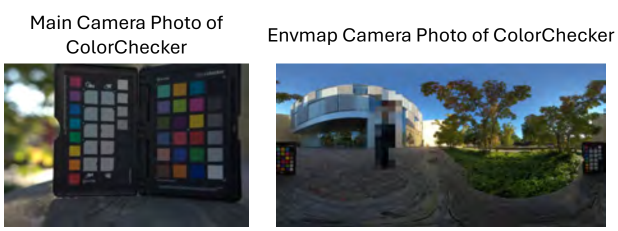

Figure 4. Photos of Xrite ColorChecker are used to calibrate color between main camera Sony A7 and envmap camera Insta360 Pro2

3.3. Color and Radiometric Calibration

To create a radiometrically accurate and color consistent dataset, we implemented a careful calibration and processing pipeline.

Cross-Camera Color Calibration. To correct for the inherent color discrepancies between the Sony and Insta360 sensors, we performed a one-time color calibration. Under diffuse daylight (without direct illumination), we photographed an X-Rite ColorChecker target simultaneously

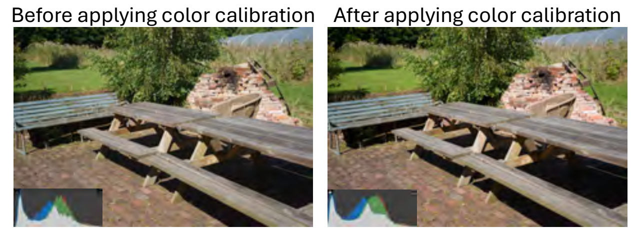

Figure 5. Comparison examples of color calibration effects. Due to the inherently sophisticated color science in professional grade products such as the Sony A7 and Insta360 Pro2, visual differences between pre- and post-color calibration images are negligible. To quantitatively analyze the calibration effects, a histogram was incorporated into the lower left corner of the image. While visual inspection reveals minimal variation, the histogram demonstrates discernible differences in pixel distribution before and after color calibration.

with both cameras (see Fig. 4). Using the ColorChecker's accompanying software, we generated custom color profiles for each camera. These profiles were then applied during post-processing to all images in the dataset, unifying the color rendition across both capture devices. As seen in Fig. 5, the visual effect of the color calibration is minute.

Linear HDR Synthesis. Our entire pipeline operates in a linear color space, starting from the RAW sensor data. By using RAW files, we circumvent the need for estimating a non-linear camera response function (CRF), a common source of error when working with processed image formats like JPEG.

We selectively perform HDR synthesis: multiple bracketed exposures are merged into a single HDR image only when the scene's dynamic range exceeds the exposure latitude of a single capture (i.e., when pixel values saturate the CMOS sensor's upper limit). If a single RAW image successfully records the scene without clipping, this merging step is bypassed. When merging is required, given the linearity of the RAW data, the radiance L_i from a single exposure i is directly proportional to its recorded linear pixel value Z_i (scaled to $[0, 1]$) divided by the exposure time Δt_i.

To robustly merge the bracketed set, we compute a weighted average of the radiance from all exposures. This approach is crucial for handling regions that are underexposed (noisy) in one shot or overexposed (clipped) in another. We employ a triangular weighting function [6] $w(Z_i)$ that assigns maximum weight to mid-range pixels (peaking at $Z_i = 0.5$) and zero weight to pixels that are fully underexposed ($Z_i = 0.0$) or saturated ($Z_i = 1.0$). The final HDR radiance L for each pixel is calculated as the weighted sum of the individual radiance contributions, normalized by the sum of their weights:

$$L = \frac{\sum_i w(Z_i) \cdot L_i}{\sum_i w(Z_i)} = \frac{\sum_i w(Z_i) \cdot (Z_i/\Delta t_i)}{\sum_i w(Z_i)}$$

where the weighting function is $w(z) = 1.0 - |2z - 1.0|$.

The resulting HDR image is stored in the linear, floating-point EXR format, using high-precision float64 accumulation to ensure accuracy. This process ensures that the final images and envmaps are a faithful, linear representation of the real-world scene radiance, which is crucial for physically-based rendering and relighting tasks.

4. Baseline Experiments

To validate the challenges posed by our new *WildRelight* dataset and to establish a performance benchmark, we evaluated several state of the art single-image relighting methods. These experiments are designed to quantify the domain gap between existing models, which are predominantly trained on synthetic data, and the complexities of real-world scenes. Furthermore, we demonstrate the value of our dataset by finetuning a representative model and measuring the resulting performance gains.

4.1. Methods

Dataset and Metrics All experiments are conducted using our *WildRelight* dataset, which is partitioned into a 21-scene training set, a 4-scene validation set, and a 5-scene hold-out test set. We evaluate performance using three standard image quality metrics: Peak Signal to Noise Ratio (PSNR), the Structural Similarity Index (SSIM), and the Learned Perceptual Image Patch Similarity (LPIPS).

Baseline Models We selected three representative methods that support single-image relighting: RGB↔X [36], DiffusionRenderer [16], and Materialist [34].

Evaluation Protocol (Diffusion-Based) The RGB↔X and DiffusionRenderer models share a similar architecture. They first employ a model trained on synthetic data to decompose a single input image into G-buffer components (albedo, roughness, metallic, normal). Then, they use a diffusion-based model as a neural renderer to relight the scene using these G-buffers and a novel envmap.

For our evaluation, we use the 'time 0' photograph from each scene as the input to their publicly available, pretrained decomposition models. We then use the resulting G-buffers, along with the GT envmaps from the other time lapses in that scene, as input to their respective rerendering diffusion models. The rendered images are then compared against the corresponding GT photographs. Since the original GT photographs are provided in HDR linear color space, we first process them with tone mapping [26] followed by gamma correction ($\gamma \approx 2.2$) before computing the comparison metrics against the model outputs.

Evaluation Protocol (Optimization-Based) Materialist [34] utilizes a different, optimization-based approach. It first predicts an initial G-buffer set, similar to the other methods. However, it then refines these G-buffers by optimizing them within a physical renderer (Mitsuba [10]) to match the input image. Given that our dataset provides high fidelity GT envmaps, we adapt their protocol to isolate material estimation. Instead of jointly optimizing for a coarse envmap, we provide the known GT envmap directly to the Mitsuba renderer during optimization. This allows the model to focus solely on optimizing the G-buffers (albedo, roughness, metallic, normal) to be physically consistent with the known input image and its corresponding lighting. For each scene, we provide two GT envmaps to facilitate this optimization. The resulting, refined G-buffers are then used to relight the scene with the remaining, previously unseen GT envmaps from that scene.

Finetuning Protocol To demonstrate the utility of *WildRelight* for domain adaptation, we finetuned the DiffusionRenderer [16] model. Since the official training code was not publicly available, we implemented a custom finetuning pipeline based on its described architecture, which consists of a VAE, a dedicated environment encoder, and the core spatio temporal UNet.

For efficient adaptation, we froze the pre-trained weights of both the VAE and the environment encoder, isolating the finetuning process to the UNet. We employed Low-Rank Adaptation (LoRA) [9] to update the UNet's weights. Specifically, we applied LoRA adapters with rank $r = 8$ to the attention layers (to k, to q, to v, to out.0) and the main projection layers (proj in, proj out).

During training, model operates in the VAE's latent space. The G-buffer components (basecolor, normal, depth, roughness, metallic) are individually encoded by the frozen VAE, and their resulting latents are concatenated to form the spatial conditioning input for the UNet. Concurrently, the GT HDR envmap is processed into three representations (LDR, log space, and a panoramic normal map). These are encoded by VAE, concatenated, passed through the frozen environment encoder to produce multi scale features, which are injected into UNet via cross attention.

The model was trained to predict the noise added to the target image's latent representation, following the DDPM-Scheduler sampling, and optimized using an L_2 (MSE) loss. We used the AdamW optimizer [20] with a learning rate of 1×10^{-4} and a linear warmup schedule. The model was trained on our 21-scene training set for 48 hours on a single NVIDIA H100 GPU. The final checkpoint was selected

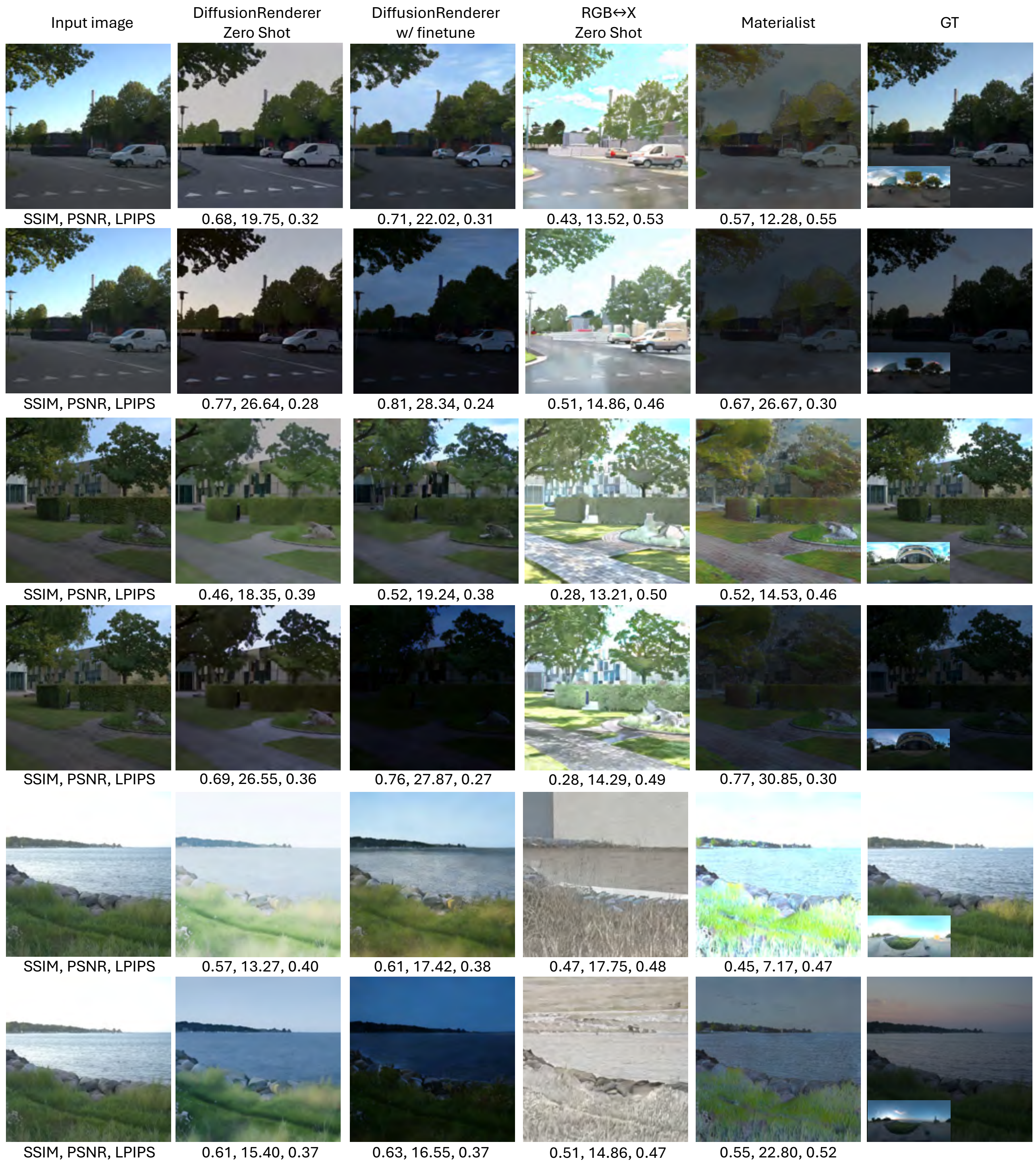

Figure 6. Qualitative comparison of different methods on the test set relighting. The numerical values displayed beneath the image correspond to the following metrics: SSIM, PSNR, and LPIPS. In zero shot scenarios, both DiffusionRenderer [16] and RGB↔X [36] struggle to accurately render image brightness. After finetuning, DiffusionRenderer demonstrates improved alignment with GT in the test set, achieving more accurate brightness reproduction. Materialist [34], leveraging an optimization-based approach combined with a physical renderer, though exhibits good brightness representation in rendered images, its reliance on geometric accuracy and the inherent limitations of physical renderers lead to unrealistic outputs compared to Diffusion Model-based methods. The envmap used for relighting is in the GT column (lower left).

Table 2. Performance of pre-trained baseline models on the *WildRelight* dataset (Overall metrics). We report two distinct experimental setups. The first group assesses the zero-shot performance of neural relighting methods. The second group evaluates an optimization-based method provided with GT illumination.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Zero-Shot Neural Relight			
RGB↔X [36]	15.87	0.4507	0.4917
DiffusionRenderer [16]	22.81	0.6218	0.3927
Opt w/ GT Illumination			
Materialist [8]	24.19	0.5819	0.3639

based on the best performance on the 4-scene validation set, and all final metrics are reported on the 5-scene test set.

Application: Instance-Specific Adaptation. To further demonstrate the unique utility of the *WildRelight* dataset, we propose a novel algorithmic framework designed explicitly to bridge the synthetic-to-real domain gap. This framework integrates Physics-Guided Diffusion Posterior Sampling (DPS) and Sampling-Aware Test-Time Adaptation (TTA). By explicitly leveraging the instance-specific, temporal illumination variations captured in our dataset as self-supervised constraints, this approach calibrates synthetic-trained generative models to real-world physical dynamics without requiring dataset-scale retraining. Detailed theoretical formulations, implementation specifics, and comprehensive evaluations regarding this methodology are provided in the Supplementary Material.

4.2. Results

The quantitative results of our baseline experiments are presented in Table 2 and Table 3. Some qualitative comparisons are available in Fig. 6.

Table 2 compares the zero shot performance of the three pre-trained models, aggregated over all splits of our dataset. As anticipated, the diffusion-based models (RGB↔X [36] and DiffusionRenderer [16]), which were trained exclusively on synthetic data, perform poorly on our real world images. This highlights a significant domain gap, with both methods achieving low PSNR and SSIM scores. Materialist [34] achieves better average metrics results. This is expected, as our protocol for this method provides the GT envmap during its optimization step, allowing it to produce more accurate, physically-based G-buffers. Additionally, the method's reliance on a physical renderer enables exceptional performance in low-light conditions, which contributes to the elevated average score. This result validates the efficacy of its optimization pipeline when illumination is known, but it is not a direct comparison to the zero shot relighting task faced by the other models.

Table 3. Effect of finetuning DiffusionRenderer [16] on the *WildRelight* dataset. Metrics are reported on the testset.

Model	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Zero-shot	23.28	0.6165	0.3790
Finetuned	25.95	0.6687	0.3368

Table 3 illustrates the practical value of our dataset. By finetuning DiffusionRenderer on our training split, its performance on the unseen test set improves dramatically. The PSNR increases from 23.28 dB to 25.95 dB, and the SSIM score improves from 0.6165 to 0.6687. This significant boost demonstrates that *WildRelight* provides a valuable resource for adapting existing models to the complex and diverse appearances of real-world materials and illumination, effectively bridging the synthetic-to-real domain gap.

5. Conclusion and Future Work

In this work, we introduced **WildRelight**, a publicly available benchmark for single-image relighting designed to bridge the critical synthetic-to-real domain gap. Our dataset provides meticulously aligned pairs of single-view images and their corresponding GT, co-located HDR envmaps. We established baselines by quantifying the zero shot performance of existing models, revealing a significant domain gap. We demonstrated the dataset's practical value for domain adaptation, finetuning the DiffusionRenderer [16] model on our training split led to a significant performance improvement on the test set.

Our dataset is characterized by its inclusion of complex, real-world light transport phenomena, such as intricate refractions, specular reflections, and soft shadowing dynamics, that remain exceptionally difficult for current models to handle. While we have endeavored to cover diverse scenarios, we acknowledge several limitations that open avenues for future work. Certain conditions, such as post-precipitation scenes with evaporating water are challenging to capture without violating our static scene assumption. A compelling future direction is the capture of longitudinal seasonal changes, though this would introduce more extensive dynamic scene elements, requiring new solutions for data alignment. We hope that WildRelight will serve as a valuable and challenging benchmark, catalyzing future research toward models that are not only photorealistic but also physically robust to the complexities of our natural world.

References

[1] Yagiz Aksoy, Changil Kim, Petr Kellnhofer, Sylvain Paris, Mohamed Elgharib, Marc Pollefeys, and Wojciech Matusik. A dataset of flash and ambient illumination pairs from the

crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 634–649, 2018. 2, 3

[2] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2014. 2

[3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural reflectance decomposition from image collections. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 12684–12694, 2021. 3

[4] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[5] Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982. 1

[6] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, page 369–378, USA, 1997. ACM Press/Addison-Wesley Publishing Co. 5

[7] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry-and image-based approach. In *Proceedings of SIGGRAPH '96*, pages 11–20, 1996. 2

[8] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 19740–19750, 2023. 2, 8

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[10] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 6

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4): 139, 2023. 3

[12] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-ORB: a real-world 3D object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36:46938–46957, 2023. 3

[13] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. OpenRooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7199. IEEE, 2021. 2

[14] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *European Conference on Computer Vision (ECCV)*, pages 555–572. Springer, 2022. 2

[15] Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi Yang. PhyIR: Physics-based inverse rendering for panoramic indoor images. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 12713–12723, 2022. 2

[16] 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 *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 26069–26080, 2025. 2, 6, 7, 8, 1

[17] Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, et al. OpenIllumination: A multi-illumination dataset for inverse rendering evaluation on real objects. *Advances in Neural Information Processing Systems*, 36:36951–36962, 2023. 2, 3

[18] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. NeRO: Neural geometry and BRDF reconstruction of reflective objects from multiview images. *ACM Transactions on Graphics*, 42(4):114, 2023. 2

[19] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):129–141, 2015. 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 6

[21] Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Wang. IntrinsicDiffusion: Joint intrinsic layers from latent diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 74:1–74:11, 2024. 2

[22] Wojciech Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003. 2

[23] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[24] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 4080–4089, 2019. 2, 3

[25] Fan Pei, Jinchen Bai, Xiang Feng, Zoubin Bi, Kun Zhou, and Hongzhi Wu. Opensubstance: A high-quality measured dataset of multi-view and-lighting images and shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5221–5231, 2025. 2

[26] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002. 6

[27] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022. 2, 3

[28] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 8598–8607. IEEE, 2019. 2

[29] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 3707–3716, 2016. 2, 3

[30] Timo Teufel, Pulkit Gera, Xilong Zhou, Umar Iqbal, Pramod Rao, Jan Kautz, Vladislav Golyanik, and Christian Theobalt. HumanOLAT: A large-scale dataset for full-body human relighting and novel-view synthesis. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 29131–29141, 2025. 2

[31] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my NeRF: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 20762–20772, 2023. 2, 3

[32] Benjamin Ummenhofer, Sanskar Agrawal, Rene Sepúlveda, Yixing Lao, Kai Zhang, Tianhang Cheng, Stephan R. Richter, Shenlong Wang, and Germán Ros. Objects with lighting: A real-world dataset for evaluating reconstruction and rendering for object relighting. In *3DV*. IEEE, 2024. 2, 3

[33] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. *Rendering techniques*, 2007:18th, 2007. 1

[34] Lezhong Wang, Duc Minh Tran, Ruiqi Cui, Thomson TG, Anders Bjorholm Dahl, Siavash Arjomand Bigdeli, Jeppe Revall Frisvad, and Manmohan Chandraker. Materialist: Physically based editing using single-image inverse rendering. *arXiv preprint arXiv:2501.03717*, 2025. 6, 7, 8

[35] Renjiao Yi, Chenyang Zhu, and Kai Xu. Weakly-supervised single-view image relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8402–8411, 2023. 2

[36] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 75:1–75:11. ACM, 2024. 2, 6, 7, 8

[37] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical Gaussians for physics-based material editing and relighting. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 5453–5462, 2021. 2, 3

[38] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T. Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics*, 40(6):237, 2021. 2, 3

[39] Xianling Zhang, Nathan Tseng, Ameerah Syed, Rohan Bhasin, and Nikita Jaipuria. Simbar: Single image-based scene relighting for effective data augmentation for automated driving vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3718–3728, 2022. 2

[40] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 18643–18652, 2022. 2

[41] Xiaoming Zhao, Pratul Srinivasan, Dor Verbin, Keunhong Park, Ricardo Martin Brualla, and Philipp Henzler. Illuminerf: 3d relighting without inverse rendering. *Advances in Neural Information Processing Systems*, 37:42593–42617, 2024. 3

[42] Xilong Zhou, Jianchun Chen, Pramod Rao, Timo Teufel, Linjie Lyu, Tigran Minasian, Oleksandr Sotnychenko, Xiaoxiao Long, Marc Habermann, and Christian Theobalt. Olatverse: A large-scale real-world object dataset with precise lighting control. *arXiv preprint arXiv:2511.02483*, 2025. 2

[43] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. IRISformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 2822–2831. IEEE, 2022. 2

WildRelight: A Real-World Dataset and Benchmark for Single-Image Relighting

Supplementary Material

1. Physics-Guided Test-Time Adaptation

To illustrate how WildRelight's real-world supervision can be practically exploited, we design a reference framework that integrates physics-guided posterior sampling for inverse decomposition regularization, together with sampling-aware TTA to better align forward relighting dynamics.

Rather than aiming to introduce a fully optimized solution, this framework serves as a structured case study demonstrating how dataset-driven supervision can be incorporated at both inference and adaptation stages. The resulting bidirectional consistency between scene representation and image formation highlights the practical value of WildRelight, while leaving substantial room for future methodological improvements.

1.1. Physics-Guided Inverse Rendering via Diffusion Posterior Sampling

To enforce physical validity in G-buffer prediction without retraining, we introduce a physics-guided inference strategy based on Diffusion Posterior Sampling [4]. At diffusion timestep t, the estimated clean latent $\hat{x}_0$ is decoded by the VAE $D(\cdot)$ into intrinsic components and rendered under illumination L via a differentiable Cook–Torrance renderer $\mathcal{R}(\cdot)$ [5, 33]. We enforce consistency with the observed image I_{gt} via a measurement loss:

$$\mathcal{L}_{\text{render}} = \left\|\mathcal{R}\big(D(\hat{x}_0),\, L\big) - I_{\text{gt}}\right\|_2^2 . \tag{1}$$

Guided by the gradient $g_t = \nabla_{x_t}\mathcal{L}_{\text{render}}$, the DDIM sampling trajectory is refined as:

$$x_{t-1} \leftarrow x_{t-1} - \zeta_t\, g_t, \tag{2}$$

where ζ_t is the guidance strength. To ensure stability and efficiency during this optimization, we employ a split-sum approximation for image-based lighting while keeping the diffusion network parameters frozen.

1.2. Evaluation of Proposed Methodology

Evaluation Protocol. Unlike global finetuning which relies on fixed splits, our inference-time methodology (DPS + TTA) enables instance-specific adaptation. We evaluate this universality across all 30 scenes using a "leave-one-lighting-out" protocol: for N lightings, one is the test target while the other $N-1$ serve as self-supervised signals. This rigorously simulates real-world deployment without ground-truth supervision.

1.2.1. Ablation Study and Analysis

We dissect the contribution of each component by incrementally integrating them into the pre-trained DiffusionRenderer [16] baseline. The quantitative and qualitative results, averaged over the full dataset, are presented in Table 4 and Figure 7.

Baseline (Zero-Shot). The synthetic-trained baseline suffers a substantial sim-to-real gap (21.63 dB PSNR), failing to disentangle complex outdoor illumination and reflectance. This deficit quantifies the limitations of synthetic data rather than the architecture. We use this baseline to represent synthetic priors, demonstrating that real-world adaptation—enabled by our data—is critical for bridging this gap.

Impact of Physics-Guided DPS. Introducing Diffusion Posterior Sampling (+DPS) injects physical constraints, yielding consistent gains (+0.95 dB PSNR). Crucially, DPS acts as a geometric anchor: by enforcing rendering equation consistency, it prevents hallucinated shadows, ensuring generation adheres to the underlying scene structure.

Impact of Temporal TTA. TTA alone drastically minimizes photometric error (+2.5 dB PSNR) but slightly degrades LPIPS ($0.390 \rightarrow 0.392$). This reveals a photometric-perceptual trade-off in unconstrained self-supervision: optimization overfits pixel intensities at the cost of the natural image manifold, leading to artifacts that penalize perceptual metrics.

Synergy: Constrained Adaptation. The full pipeline (+DPS & TTA) resolves this trade-off, achieving the best overall performance (25.04 dB PSNR, 0.345 LPIPS). Here, DPS regularizes the TTA trajectory, ensuring the model adapts to scene-specific lighting dynamics without sacrificing physical plausibility or high-frequency details.

Comparison with Global Finetuning. Remarkably, our inference-time approach rivals fully supervised global finetuning reported in Sec. 4.2 (25.04 vs. 25.95 dB). This underscores our framework's efficiency: effectively bridging the sim-to-real gap with near-supervised performance without expensive retraining.

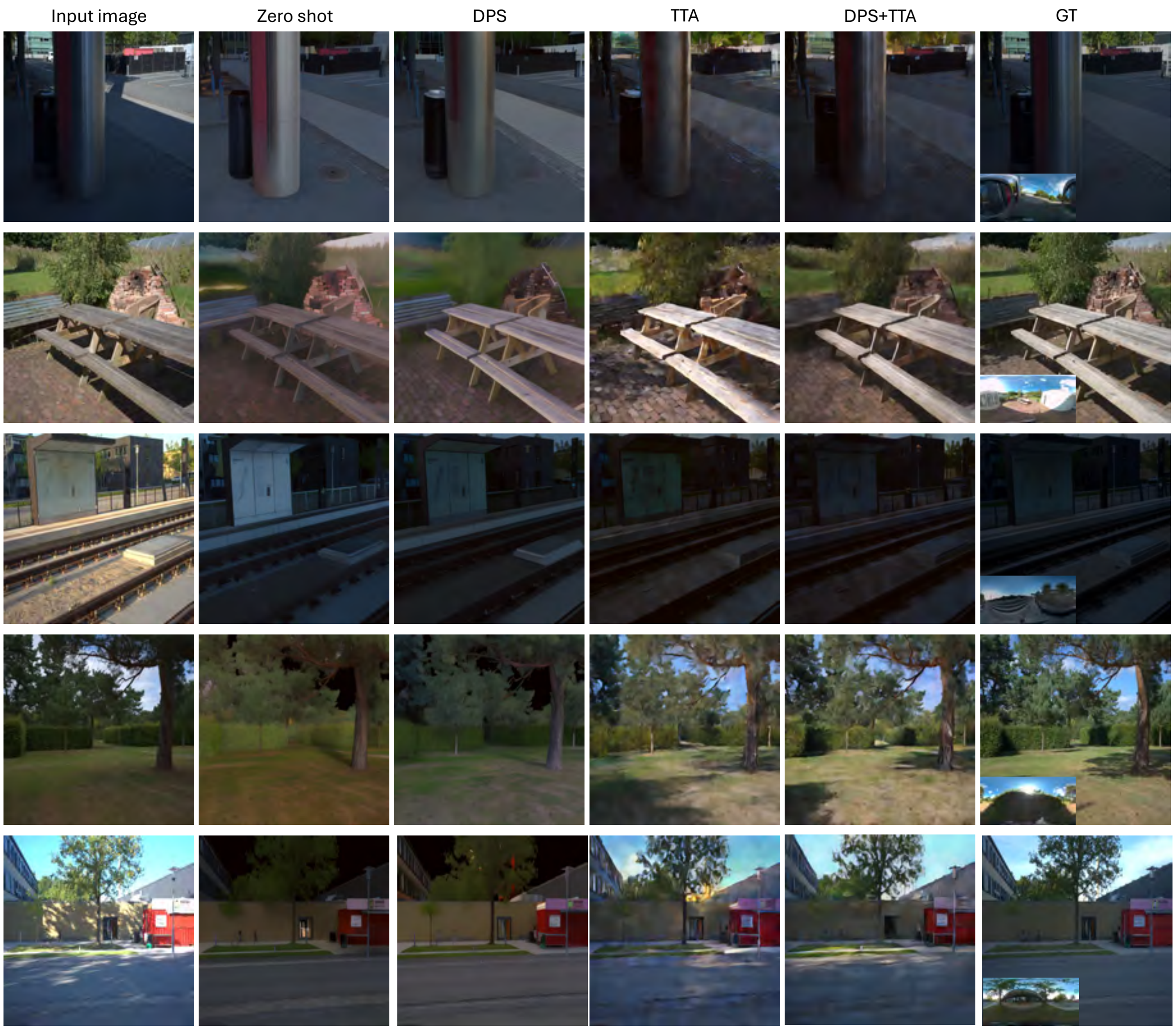

Figure 7. Qualitative Ablation Study Results. We visualize the results of our proposed framework on the *WildRelight* dataset. The envmap on bottom left of GT image correspond to target illuminations used for relighting.

Table 4. **Ablation Study of the Proposed Methodology.** Evaluated on the full 30-scene *WildRelight* dataset. While TTA alone significantly improves PSNR, it inadvertently degrades perceptual quality (higher LPIPS) due to unconstrained optimization overfitting to pixel-level intensities. Incorporating DPS restores physical consistency, and their combination yields the best overall performance, effectively bridging the sim-to-real gap without supervised training.

Configuration	Mechanism	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline (Pre-trained)	Zero-shot	21.63	0.6311	0.3901
+ DPS	Inference Prior	22.58	0.6578	0.3825
+ TTA	Optimization	24.10	0.6451	0.3923
+ DPS & TTA	**Constrained Adapt.**	**25.04**	**0.6829**	**0.3453**

2. Quantitative Validation of Illumination Alignment

We provide a rigorous quantitative validation based on metadata timestamp statistics and solar angular displacement analysis. This proves that the temporal gap in our acquisition pipeline results in physically negligible illumination misalignment for the task of relighting.

2.1. Temporal Synchronization Statistics

We extracted and analyzed the acquisition timestamps from the metadata of all image pairs (Sony A7 scene images and Insta360 envmaps). The distribution of the time delay Δt is summarized in Table 5.

Table 5. Statistics of Temporal Synchronization (Δt) between scene and environment map capture.

Metric	Value
Median Time Delta	38.00 seconds
Mean Time Delta	40.14 seconds
Max Time Delta	114.00 seconds

While the maximum delay is approximately 1.7 minutes, we demonstrate below that this is well within the tolerance for accurate outdoor lighting estimation.

2.2. Physical Error Analysis: Solar Angular Displacement

The primary source of directional illumination outdoors is the sun. The Earth rotates at approximately $15°$ per hour ($0.00417°$ per second). We calculate the angular displacement of the sun, θ_{shift}, caused by the capture delay:

$$\theta_{\text{shift}} = \Delta t \times 0.00417°/s \tag{3}$$

For our mean delay (40.14s):

$$\theta_{\text{mean}} \approx 0.17° \tag{4}$$

For our worst-case delay (114s):

$$\theta_{\text{max}} \approx 0.48° \tag{5}$$

To contextualize these values, the angular diameter of the sun is approximately $0.5°$. In the average case, the sun moves only $1/5$ of its own diameter, which is perceptually imperceptible in lighting effects. Even in the worst-case scenario, the displacement ($0.48°$) is still less than the angular size of the light source itself ($0.5°$).

2.3. Impact on Relighting Tasks

Modern single-image relighting algorithms typically operate on environment maps that are downsampled (e.g., to 256×128) to estimate lighting coefficients (like Spherical Harmonics) or to condition diffusion models.

At a width of 256 pixels, the horizontal angular resolution is $360°/256 \approx 1.4°$ per pixel.

$$\text{Pixel Shift}_{\text{max}} = \frac{0.48°}{1.4°/\text{pixel}} \approx 0.3 \text{ pixels} \tag{6}$$

Thus, even our maximum temporal delay results in a sub-pixel shift (0.3 pixels) in the effective resolution used by state-of-the-art relighting models. This quantitatively confirms that our dataset maintains high-quality illumination alignment suitable for training and evaluating inverse rendering methods.

We also plot the distribution of the absolute time differences (Δt) in Figure 8. As illustrated, the distribution is strongly skewed towards zero. The dominant peak in the first 3 bins (< 40s) corroborates that for the majority of our collected scenes, the temporal gap is minimal. The long tail extending to ~ 100s represents a small fraction of outlier cases. Combined with the solar angular displacement analysis, this distribution confirms that the impact of these delays on illumination alignment remains physically negligible for the purpose of relighting.

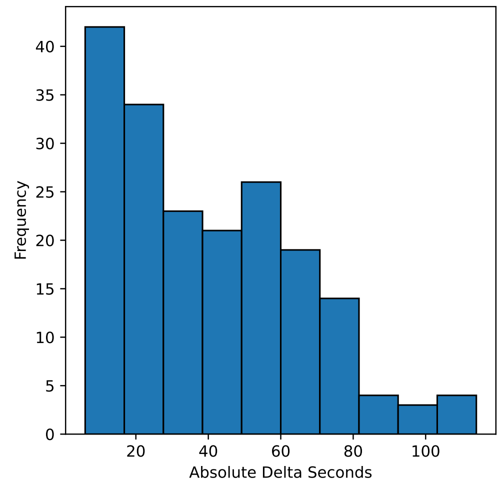

Figure 8. Distribution of capture time differences (Δt). The histogram shows the frequency of absolute time delays between the scene image and the environment map capture across the dataset. The distribution is heavily right-skewed, with the vast majority of samples having a delay of less than 20 seconds, confirming that large delays are rare.

3. Dynamic Element Masking

A significant challenge in our longitudinal, "in-the-wild" capture is the inevitable presence of dynamic scene elements, such as wind-blown foliage and moving cloud formations, despite a static camera rig. To preserve the photometric integrity of our GT images, we avoid computational alignment (e.g., warping) which would alter the pixel data. Instead, we provide meticulously hand-annotated binary masks for all non-static regions. This approach allows researchers to optionally exclude these dynamic areas during metric computation, thereby isolating the evaluation of relighting performance from artifacts caused by scene motion. After determining that automated methods were unreliable for our complex natural scenes, we developed a rigorous manual

Figure 9. Advantage of RAW format with HDR. When using HDR photo, the details of the photo are preserved. Therefore, by adjusting the exposure settings, the original colors of the photo can be accurately restored.

Figure 10. Example of Dynamic Scene Elements. As showed in the figure, the left and right images were captured at different times with a fixed camera position. Despite the static camera setup, subtle movements of the leaves occur due to external factors such as wind. To address this, we manually created masks for these dynamic regions, allowing researchers to determine whether to include them when computing metrics.

annotation pipeline based on pairwise temporal comparisons. The complete details of this pipeline, including our annotation interface and specific exclusion criteria (e.g., for water and reflections), are available in the supplementary materials.

4. Advantages of RAW-Based HDR Image

The data recorded by a camera sensor in RAW format exhibits a fundamentally linear relationship with the light intensity of the actual scene. This linearity is a core advantage in computational photography, particularly for HDR image synthesis. Leveraging this linear property of RAW data allows us to directly synthesize an HDR image by simply performing a linear combination of multiple exposures. This method circumvents the complex process of calculating and inverting the Camera Response Function (CRF). It significantly simplifies the HDR generation pipeline while improving fidelity.

In contrast to the RAW format, the conventional JPG format has inherent limitations for HDR applications. JPG files are typically stored and compressed using an 8-bit color depth, a process that results in a significant loss of color and luminance information. This loss of information is particularly severe under extreme lighting conditions, as illustrated in Figure 9. The key differences are:

1. **Shadow Detail and Color Fidelity:** In low-light environments, the RAW format, with its high bit depth (typically 12 or 14 bits), captures extensive detail in the dark regions. By increasing the exposure in post-processing, the original information can be recovered with minimal loss. Conversely, since this information is already discarded during the in-camera processing of a JPG image, attempting to boost its exposure does not recover meaningful detail and instead leads to severe color distortion, banding, and noise, as shown in Fig. 9.
2. **Highlight Information Retention:** In highlight regions, while both a RAW-based HDR image and a JPG image may appear as pure white on a Standard Dynamic Range (SDR) display due to exceeding the display's maximum brightness, the amount of information they contain is fundamentally different. The RAW data fully retains the color and tonal information within these bright areas. When the exposure is reduced in post-processing, the details in these "clipped" areas can be clearly recovered. As shown in the cloud section of Fig. 9, at an exposure compensation of -2EV, the RAW-based HDR image reveals subtle gradations in the clouds. The corresponding area in the JPG image, however, remains a flat white expanse because the information was permanently lost.

RAW-based HDR synthesis offers unparalleled advantages over the JPG format, both in the recovery of shadow detail and the preservation of highlight information.

Figure 11. Side-by-side camera setup, a non-aligned envmap camera will record direct sun light, but scene camera records a shadow.

4.1. Necessity of Strict Spatial Alignment

Precise co-location of the environment map camera (Insta360) and the scene camera (Sony A7) is a prerequisite for pixel-aligned evaluation.

Spatial Parallax is Fatal. A non-confocal setup (e.g., a standard side-by-side placement with a 10cm baseline) fundamentally alters occlusion relationships between the scene geometry and the light source. For instance, foliage that occludes the sun in the environment camera's view may not occlude it in the scene camera's view. This discrepancy creates "false" shadows in the Ground Truth, shadows that exist in the illumination map but not in the photograph (or vice versa). Such geometric inconsistencies render strict, pixel-aligned quantitative evaluation impossible.

5. Methodology for Determining the Nodal Point (No-Parallax Point)

In the domain of photographic composition, the elimination of parallax error is of paramount importance. Parallax manifests as an apparent displacement of foreground objects relative to the background as the camera is rotated. To mitigate this artifact, it is crucial to rotate the camera system around a specific pivot point. While colloquially referred to as the "nodal point," the technically precise term for this locus is the center of the entrance pupil. The entrance pupil represents the virtual image of the physical aperture stop as viewed through the front elements of the lens. It is the conjugate point through which all chief rays appear to pass before refraction. Rotating the camera and lens assembly around the center of the entrance pupil ensures that the perspective remains consistent across multiple exposures, thereby facilitating seamless image stitching.

The empirical determination of the entrance pupil's location, or the no-parallax point, is a foundational procedure in panoramic photography. The following methodology outlines a systematic approach to identifying this point.

5.1. Experimental Setup

The camera must be securely mounted on a panoramic head affixed to a stable tripod. This specialized head allows for the adjustment of the camera's position along the longitudinal axis of the lens. For the purpose of this procedure, a focal length of 40mm was selected. Two distinct, vertically-oriented objects, positioned at different distances from the camera, were chosen as reference points. A lamppost and a more distant utility pole served as suitable subjects. The camera's position was initially adjusted so that, when viewed through the camera's live-view display, the nearer reference object precisely occluded the more distant one at the center of the frame.

5.2. Procedure for Parallax Elimination

The core of the methodology lies in an iterative process of rotation and observation. With the reference objects aligned, the camera is panned horizontally by an angle of approximately 30 degrees to the left and then to the right. The relative position of the two reference objects is carefully observed during this rotation. If the nearer object appears to shift its position relative to the farther object, parallax is present. This indicates that the axis of rotation is not coincident with the entrance pupil. Adjustments must then be made to the fore-aft position of the camera on the panoramic head, and the rotational test is repeated.

The objective is to achieve a state where, upon panning the camera to the left and right, the two reference objects remain in perfect alignment, with no discernible relative displacement, as shown in Fig. 12. When this condition is met, the camera's axis of rotation is correctly aligned with the no-parallax point of the lens at the selected focal length. This position ensures that images captured from different angles will be free of parallax-induced stitching errors.

5.3. Details of Dynamic Scene Elements Annotation

A significant challenge in capturing longitudinal, "in-the-wild" datasets is the presence of dynamic scene elements. While our capture rig ensures a static viewpoint, the long temporal intervals between acquisitions mean that elements such as wind blown foliage, grass, and cloud formations inevitably move.

Although computational methods exist for image alignment (e.g., optical flow warping), applying such modifications would alter the pixel values and compromise the photometric integrity of our ground truth (GT) images. To preserve the dataset as a true GT reference, we opted instead to provide binary masks that identify these non-static regions. This approach allows researchers to optionally exclude these dynamic areas during metric computation, thereby isolating relighting performance from inconsistencies caused by scene motion.

We initially explored automated segmentation methods, including optical flow and contour detection, to identify these moving regions. However, these approaches produced unsatisfactory results, struggling with the subtle motions and complex natural textures present in our scenes. Consequently, we adopted a meticulous manual annotation process.

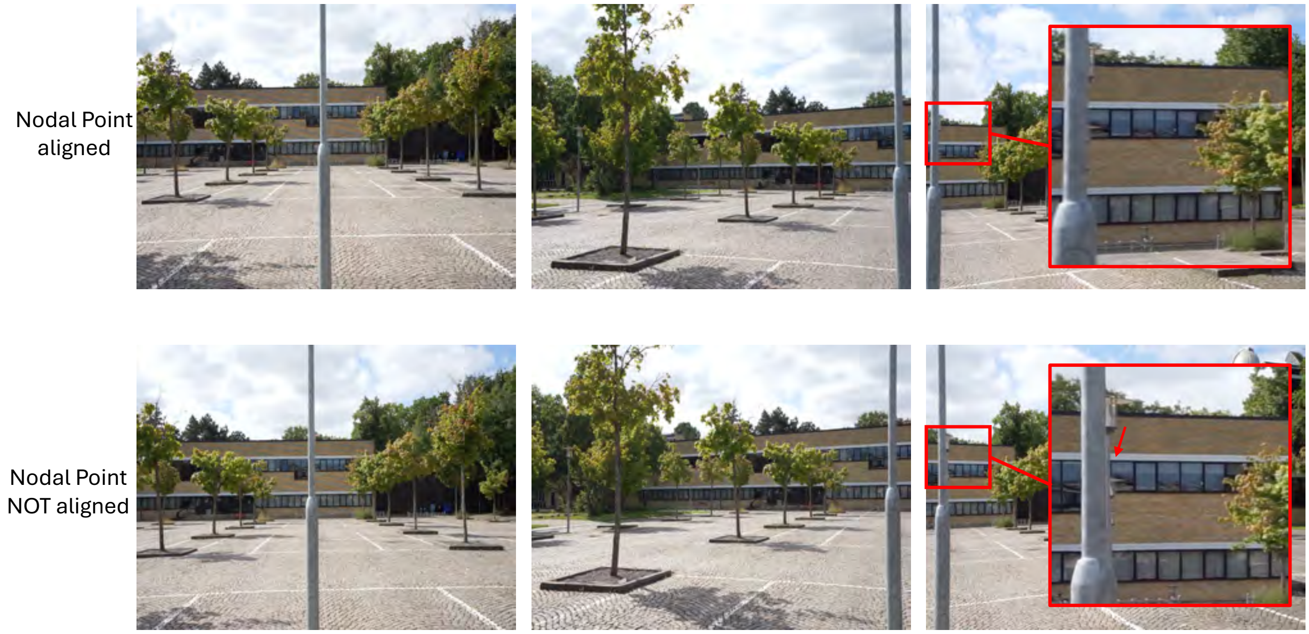

Figure 12. Nodal point alignment. When the camera is not positioned at the nodal point, rotating the camera causes the nearby utility pole to fail to occlude the poles behind it. In contrast, when the camera is at the nodal point, the nearby utility pole can occlude the poles behind it.

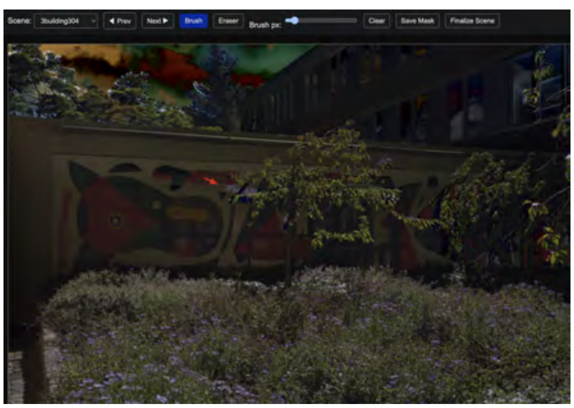

Figure 13. UI Interface for Marking Dynamic Scene Elements. Volunteers utilize a brush tool to create masks or an eraser tool to remove incorrectly marked regions. In the illustrated example, the areas indicated by red arrows correspond to discrepancies between the two photographs, highlighting regions where inconsistencies exist.

Our annotation pipeline is as follows:

1. **Pairwise Comparison:** For each scene, annotators performed a sequential, pairwise comparison of adjacent time steps (e.g., t_0 vs. t_1, t_1 vs. t_2, etc.).
2. **Difference Visualization:** To aid the human annotators, we generated absolute pixel-difference images for each pair. This visualization technique effectively accentuates the contours of misaligned objects, where pixel gradients are highest, making the boundaries of dynamic elements more conspicuous.
3. **Manual Annotation:** Annotators manually painted masks over all identified dynamic regions for each image pair. The primary targets for masking were clouds and moving vegetation (leaves, branches, and grass).
4. **Mask Aggregation:** The final mask for the entire scene is generated by computing the union of all pairwise masks. This ensures that any element that moved at any point during the capture sequence is included in the aggregate mask.

We explicitly excluded two categories of dynamic effects from masking. First, water surfaces (e.g., lakes and seas) were not annotated due to their highly complex and stochastic textures, which are intractable to mask reliably. Second, dynamic reflections and refractions were intentionally left unmasked, as we consider the ability to model these complex, illumination dependent light transport phenomena to be a core aspect of the relighting challenge itself. To facilitate the annotation of standard masks, we employ a dedicated user interface (UI) that enables manual annotation tasks (Fig. 13). The interface provides interactive tools for adjusting the brush size to generate masks, as well as an eraser function to correct erroneous regions.