أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Luma: Self-Supervised Realistic Lighting Estimation from Real Indoor Images

Deem Alwliaee, Lujain Bajunayd, Muhammad Mubashar

deem.alwliaee@kaust.edu.sa, lujain.bajunayd@kaust.edu.sa, muhammad.mubashar@strath.ac.uk

Abstract

Determining physically accurate indoor lighting from a limited number of images is a challenging problem that often involves balancing expensive scene-specific optimization with the physical precision of efficient feed-forward models, because lighting is entangled with material and geometry, where a single pixel could be the result of different combinations of illumination and albedo. In addition, no ground-truth lighting exists for real scenes. While recent approaches generate consistent illumination, they are constrained by high computational cost. To address this limitation, we introduce a physics-grounded teacher-student framework that derives high-fidelity illumination from multi-view inverse rendering and distilling this knowledge into a streamlined cross-scene lighting estimator. The approach consists of two main phases. Initially, a scene-specific Neural Incident Light Field (NeILF) [1] teacher, combined with a gated residual material corrector, is optimized using multi-view RGB-D observations and MVInverse material priors. This provides illumination pseudo-labels for the student network, which includes global environment maps, spatially-varying directional light fields, and teacher's physical renders. In the second phase, a single feed-forward student network is trained to predict these representations. The student utilizes a ConvNeXt-based visual encoder, gated cross-attention for fusing features from multiple views, and explicit geometric tokenization to separate and characterize localized emissive sources. The resulting feed-forward network generates physically consistent indoor lighting and flexible inference, seamlessly scaling from rich multi-view context.

Date: August 24, 2026

Code: https://github.com/LUJxAIN/luma

Contents

1 Introduction **2**

2 Related Work **2**

2.1 Physics-Based Inverse Rendering . . . 2

2.2 Single-Image Indoor Inverse Rendering . . . 3

2.3 Single-Image Illumination Estimation . . . 3

2.4 Diffusion-Based Illumination Estimation . . . 3

2.5 Diffusion-Based Inverse/Forward Rendering and Object Insertion . . . 3

2.6 Diffusion-Based Relighting . . . 3

2.7 Summary . . . 4

3 Methodology **4**

3.1 Method Details . . . 4

3.2 Dataset Preparation . . . 4

3.3 Data Processing . . . 4

3.4 Network Architecture . . . 5

3.5 Evaluation Procedure . . . 9

4 Results **9**

4.1 Results of Per-Scene Optimization . . . 9

4.2 Qualitative comparison of relighting results across four scenes. . . . 10

5 Discussion **11**

5.1 Correction Head Placement . . . 11

5.2 Student Results . . . 11

6 Limitations and Future Work **11**

7 Conclusion **12**

References **12**

1 Introduction

Many people rely on interior design software to visualize and plan their homes. Although these applications have advanced significantly with modern technology and Artificial Intelligence (AI), they often fail to accurately represent how a design will appear in the real environment. Current rendering applications tend to overestimate lighting and do not fully account for the physical behavior of light, resulting in unrealistic and overly optimistic visualizations. This often leads to inaccurate material appearance, unrealistic reflections, and lighting that does not match the actual scene.

The main challenges of this problem are:

- **Lighting ambiguity:** Recovering lighting from a single real indoor image is inherently challenging because the color of each pixel is influenced by multiple factors, including light intensity, light direction, surface material, and object geometry. Different combinations of these factors can produce the same observed pixel value, making the problem highly ambiguous.
- **Lack of ground-truth data:** Indoor lighting typically originates from multiple light sources with varying intensities and directions. As a result, creating a large-scale data set with accurate ground-truth lighting information is extremely difficult, limiting the availability of supervised training data.

To address these challenges, we propose a physics-grounded lighting model trained with a physically-based differentiable renderer, which re-renders the scene from the estimated lighting to rule out implausible solutions. Since no ground-truth lighting exists for real scenes, we train it in a self-supervised manner, using only the constraint that the re-rendered image must reproduce the input.

2 Related Work

Research on indoor lighting estimation has shifted from physics-based inverse rendering to learning-based single-image methods, and more recently to diffusion-based generative approaches. These techniques recover illumination from an indoor image to support applications such as realistic object insertion, scene relighting, and image editing.

2.1 Physics-Based Inverse Rendering

Early neural relighting methods build on inverse rendering models such as PhySG [2] and NeRFactor [3], which jointly decompose a scene into geometry, spatially-varying materials, and lighting. Using multi-view observations similar to NeRF [4], they enable physically accurate relighting and material editing. However, this accuracy comes at a high computational cost: these methods require dense multi-view captures and per-scene optimization, are mostly applied to isolated objects rather than full indoor scenes, and struggle with near-field or spatially varying indoor lighting because they typically assume distant illumination. Neural Incident Light Field (NeILF) [1], jointly estimates materials and incident lighting from multi-view images in a self-supervised manner, supervised only by image re-rendering. NeILF was evaluated on synthetic, DTU MVS, and BlendedMVS scenes, which are all clean and controlled multi-view settings.

2.2 Single-Image Indoor Inverse Rendering

A related line of work recovers spatially-varying lighting, geometry, and SVBRDF from a single RGB image rather than multi-view captures. Li et al. [5] introduce a large-scale synthetic dataset with ground-truth lighting, geometry, and materials for training and evaluation. Zhu et al. [6] propose an inverse rendering method for complex indoor scenes using differentiable Monte Carlo ray tracing, predicting geometry, spatially-varying lighting, and materials from a single image, trained on their synthetic InteriorVerse dataset. Wang et al. [7] estimate HDR spatially-varying lighting from a single LDR image using a Volumetric Spherical Gaussian representation with differentiable rendering for physically consistent lighting. These methods remain constrained by relying on synthetic data, which can limit their applicability on real indoor scenes.

2.3 Single-Image Illumination Estimation

To avoid multi-view reconstruction, learning-based methods estimate illumination directly from one image. Gardner et al. [8] established single-image HDR lighting regression using a feed-forward network trained on real SUN360 LDR panoramas and 2,100 self-captured HDR panoramas. Bai et al. [9] predict a low-dimensional spherical-Gaussian representation from a single image, giving computational efficiency and temporally stable shading, though it tends to smooth out fine lighting detail. EverLight [10] combines a parametric spherical-Gaussian model with GAN-based HDR panorama synthesis to predict explicit lighting, but remains limited by GAN artifacts. These methods support augmented-reality object insertion and relighting, but their low-dimensional or generator-driven representations limit their ability to reproduce sharp specular reflections and shadows.

2.4 Diffusion-Based Illumination Estimation

Recent methods reformulate indoor HDR lighting estimation as conditional image generation using pretrained diffusion priors. DiffusionLight [11], IllumiDiff [12], and the video-transformer-based LuxDiT [13] reconstruct HDR environment maps or inpaint HDR chrome-ball observations from a single image. Tong et al. [14] extend this to a spatiotemporal setting, estimating a continuous light field of spatially and temporally varying lighting from an image or video, fine-tuning diffusion priors to predict lighting at multiple locations for zero-shot generalization to real scenes. Diffusion models yield richer lighting with better high-frequency detail than regression-based methods, but remain computationally expensive (the original DiffusionLight requires iterative inpainting taking tens of minutes per image, prompting accelerated variants such as DiffusionLight-Turbo) and can produce lighting only partially consistent with the input image. Ensuring physical consistency and precise recovery therefore remains an open problem.

2.5 Diffusion-Based Inverse/Forward Rendering and Object Insertion

Diffusion models have also been used to model the full inverse and forward rendering pipeline. Diffusion Renderer [15] uses a video diffusion model to estimate scene attributes from a single video and re-render realistic relit images, trained on synthetic data supplemented with auto-labeled real videos. Ouroboros [16] improves speed and consistency using two single-step diffusion models linked by a cycle-consistency objective between forward and inverse rendering, for both indoor and outdoor scenes.

2.6 Diffusion-Based Relighting

Frameworks such as IC-Light [17], Neural Gaffer [18], and LumiNet [19] show that diffusion models can directly produce realistic lighting effects on images, though they are not designed for

illumination *estimation*. They generate relit images conditioned on image attributes, text prompts, or target lighting rather than computing an explicit lighting representation. While visually convincing, their lighting is baked into the output image and is not reusable for downstream tasks such as material editing, object insertion, or physically-based rendering. For such applications, explicit illumination estimation remains essential.

2.7 Summary

The methods reviewed above are validated almost exclusively on synthetic data or real data captured under controlled or object-centric conditions, leaving their behavior on cluttered real indoor scenes largely unverified. NeILF [1] has demonstrated successful self-supervised lighting and material estimation from real multi-view images, including evaluation on the DTU MVS dataset, which provides controlled multi-view observations with accurately calibrated camera poses and predefined illumination conditions. However, NeILF was not evaluated on complex, uncontrolled real indoor scenes. We therefore adapt NeILF to ScanNet++ [20], training it on real indoor data with multiple illumination sources and complex interactions between lighting, materials, and scene geometry.
We optimize NeILF using material estimates from MVInverse [21] and add a bounded, zero-initialized residual correction head to correct material errors without absorbing them into the estimated lighting. Finally, we distill the optimized per-scene model into a feed-forward student for faster single-image inference.

3 Methodology

In this section, we present the proposed framework for realistic light estimation from real indoor scenes using a physics-grounded teacher–student framework. A pretrained per-scene optimized Neural Incident Light Field (NeILF)[1] teacher provides the pseudo-labels for the student, generated from multi-view observations. The teacher's learned illumination fields, and corrected material maps are used as the training signals of a feed-forward student network. The pipeline takes RGB images as input, estimates its geometry and material properties, identifies candidate emitter regions, reconstructs the incident illumination, and renders the scene under the estimated lighting.

3.1 Method Details

The proposed pipeline is divided into several stages, including data preparation, data processing, geometry estimation, network prediction, differentiable rendering, and image refinement.

3.2 Dataset Preparation

The real-image training data were obtained from the ScanNet++ dataset [20], which provides RGB videos, LiDAR depth, and camera poses estimated using COLMAP. A preprocessing pipeline converts the compressed videos and depth streams into a frame-based format. Only frames with successfully estimated camera poses and valid alignment with the COLMAP reconstruction are retained. From these, 40 frames per scene are selected using farthest-point sampling over camera centers to maximize viewpoint diversity. For each frame, the RGB image and 16-bit depth map are extracted and stored in a CSV manifest with camera intrinsics, camera-to-world poses, and quality metrics.

3.3 Data Processing

3.3.1 Materials

To extract the material properties (albedo, metallic, roughness) of each frame, the pipeline replaces NeILF's jointly-optimized material branch with the pretrained MVInverse [21] as a frozen front-end, so that lighting is the only quantity optimized and the material and lighting ambiguity is reduced. Because MVInverse is pretrained primarily on synthetic data and is imperfect on real scenes, we add a residual correction head that fixes material errors within a limited range.

3.3.2 Geometry

Geometry is used in two ways:

- **Per-pixel positions and view directions:** Each pixel's depth is unprojected to camera coordinates and transformed by the camera-to-world pose into a 3D world position x, normalized by the scene center and scale. The direction from the camera center to that point gives the per-pixel view direction ω. Surface normals are derived from the sensor depth by smoothing and finite differencing, then normalized to unit length.
- **Mesh:** A 3D mesh of the room is reconstructed from the depth map. Provides scene geometry for shadows.

3.3.3 Candidate Emitter Detection

Candidate emitter identification involves detecting potential emissive or brightly illuminated areas within the analyzed image alongside the refined albedo. Two complementary criteria are utilized: a conventional bright-and-low-albedo test, which accounts for near-black emitters like exposed bulbs or windows, and an emission-ratio test that is independent of albedo:

$$R(x) = \frac{L(x)}{max(A(x),0.05)}$$

where L(x) is observed luminance and A(x) is albedo luminance. Connected components are identified from the combination of both criteria; any components that have a face count below a specified minimum are eliminated, while the remaining candidates are characterized by their normalized location in the image, approximate position in 3D space, average surface normal, and relative area in the image. These candidate regions fulfill two functions later in the process: they act as specific anchors for the teacher's compact lighting output and serve as the input tokens that the student's emitter head analyzes.

3.4 Network Architecture

The pipeline is divided into three stages:

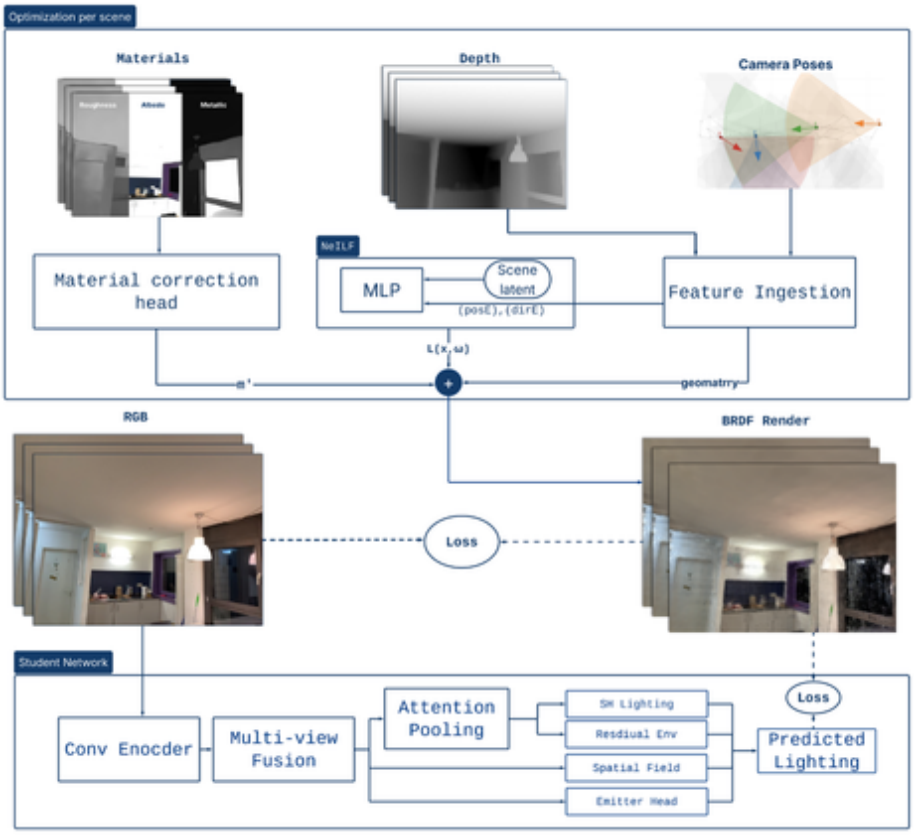


Figure 1 Overall System Architecture.

3.4.1 Per Scene Teacher Optimization

- **Incident light field.** The core object is the incident light field $\mathbf{L}(\mathbf{x}, \omega)$. It takes a 3D position $\mathbf{x}$ and incident direction ω as inputs, with both encoded using positional encoding (PosE and DirE). The network is also conditioned on a per-scene latent, which is optimized during training.
- **Material correction head.** The main addition is a bounded, zero-initialized, gated residual

$$\mathbf{m}' = \mathbf{m} + \mathbf{g} \odot \Delta\mathbf{m},$$

  which corrects known MvInverse errors without reopening the material–light ambiguity. The head predicts residuals for albedo, roughness, and metalness $(\Delta a, \Delta\rho, \Delta m)$. Two additional corrections are introduced: ΔF, which adjusts the specular reflectance at normal incidence, and Δ_{emission}, which accounts for surfaces that emit light rather than only reflect it. Corrections are activated one material component at a time. Each correction has sufficient freedom to change the material, but the change is accepted only when it reduces the reconstruction residual.
- **Two-stage optimization.** Optimization proceeds in two stages. First, the incident light field is optimized together with the per-frame exposure using the fixed MVInverse materials. The light field is then frozen while the material correction head is trained on top of it. This provides a stable radiance estimate for material correction and prevents material errors from being absorbed into the estimated lighting.
- **Differentiable rendering.** Given the optimized light field, each surface point is rendered using a differentiable, physically based shading integrator. For a point $\mathbf{x}$ with normal $\mathbf{n}$,

viewing direction $\mathbf{v}$, and material parameters (a, ρ, m), the outgoing radiance is

$$L_o(\mathbf{x}, \mathbf{v}) = \int_\Omega L(\mathbf{x}, \boldsymbol{\omega}_i) f_r(\boldsymbol{\omega}_i, \mathbf{v}; a, \rho, m)(\mathbf{n} \cdot \boldsymbol{\omega}_i)\, d\boldsymbol{\omega}_i.$$

The integral is estimated using Monte Carlo sampling with M incident directions drawn from a cosine-weighted distribution over the hemisphere around $\mathbf{n}$. The field is queried once for each sampled direction. The BRDF combines a Lambertian diffuse component, scaled by $(1 - m)$, with a Cook–Torrance GGX specular component.

- **Optimization.** The renderer is fully differentiable, allowing the photometric loss between rendered and observed images to backpropagate into the light field, per-frame exposure, and material correction head.

3.4.2 Teacher Light Supervision Pseudo-Labels

After the teacher NeILF model is trained on multi-view ScanNet++ scenes, it is used to generate the supervision targets for the student model.

For each frame, the pipeline computes and stores: global environment map, spatially varying local-light map, directional local-light field, candidate-emitter radiance estimate, global exposure value, and the teacher rendered image. The corrected material maps from the teacher is also retained and used in the pipeline. These targets combined provide a richer, both global and localized signals for the student model than a single environment map.

3.4.3 Student Network

The student model is designed as encoder-decoder architecture as shown in Figure 1, it encodes input images using ConvNext-Tiny backbone to extract hierarchal features. The prediction is a combination of low-order spherical harmonics, residual environment map, spatial light grid, and emitter values (color, intensity, and direction) for each candidate illumination source. The network is composed of:

- Visual encoder: using ConvNext-Tiny backbone, it extracts hierarchal spatial features at different resolutions. The encoder processes the RGB image along with the surface normals, and the features are projected to a common dimensionality then fused to produce a spatial feature map with pooled global latent repesentation.
- Multi-view fusion module: gated cross-attention that combines features from multiple frames of the same scene (also called companion views), conditioned on their relative camera rotations. It is initially set to zero, such that it only activates in training when useful, making the network capable of working even when only a single image is available.
- Global environment heads: two components generate the global incident light. Composed of a linear head that predicts nine low-order spherical harmonic coefficients, and a residual decoder modifies this base to generate a high-frequency environment map.
- Exposure head: predicts global scaling factor to handle a larger dynamic range.
- Directional field head: a convolutional head predicts K = 12 incoming radiance samples, which form a directional local-illumination field. The mean radiance is used to derive a spatial light value which allows to modify the rendered image based on localized illumination.
- Explicit emitter head: the candidate illumination regions are converted to tokens that combines the visual features with geometry. Then a transformer-based with cross-attention is used to predict the presence probability, reflection routing, color, intesnity, and 3D direction.

During inference, the model takes multi-view scenes when available; from a reference frame, a small number of companion views are fused together using the multi-view fusion module. However during single-view inference, since the fusion module is gated, the student can bypass it and operate on a single image.

3.4.4 Training

The student is trained through a multi-phase training curriculum that gradually teaches the network abstract representations of lighting, which are later tied to the principles of physical image formation. The optimization process utilizes difficulty-aware sampling that prioritizes challenging examples and employs gradient accumulation to ensure stability.

Phase 1: Multi-Objective Knowledge Distillation

In the initial phase, the student directly regresses the cached teacher targets to build fundamental lighting representations. During this stage, the student is trained with multi-view inputs: a reference frame paired with two randomly selected companion views from the same scene.

- Direct supervision: in this part, the environment map and directional lighting fields are optimized using logarithmic, relative (L_1), and chromatic radiometric errors. The spherical-harmonic coefficients and emitter parameters are supervised directly from the teacher predictions, while exposure is supervised in log-space to handle larger dynamic range variations. Moreover, the spatial local-light field is supervised with a combination of logarithmic and relative L_1 errors that are evaluated on normalized illumination.
- Regularization: a variance-covariance regularizer (VICReg-style) is applied to the global latent vectors, to prevent representation collapse.
- Consistency and attention: for robust feature extraction, the network is constrained by multi-view and augmentation consistency losses. Moreover, attention-distillation loss is added to ensure that the student's spatial attention map aligns with the teacher's rendered image.

Phase 2: Physics-Grounded Differentiable Rendering

The second phase establishes a direct connection between predicted lighting and image formation. In this phase, companion views are omitted, and a fully differentiable PyTorch Monte Carlo renderer is employed for each image, thereby removing the requirement for external ray tracers.

- Physical integration: the renderer uses cosine-weighted diffuse sampling and GGX-importance specular sampling to evaluate the outgoing radiance by using the student's environment map, local light field, along with the teacher-corrected materials and surface normals.
- Rendering losses: since the renderer is differentiable, the gradients flow directly from the rendered image to the lighting decoder. The primary loss is the image reconstruction loss against the real image. It is regularized by secondary reconstruction term against the teacher's render, augmented with image-gradient penalties and shadow-sensitive reweighting to ensure that boundary regions are heavily penalized.
- Anchor constraints: prevents the network from drifting far away from physically accurate light parameters.

After rendering, an optional refiner UNet is added to handle residual image-space editing. This network predicts a hyperbolic-tangent-limited residual to polish high frequency objects. It is initialized with zero such that it only activates when a refiner-training stage is explicitly run. During training, the data is split on scene-level, with a hold-out set for validation, ensuring

generalization on unseen environments.

3.5 Evaluation Procedure

To run the evaluation, a held-out set from the ScanNet++ dataset is used. For each evaluation frame, the student's predicted environment map and light field are compared against the teacher targets. The environment map mean absolute error (MAE), light field MAE, relative exposure error, and dominant light direction error are used as metrics to evaluate the model. The predicted lighting is also rendered via differentiable PyTorch renderer, and compared to both the teacher's render and the original image. The render is evalauated using a set of metrics including: MAE, PSNR, SSIM, scale-invariant RMSE (SI-RMSE) and LPIPS. SSIM and LPIPS are computed on the masked tonemapped images, while the SI-RMSE is computed in linear space. Finally, the metrics for each frame, how many views were used, and inference time are recorded to assess the performance of the network.

4 Results

In this section, we present the results of our experiments and evaluate the performance of the proposed approach.

4.1 Results of Per-Scene Optimization

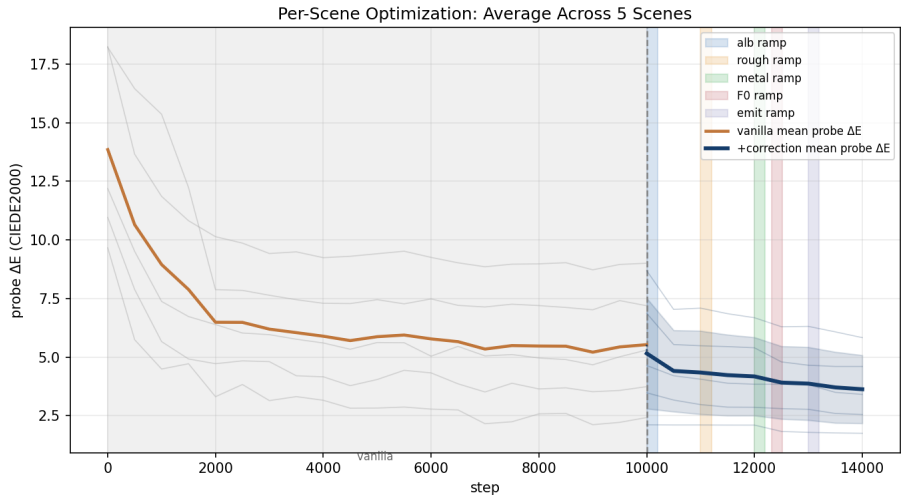


Figure 2 Per-scene optimization results averaged across five scenes. The plot shows the mean probe ΔE (CIEDE2000) for the vanilla model and the model with correction. Shaded regions indicate the optimization intervals for the different ramp materials

4.2 Qualitative comparison of relighting results across four scenes.



Figure 3 Qualitative comparison of relighting results across four scenes. Target images are compared with the results produced by NeILF and NeILF augmented with the proposed correction head.

Table 1 shows the evaluation metrics on the student model across across 754 frames. Each record is evaluated of four reconstruction error metrics (PSNR, SSIM, Render MAE, LPIPS) illumination/geometry metrics (Environment MAE, Field MAE, Light Direction Error), and runtime performance (Inference Time). All frames were evaluated using multi-view input (3 views per frame).

Table 1 Evaluation metrics over the test set.

Metric	Count	Mean	Std	Min	Median (50%)	Max
PSNR (dB)	754	**17.79**	2.60	8.42	18.06	25.49
SSIM	754	**0.725**	0.101	−0.011	0.742	0.907
Student Render MAE	754	**0.099**	0.032	0.041	0.093	0.280
Teacher Render MAE	754	**0.294**	0.055	0.102	0.302	0.397
LPIPS (Perceptual)	754	**0.320**	0.080	0.089	0.311	0.806
Inference Time (s)	754	**0.240**s	0.623	0.171	0.219	17.31s

5 Discussion

In this section, we discuss and interpret the results presented in Section 4.

5.1 Correction Head Placement

The results in Section 4 show that placing the correction head after NeILF provides a better separation between lighting estimation and error correction. Applying the correction during NeILF's lighting estimation introduced interference between the two processes, making it difficult to distinguish lighting errors from material-related artifacts. In contrast, the post-NeILF correction reduces the residual ΔE, as shown in Figure 2, while improving visual details such as the bedspread, board reflections, and metallic kettle in Figure 3. These improvements are important for student training, as reducing material-related artifacts provides a cleaner supervision signal and allows the student to focus on learning the underlying lighting rather than unnecessary material variations.

5.2 Student Results

The evaluation results show that the proposed method achieves a good reconstruction quality and improves the rendering accuracy in comparison to the teacher model. Across the 754 samples, the student achieves a mean PSNR of 17.79 dB and a mean SSIM of 0.725, which indicates a reasonable structural and pixel-level similarity between the rendered outputs and the reference images. The SSIM value proves that the model is generally able to preserve the major structural and illumination characteristics of the scenes, although the relatively moderate PSNR indicates that noticeable pixel-level differences remain.

The most significant improvement is the rendering error. The Student Render MAE is 0.099, compared with 0.294 for the teacher model. This represents approximately a 66.3% reduction in MAE, suggesting that the student produces rendered images that are considerably closer to the target appearance. This improvement is particularly important because the rendering error directly reflects the quality of the estimated lighting when applied to the scene. The lower student MAE therefore indicates that the knowledge transferred from the teacher has enabled the student to reproduce scene illumination more accurately while potentially avoiding some of the errors produced by the more computationally expensive teacher.

6 Limitations and Future Work

The proposed physics-grounded framework successfully addresses the gap between the expensive multi-view optimization and efficient feed-forward architecture. However, there are still opportunities for further exploration:

- **Dependency on material priors:** the pipeline utilizes MVInverse generated material estimates for the rendering. Although the teacher optimization stage has a material-correction head, it is still bounded to remain conservative, which could have a negative effect on sample especially for metallic and transparent objects. Future work could extend by jointly optimizing the student predictions with a feed-forward material estimation network during the differentiable rendering phase.
- **Heuristic Emitter Detection:** during data pre-processing, fixed heuristic thresholds are used to identify localized light sources. This effectively identifies standard light sources such as light bulbs, however, it might struggle to identify subtle emissive surfaces. Exploring a fully learnable attention-based emitter within a student network could be a further improvement on the architecture.

7 Conclusion

This work presented a comprehensive physically-grounded teacher-student framework for real indoor scenes light estimation. The proposed method provided a combination between high-fidelity consistent illumination from multi-view observations using per-scene optimized NeILF model, and distills this knowledge into a faster student with encoder-decoder architecture. Through this work, we addressed the gaps observed in previous methods, by allowing the student model to learn global and localized light fields and emitter values with a feed-forward manner enabling faster inference, while achieving a 66.3% reduction in MAE which indicates the ability of the model to reconstruct images relative to the original. Moreover, incorporating a fully differentiable Monte Carlo renderer in the student's optimization loop ensures the network predictions are bounded by the ability to reconstruct the original ground-truth image.

References

[1] Yao Yao et al. *NeILF: Neural Incident Light Field for Physically-based Material Estimation*. 2022. arXiv: 2203.07182 [cs.CV]. URL: https://arxiv.org/abs/2203.07182.

[2] Kai Zhang et al. "Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5453–5462.

[3] Xiuming Zhang et al. "Nerfactor: Neural factorization of shape and reflectance under an unknown illumination". In: *ACM Transactions on Graphics (ToG)* 40.6 (2021), pp. 1–18.

[4] Ben Mildenhall et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[5] Zhengqin Li et al. "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2475–2484.

[6] Jingsen Zhu et al. "Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing". In: *Siggraph asia 2022 conference papers*. 2022, pp. 1–8.

[7] Zian Wang et al. "Learning indoor inverse rendering with 3d spatially-varying lighting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12538–12547.

[8] Marc-André Gardner et al. "Learning to predict indoor illumination from a single image". In: *arXiv preprint arXiv:1704.00090* (2017).

[9] Jiayang Bai et al. “Deep graph learning for spatially-varying indoor lighting prediction”. In: *Science China Information Sciences* 66.3 (2023), p. 132106.

[10] Mohammad Reza Karimi Dastjerdi et al. “Everlight: Indoor-outdoor editable hdr lighting estimation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 7420–7429.

[11] Pakkapon Phongthawee et al. “Diffusionlight: Light probes for free by painting a chrome ball”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 98–108.

[12] Shiyuan Shen et al. “Illumidiff: Indoor illumination estimation from a single image with diffusion model”. In: *IEEE transactions on visualization and computer graphics* (2025).

[13] Ruofan Liang et al. “Luxdit: Lighting estimation with video diffusion transformer”. In: *Advances in Neural Information Processing Systems* 38 (2026), pp. 5015–5046.

[14] Mutian Tong, Rundi Wu, and Changxi Zheng. “Spatiotemporally Consistent Indoor Lighting Estimation with Diffusion Priors”. In: *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 2025, pp. 1–11.

[15] Ruofan Liang et al. “Diffusion renderer: Neural inverse and forward rendering with video diffusion models”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 26069–26080.

[16] Shanlin Sun et al. “Ouroboros: Single-step diffusion models for cycle-consistent forward and inverse rendering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 10386–10397.

[17] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport”. In: *The Thirteenth International Conference on Learning Representations*. 2025.

[18] Haian Jin et al. “Neural gaffer: Relighting any object via diffusion”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 141129–141152.

[19] Xiaoyan Xing et al. “Luminet: Latent intrinsics meets diffusion models for indoor scene relighting”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 442–452.

[20] Chandan Yeshwanth et al. “ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 12–22.

[21] Xiangzuo Wu et al. “MVInverse: Feed-forward Multiview Inverse Rendering in Seconds”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2026, pp. 37343–37357.