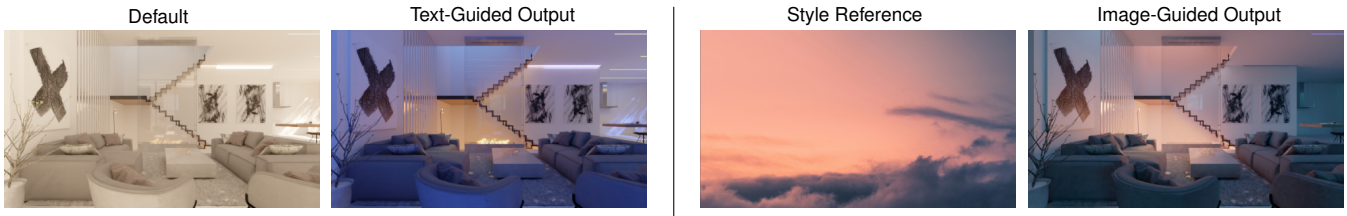
Differentiable Objectives for 3D Scene Relighting via Gradient Descent on OLAT Basis Coefficients

Anson Savage, Parris Egbert, Seth Holladay

Department of Computer Science, Brigham Young University, USA



(a) *Lighting optimized using a target prompt: "an interior photo at night"*

(b) *Lighting optimized using guidance from a reference image*

Figure 1: *We propose a render-engine-agnostic, automatic lighting routine that can be guided by text prompts (a) or a reference image (b).*

Abstract

Designing effective lighting is an iterative and often time-consuming process. This work contributes to automatic lighting design research by presenting a render-engine-agnostic optimization routine: gradient descent on RGB multipliers of one-light-at-a-time (OLAT) basis images. We propose several objective functions to accomplish lighting tasks and show that our method is capable of quickly and effectively exploring different lighting styles using either text prompts or reference images.

CCS Concepts

• ***Computing methodologies*** → *Rendering; Graphics systems and interfaces; Machine learning;*

1. Introduction and background

Lighting plays a critical role in shaping the mood and aesthetic of visual art. The lighting process is iterative and can require significant time to achieve a desired outcome, especially when using path-tracing. Adjusting lighting in compositing or using real-time prototyping can help, but rapid exploration of styles is still challenging.

Recent progress in generative AI has prompted research in relighting images using neural networks (e.g., [MHT*25]). Although such methods are capable image relighters, because they do not rely on an underlying scene structure or light transport simulation, they face limitations, including difficulty of control, a lack of content consistency, physically inaccurate lighting, and temporal instability, making them challenging to use in animation.

To address this, we optimize interpretable lighting parameters (RGB multipliers of OLAT basis images) via gradient descent. We introduce novel objectives using modern neural networks to evaluate similarity with a text prompt or reference image. Our workflow allows artists to quickly explore styles and further adjust lighting after optimization. The source code is available on GitHub.

2. Related work

Automatic lighting design research has long pursued "lighting-by-example," where a user provides a reference image with a lighting style they wish to emulate. Recent work has also explored guiding this process with text prompts.

2.1. Lighting-by-example

Hudon et al. [HCB16] demonstrate one method of lighting-by-example in which the objective is to match the luminance histogram of a 3D render with that of a reference image. Similar to our work, they optimize multipliers of OLAT basis images. However, likely due to a non-differentiable objective, they use a genetic algorithm for optimization. Galvane et al. [GLCC18] present an approach in which the objective is to match both luminance histograms and gradient histograms, improving their system's ability

DOI: 10.2312/egs.20261020

to match lighting direction. They use particle swarm optimization, which requires many re-evaluations of the objective, so using a real-time engine to achieve manageable optimization speeds was key for their method. Most recently, Ecormier-Nocca et al. [ELU*25] perform lighting-by-example by first synthesizing light textures that match those of lights in a reference image, and then using gradient-based optimization to match the average color of the non-light surfaces from the reference image. Their method effectively transfers light source style, but depends on a specialized differentiable renderer [LHE*24].

2.2. Text-guided lighting design

Little prior work has explored text-conditioned lighting optimization for view-dependent rendering. Oh et al. [OKK24] indirectly use text to guide optimizing for a lighting mood. In their method, they use a diffusion model to generate a lighting example image conditioned on a depth map of the 3D scene. Optimization is performed to minimize pixel-wise MSE with the reference image using a grid search.

Contrastive learning on image and text pairs (e.g., CLIP [RKH*21]) has a side effect of producing rich priors for downstream aesthetic evaluation tasks [HKH22], suggesting the utility of using "CLIP-like" models for aesthetic, text-guided lighting optimization. In contrast to prior work, we leverage such developments in modern vision and language models for use with gradient-based optimization of lighting parameters.

3. Methodology

Our framework pairs a core auto-diff optimization method ("OLAT-AD", algorithm 1) with arbitrary differentiable objective functions.

3.1. OLAT-AD

OLAT images are first rendered in a linear RGB color space. Each render I_i is multiplied by a learned RGB coefficient θ_i and summed to create a linear combination (X, Line 3). Optional random augmentations (Line 4) can be applied to encourage diversity and help avoid adversarial optima. Tone mapping (Line 5) brings the image into an sRGB space, which most vision models are trained in; a differentiable approximation of the same tone mapping used in the lighting artist's 3D software should be applied. Users can control various optimization parameters, including multiplier bounds ($\theta_{\min}, \theta_{\max}$; by default $\theta_{\min} = 0$ to keep results physically plausible) and the initial standard deviation σ, which can be increased to encourage batch exploration. Intermediate results can be displayed, allowing the user to visually compare candidate results both within batches and across iterations. Upon selecting a preferred result, the corresponding coefficients $\{\theta^\star\}$ can be applied to the light intensities in the 3D software for further adjustment.

3.2. Neural cost functions

OLAT-AD can work with any differentiable cost function that can score an image as its lights are being optimized. Here we present cost functions that we found to be effective.

Algorithm 1: OLAT-AD

Input : $\{I_i\}$ images; $\mathcal{L}$ loss; $\mathbf{g}$ guidance image/text; σ std. dev.
Output: Learned multipliers $\{\theta^\star\}$

1 Sample $\theta^{(0)} \sim \mathcal{N}(1, \sigma^2\mathbf{I})$ where $\theta_i \in \mathbb{R}^3$ is a per-light multiplier
2 **repeat**
3 $\quad X \leftarrow \sum_{i=1}^{n} \theta_i \odot I_i$
4 $\quad \tilde{X} \leftarrow \textsc{Augment}(X)$
5 $\quad \hat{X} \leftarrow \textsc{ToneMapping}(\tilde{X})$
6 $\quad L \leftarrow \mathcal{L}(\hat{X}, \mathbf{g})$
7 $\quad \theta \leftarrow \mathsf{Opt}(\theta, \nabla_\theta L)$ `// Adam optimizer [KB17]`
8 $\quad \theta \leftarrow \textsc{Clamp}(\theta; \theta_{\min}, \theta_{\max})$
9 **until** *converged*
10 **return** $\{\theta^\star\}$

(a) *Default (pre OLAT-AD)* (b) *Optimized with* $\mathcal{L}_{Absolute}$ (c) *Optimized with* $\mathcal{L}_{Relative}$

Figure 2: *Results using the original CLIP (ViT) model with the target prompt: "a 3D rendering of a car with beautiful, aesthetic lighting". The default render before applying OLAT-AD is shown in (a). Using* $\mathcal{L}_{Absolute}$ *in (b) produces an overexposed, adversarial or local solution, while using* $\mathcal{L}_{Relative}$ *with the same target prompt plus the initial prompt: "a 3D rendering of a car with boring, ugly, overexposed, flat lighting" produces a more aesthetic result in (c).*

3.2.1. Similarity to a text prompt

A basic cost function for similarity with a text prompt is the cosine distance between the image embedding ($\mathbf{I}$) and the target prompt embedding ($\mathbf{p}$), as embedded by a "CLIP-like" model:

$$\mathcal{L}_{\text{Absolute}} := \text{CosDist}(\mathbf{I}_{\text{cur}}, \mathbf{p}_{\text{tgt}}) = 1 - \frac{\mathbf{I}_{\text{cur}} \cdot \mathbf{p}_{\text{tgt}}}{\|\mathbf{I}_{\text{cur}}\| \|\mathbf{p}_{\text{tgt}}\|} \tag{1}$$

Gal et al. [GPM*21] showed that using an initial prompt and a target prompt and maximizing cosine similarity between the deltas can help to avoid adversarial solutions (see Figure 2). We define:

$$\mathcal{L}_{\text{Relative}} := \text{CosDist}(\mathbf{I}_{\text{cur}} - \mathbf{I}_{\text{init}}, \mathbf{p}_{\text{tgt}} - \mathbf{p}_{\text{init}}) \tag{2}$$

Frans et al. [FSW21] discovered that applying random augmentation at each optimization step is one way to avoid adversarial optima when optimizing Bézier curves for similarity with a CLIP embedding. Applying this method (using random resize and crop) at each optimization step, we find that our results can improve beyond only using $\mathcal{L}_{\text{Relative}}$ (see Figure 3). Using stochastic augmentations can also encourage result diversity within a batch. (Note: both Figure 2 and Figure 3 use a standard sRGB Gamma curve instead of the AgX tone mapping used in the remaining figures; its harsher highlight clipping better illustrates the effect of $\mathcal{L}_{\text{Relative}}$ and random augmentations.)

(a) $\mathcal{L}_{Relative}$ *only* (b) *With random resize/crop*

Figure 3: *Here, the default original CLIP (ViT) model is used with a target prompt: "a 3D rendering of a car at night" (and the same initial prompt as Figure 2c). Applying random augmentations enables the optimizer to produce a more semantically correct result.*

3.2.2. Similarity to an image

Any differentiable image similarity metric can be used as a loss function for lighting-by-example. We found good results with cosine similarity in "CLIP-like" embedding spaces, and with the neural style loss proposed by Gatys et al. [GEB15].

We also tested traditional metrics for image similarity, including mean SSIM [WBSS04] (used in Figure 1b) and LPIPS [ZIE*18]. While random augmentations can help when optimizing using a "CLIP-like" model, we found that they do not typically improve results for general image-guided lighting tasks.

Off-the-shelf vision-language models have significant world knowledge, but they have not been trained specifically for lighting tasks. A model more suitable for lighting tasks would ideally be able to recognize lighting concepts while mostly ignoring image content. We fine-tuned ViT-B-16-SigLIP-512 [ZMKB23] on pairs of synthetic renders with the same lighting but different content and found that the resulting "lighting embedding" can sometimes improve results for both text and image guidance. See the supplemental video for comparisons and more details.

4. Results

The main cost of our method is the initial OLAT rendering. When using the ViT-B-16-SigLIP-512 model, a single NVIDIA H100 GPU, a batch size of 4, a learning rate of 0.06, and a max iteration count of 230, typical optimization time and peak memory are 20 seconds and 5.6 GB. The models used here have similar parameter counts and thus yield similar computational performance. Optimization speed is independent of image size, as all input images are center-cropped to the model's native resolution. Though OLAT-AD scales linearly with the number of lights, runtime is primarily determined by the model size (models larger than those used here can significantly increase optimization time).

4.1. Text-guided lighting

Text-guided lighting results and their corresponding prompts ("*initial*"→"*target*") are shown in Figure 4. The ability to optimize for more abstract concepts (like mood) is shown in Figure 5.

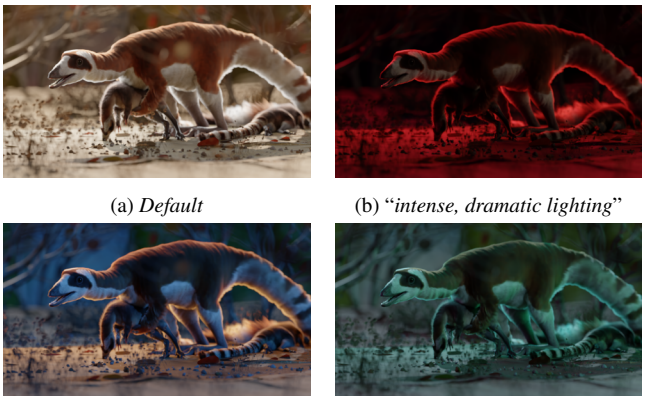
4.2. Lighting with a reference image

Figure 6 shows the ability of neural style transfer and LPIPS to mimic diverse lighting styles, while Figure 7 highlights the generalization of neural style loss across different scenes. Though each loss metric we tried can produce good relighting results, we found it challenging to determine which will perform the best for a given scene/reference before testing it. However, some patterns did emerge. For example, using cosine similarity in a SigLIP embedding space maintains attributes like light direction more than the style transfer loss (compare Figure 6b and Figure 8a), likely due to the spatially invariant construction of the style term [GEB15]. Figure 8 shows that using LPIPS or SSIM loss functions fails to effectively relight the scene; however, the colors that are present in the LPIPS optimization more closely match the reference colors than SigLIP's.

(a) *Default (before OLAT-AD)* (b) "*harsh fluorescent lighting*" → "*soft incandescent lighting*"

(c) "*soft, pale lighting*" → "*vivid, neon lighting*" (d) "*cold, dull blue lighting*" → "*warm, golden lighting*"

Figure 4: *Results using* $\mathcal{L}_{Relative}$ *("initial"→"target")*



(a) *Default* (b) "*intense, dramatic lighting*"

(c) "*hopeful, happy, lighting on subject, blue lighting in background*"* (d) "*sad, rainy, gloomy lighting*"*

Figure 5: *Results using more abstract prompts. Initial prompts omitted for brevity. (*) Optimized using a fine-tuned model.*

5. Conclusion

as We have presented a render-engine-agnostic framework for automatic lighting design that offers rapid exploration of lighting styles due to its differentiable configuration and advances in modern neural networks. Unlike "black-box" generative methods, our approach

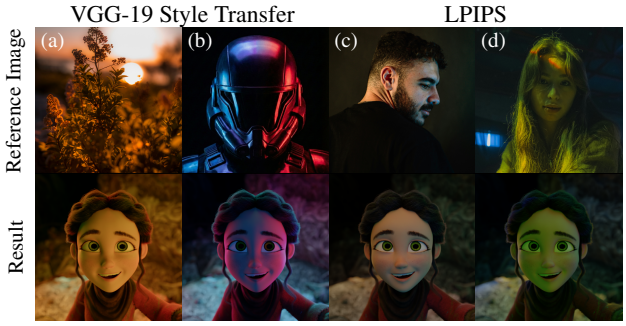


Figure 6: *Image-guided optimization w/ style transfer and LPIPS.*

Figure 7: *Neural style transfer loss generalizes well across a variety of scenes. Reference image from Figure 6a.*

yields editable light parameters, making it more useful to artists for refinement. Future work could explore optimizing light transforms using differentiable rendering [LHE*24] or gradient-free optimization methods like Bayesian optimization [VEPG23].

6. Acknowledgments

We thank Lamar Salama for her help generating training data for model fine-tuning and for curating results. We also thank Isaac Criddle for his guidance on color spaces and tone mapping. Assets are CC0/royalty-free, excluding the *house*, *girl*, *girl and dog*, and *robot* (CC-BY), and the *dinosaur* (CC-BY-SA 4.0).

References

[ELU*25] ECORMIER-NOCCA, P., LIPP, L., ULSCHMID, A., et al. "Single-Exemplar Lighting Style Transfer via Emissive Texture Synthesis and Optimization:" *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. Porto, Portugal: SCITEPRESS - Science and Technology Publications, 2025, 113–126. ISBN: 978-989-758-728-3. DOI: 10.5220/0013193900003912 2.

[FSW21] FRANS, K., SOROS, L. B., and WITKOWSKI, O. *CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders*. arXiv:2106.14843 [cs]. June 2021. DOI: 10.48550/arXiv.2106.14843 2.

[GEB15] GATYS, L. A., ECKER, A. S., and BETHGE, M. *A Neural Algorithm of Artistic Style*. arXiv:1508.06576 [cs]. Sept. 2015. DOI: 10.48550/arXiv.1508.06576 3.

[GLCC18] GALVANE, Q., LINO, C., CHRISTIE, M., and COZOT, R. "Directing the Photography: Combining Cinematic Rules, Indirect Light Controls and Lighting-by-Example". en. *Computer Graphics Forum* 37.7 (Oct. 2018), 45–53. ISSN: 0167-7055, 1467-8659. DOI: 10.1111/cgf.13546 1.

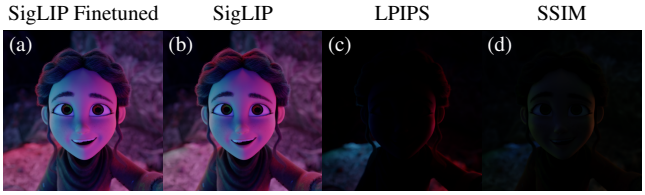


Figure 8: *Ablation across similarity metrics, revealing failure cases. Reference image from Figure 6b.*

[GPM*21] GAL, R., PATASHNIK, O., MARON, H., et al. *StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators*. arXiv:2108.00946 [cs]. Dec. 2021. DOI: 10.48550/arXiv.2108.00946 2.

[HCB16] HUDON, M., COZOT, R., and BOUATOUCH, K. "Automatic light compositing using rendered images". *2016 Digital Media Industry & Academic Forum (DMIAF)*. July 2016, 176–179. DOI: 10.1109/DMIAF.2016.7574927 1.

[HKH22] HENTSCHEL, S., KOBS, K., and HOTHO, A. "CLIP knows image aesthetics". eng. *Frontiers in Artificial Intelligence* 5 (2022), 976235. ISSN: 2624-8212. DOI: 10.3389/frai.2022.976235 2.

[KB17] KINGMA, D. P. and BA, J. *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980 [cs]. Jan. 2017. DOI: 10.48550/arXiv.1412.6980. URL: http://arxiv.org/abs/1412.6980 2.

[LHE*24] LIPP, L., HAHN, D., ECORMIER-NOCCA, P., et al. "View-Independent Adjoint Light Tracing for Lighting Design Optimization". *ACM Trans. Graph.* 43.3 (May 2024), 35:1–35:16. ISSN: 0730-0301. DOI: 10.1145/3662180 2, 4.

[MHT*25] MAGAR, N., HERTZ, A., TABELLION, E., et al. "LightLab: Controlling Light Sources in Images with Diffusion Models". *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. arXiv:2505.09608 [cs]. Aug. 2025, 1–11. DOI: 10.1145/3721238.3730696 1.

[OKK24] OH, J., KIM, S., and KIM, S. "LumiMood: A Creativity Support Tool for Designing the Mood of a 3D Scene". *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. New York, NY, USA: Association for Computing Machinery, May 2024, 1–21. ISBN: 979-8-4007-0330-0. DOI: 10.1145/3613904.3642440 2.

[RKH*21] RADFORD, A., KIM, J. W., HALLACY, C., et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020 [cs]. Feb. 2021. DOI: 10.48550/arXiv.2103.00020 2.

[VEPG23] VITSAS, N., EVANGELOU, I., PAPAIOANNOU, G., and GKARAVELIS, A. "Opening Design using Bayesian Optimization". *Virtual Reality & Intelligent Hardware* 5.6 (Dec. 2023), 550–564. ISSN: 2096-5796. DOI: 10.1016/j.vrih.2023.06.001 4.

[WBSS04] WANG, Z., BOVIK, A., SHEIKH, H., and SIMONCELLI, E. "Image quality assessment: from error visibility to structural similarity". *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), 600–612. ISSN: 1941-0042. DOI: 10.1109/TIP.2003.819861 3.

[ZIE*18] ZHANG, R., ISOLA, P., EFROS, A. A., et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". en. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, June 2018, 586–595. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00068 3.

[ZMKB23] ZHAI, X., MUSTAFA, B., KOLESNIKOV, A., and BEYER, L. "Sigmoid Loss for Language Image Pre-Training". en. 2023, 11975–11986. DOI: 10.1109/ICCV51070.2023.01100 3.