Generating Synthetic Illumination Variation from Sparse Co-Located Captures

Anonymous CVPR submission

Paper ID 74

Abstract

Low-light robot vision is often improved by simply adding an onboard light, but the chosen light level changes image formation, can introduce failure modes such as saturation or glare, and is difficult to study exhaustively with real recollection. Further, it is challenging to completely model the scene geometry and composition to completely model external illumination since the robot is moving. We address these two challenges with CLID, a co-located illumination decomposition method for efficient synthetic data generation from sparse real captures. Given an observed image and a target light intensity, CLID predicts a compact decomposition and analytically synthesizes the corresponding scene appearance under new onboard-light settings, turning one captured sequence into a dense family of counterfactual observations. This yields a scalable synthetic-data pipeline for controlled illumination variation while staying grounded in real sensor measurements. We show that CLID expands 10,153 robot frames into 101,530 relit images, produces visually faithful illumination sweeps, and achieves strong paired-scene fidelity on challenging low-light data. These results position CLID as a practical synthetic-data primitive for studying illumination-conditioned perception in robotics.

1. Introduction

Robots are often expected to operate in visually degraded environments such as tunnels, mines, caves, industrial interiors, and other low-light spaces. This is especially clear in subterranean robotics settings, exemplified by the DARPA Subterranean Challenge, where systems must perceive, navigate, and search despite darkness and severe visibility constraints. In practice, a common engineering response is simple: attach an onboard light and increase the output until the scene becomes visible enough for downstream perception. That intervention is effective, but it also changes the image formation process in ways that matter for robot vision: added light can reveal useful structure, but it can also create glare, specularities, and saturation.

This creates an under-studied active perception problem.

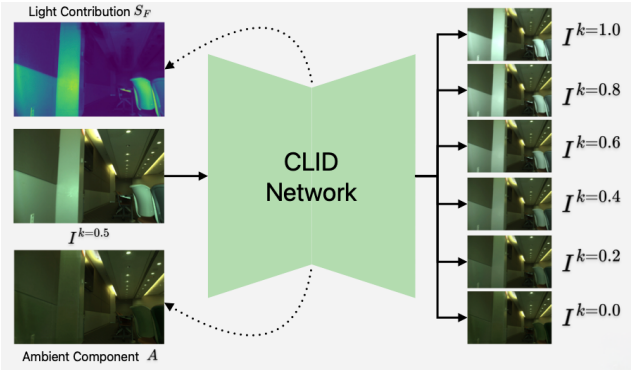

Figure 1. Overview of co-located relighting with CLID. Starting from an observation captured at a single illumination level $I^{k=0.5}$, the scene is represented in terms of an ambient component A and a light contribution S_F. Conditioned on these components, the CLID network synthesizes the scene appearance at target light intensities I^k, producing a dense set of relit images from a sparse capture.

For a mobile robot, illumination is not merely a passive imaging condition; it is a controllable variable that affects what geometry, texture, and failure modes are exposed to the camera as the platform moves. Unlike exposure control, onboard light changes the photons reaching the sensor, making it a powerful but hard-to-study intervention. A dense study of onboard-light variation requires repeated capture of the same scene or trajectory under many lighting settings while camera pose, exposure, and scene content remain fixed. That process is already slow for static scenes and becomes especially challenging for robotics data, where trajectories must be replayed and aligned. As a result, many low-light datasets sparsely sample illumination or avoid controlled onboard-light variation altogether. This paper studies a practical alternative: generate dense illumination variation synthetically from sparse real captures. We focus on a common imaging regime in which the camera carries a co-located light source whose intensity can be varied. This setup is narrower than unconstrained relighting, but it is precisely the structure that makes it valuable for robotics and dataset construction: the source geometry is fixed, the control variable is one-dimensional, and the desired output is a bank of counterfactual observations of the same underlying scene under different onboard-light settings. The main challenge

is to preserve scene structure from real measurements while synthesizing realistic photometric changes over that intensity sweep.

We address this problem with *co-located illumination decomposition*. Given one observed image at source intensity k_s, CLID predicts an ambient image and a reusable source-dependent contribution. Images at arbitrary target intensities are then produced analytically by recombination, so a sparse capture can be expanded into a dense synthetic stack without reacquiring the scene. This ties the generation process to additive flash/ambient image formation [3, 4, 8, 12, 14, 15] while remaining much more structured than generic relighting networks [21, 22].

Our contribution is therefore not a new active-control policy, but a synthetic data generation primitive for studying controlled illumination variation. We preserve the current CLID method and evidence, but frame them around sparse-to-dense illumination synthesis: CLID is trained on sparse real illumination captures, then used to synthesize dense intensity sweeps for static scenes and robot trajectories. This makes it possible to evaluate how different onboard-light settings would affect the same environment without repeatedly recollecting data, which is particularly valuable for scalable synthetic-data generation in low-light robotics.

2. Related Work

Synthetic data pipelines often vary geometry, materials, or weather in simulation [7, 17, 18], but controlled illumination variation from real captures remains comparatively underexplored because repeated acquisition is expensive [13]. Our setting is motivated by low-light robotics with onboard lighting, where the practical answer is often to "just throw a light on" and tune the output until perception becomes usable. That makes illumination an actionable control variable, yet its effect on mobile robot perception is hard to study systematically because recollecting the same scene or trajectory at many source powers is tedious and poorly scalable. We therefore seek to expand sparse real measurements into dense synthetic illumination sweeps. Active illumination is relevant as an application domain [1, 2, 5, 6, 9–11, 19], but we use it here primarily as a robotics application that benefits from efficient, structured synthetic data.

The closest image-formation priors come from intrinsic decomposition and flash/ambient imaging. Classical intrinsic methods separate reflectance and shading [3, 4, 14], while flash/no-flash methods use additive ambient and flash components for enhancement or factorization [8, 12, 15]. CLID inherits that additive view, but uses it as a re-lightable latent representation for dataset expansion from sparse captures.

Learning-based relighting methods [21, 22] aim for broader illumination transfer across diverse source configurations. Our problem is narrower and more data-centric: the source geometry is fixed and only intensity changes, so the goal is efficient sparse-to-dense synthesis of a one-dimensional illumination family grounded in real sensor data. That restriction lets CLID infer a decomposition once and then generate dense synthetic samples analytically.

3. CLID for Synthetic Data Generation

We consider a fixed camera–light rig in which the onboard source is mounted close to the optical axis and only its intensity is controlled. The objective is to turn sparse real illumination measurements into a dense synthetic family of images for the same scene while camera pose and exposure remain fixed. CLID addresses this problem through a compact decomposition that separates ambient appearance from the source-dependent contribution of the onboard light.

3.1. Co-Located Illumination Decomposition

We formulate relighting in a co-located camera–light setup as a relaxed intrinsic decomposition problem, which we term *Co-Located Illumination Decomposition* (CLID). Our goal is not full inverse rendering under arbitrary lighting, but a compact and controllable representation that supports sparse-to-dense synthesis as the intensity of a fixed onboard light source is varied.

We begin from the classical intrinsic-image model

$$I = R \odot S, \tag{1}$$

where I is the image, R is reflectance, and S is shading [3, 4, 14]. In the presence of a co-located light source, flash image formation can be written as

$$I = R \odot (c_A S_A + S_F), \tag{2}$$

where S_A and S_F denote ambient and flash shading, respectively, and $c_A \in \mathbb{R}^{1\times 3}$ is the ambient illuminant color [12]. Under illumination superposition in linear sensor space, this can be rewritten as

$$I = A + F, \tag{3}$$

where $A = R \odot c_A S_A$ is the ambient image contribution and $F = R \odot S_F$ is the contribution of the co-located source.

CLID builds on Equations (1) to (3) but relaxes the factorization to better suit controllable relighting. Rather than estimating reflectance and shading separately, we predict an ambient image and a co-located light contribution from a single conditioned source view (I^{k_s}, k_s), and supervise the resulting decomposition against all target intensities available for that scene. Concretely, CLID predicts an ambient image $A \in \mathbb{R}^{H\times W\times 3}$, a scalar light-contribution map $S_F \in \mathbb{R}^{H\times W\times 1}$, and a global color vector $C_F \in \mathbb{R}^{1\times 3}$. We then synthesize the appearance at any target intensity $k_t \in \mathcal{K} \subseteq [0, 1]$ via

$$\hat{I}^{k_t} = A + k_t \cdot (S_F \odot C_F), \tag{4}$$

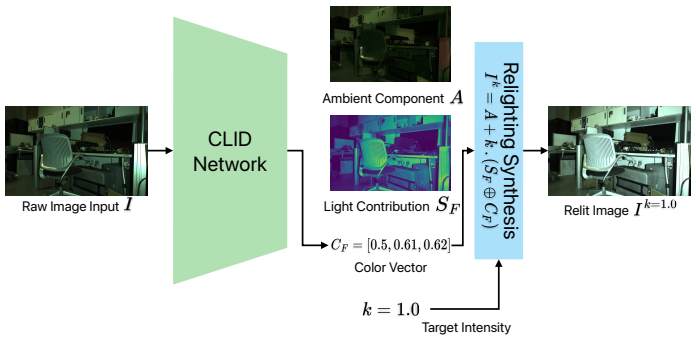

Figure 2. **CLID overview.** From one source view I^{k_s}, CLID predicts an ambient image A, a scalar light-contribution map S_F, and a global source-color vector C_F. Any target intensity k_t is then synthesized analytically as $\hat{I}^{k_t} = A + k_t(S_F \odot \bar{C}_F)$, so one sparse capture can be expanded into a dense synthetic illumination sweep.

where S_F is broadcast across color channels and C_F is the light-color vector used during synthesis.

This parameterization is deliberately compact. The map S_F captures where the co-located source contributes, while C_F captures its global chromaticity. Unlike flash formulations that assume white-balanced or achromatic light [12], we model source color explicitly. The ambient term A is a relightable ambient image, not a true albedo map: it absorbs reflectance and ambient shading under the same fixed exposure. All operations in Equation (4) are carried out in linear sensor space; our custom captures use normalized `BayerRG16` measurements without demosaicing or white balancing so the additive model remains meaningful for synthetic data generation from real sensor values.

3.1.1. Network Architecture

To estimate A, S_F, and C_F, we use a four-scale U-Net encoder–decoder [16]. The input is a linearly normalized sensor image concatenated with the source intensity k_s as a broadcast fourth channel. The encoder and decoder use standard convolution, pooling, transposed-convolution upsampling, and skip connections; the final decoder stage outputs $Y \in \mathbb{R}^{H \times W \times 4}$, whose first three channels correspond to the ambient RGB image A and whose final channel corresponds to the scalar light map S_F, i.e., $Y = [A \;\; S_F]$. We apply ReLU to both outputs to preserve non-negativity in the additive model. The color vector $C_F \in \mathbb{R}^3$ is a global learnable parameter optimized jointly with the network. At inference, the network is evaluated once per source image to estimate (A, S_F, C_F), after which arbitrary target intensities are synthesized analytically through Equation (4).

3.1.2. Training and Supervision

We train CLID in two stages. We first pre-train on the MIT Multi-Illumination dataset [13], treating direction 0 as ambient and direction 2 as the co-located-light endpoint, then fine-tune on our custom dataset of ambient observations and multiple discrete onboard-light levels under fixed exposure. Each training sample contains the full intensity stack for one scene; we choose one source intensity $k_s \in \mathcal{K}$, predict a decomposition from I^{k_s}, and supervise the synthesized outputs against all target intensities in the stack.

Our objective combines masked reconstruction, ambient reconstruction, zero-intensity consistency, delta-consistency, monotonicity, and smoothness terms. For image supervision we use

$$\ell(X, Y) = \ell_1^{\text{mask}}(X, Y) + \lambda_{\text{ssim}}\big(1 - \text{SSIM}(X, Y)\big),$$

where ℓ_1^{mask} excludes saturated target pixels. The main reconstruction terms are

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{K}|} \sum_{k_t \in \mathcal{K}} \ell(\hat{I}^{k_t}, I^{k_t}), \qquad \mathcal{L}_{\text{amb}} = \ell(A, I^0), \quad (5)$$

which supervise the relit stack and the ambient prediction. We additionally constrain the predicted low-to-high intensity difference, penalize residual light contribution when the target intensity is zero, enforce monotonicity across the sweep, and regularize S_F with total variation. In practice, pretraining disables the zero-intensity term because true ambient supervision is unavailable, and fine-tuning activates it once real ambient targets are present. We optimize with Adam and initialize fine-tuning from the pretrained checkpoint.

Discussion. The CLID formulation provides two properties that are especially desirable for synthetic dataset construction. First, it is physically grounded: the decomposition mirrors additive image formation by ambient and co-located light in linear sensor space, rather than treating relighting as unconstrained appearance translation. Second, it is efficient: once (A, S_F, C_F) are predicted from a single conditioned source view, dense multi-intensity synthesis is obtained by direct recombination, enabling structured supervision during training and practical expansion of sparse captures into relit image families at inference time.

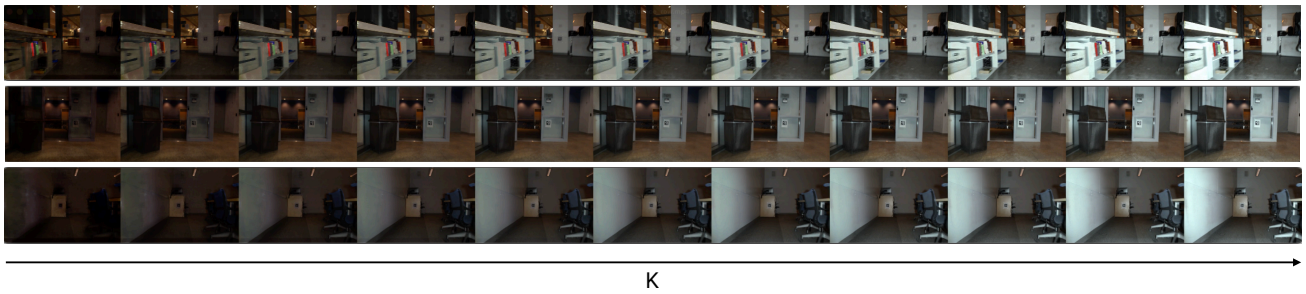

Figure 3. Qualitative CLID illumination synthesis on robot sequences. Each row starts from a captured source frame at 50% light output and synthesizes a dense 0–100% intensity sweep for the same trajectory. This is the intended data-generation use case: one real observation can be expanded into a bank of counterfactual onboard-light settings without re-running the robot.

4. Experimental Evaluation

Our evaluation asks whether CLID is useful as a synthetic-data generator for controlled illumination variation. A successful model should preserve scene structure from real sensor measurements while producing the intensity trends needed to populate dense relit datasets from sparse captures. We therefore focus on paired static scenes, where known targets at multiple illumination levels allow direct fidelity measurements, and additionally show qualitative relighting examples on real robot trajectories to illustrate the intended robotics deployment setting.

4.1. Experimental Setup

We pre-train on the MIT Multi-Illumination dataset [13] and fine-tune on custom data captured with a FLIR Blackfly S camera and a co-located COB LED under fixed exposure. All measurements are stored as linear `BayerRG16` RAW images. Our custom collection contains 160 static scenes with 5 light intensity levels and 15 robot sequences recorded at 10 FPS with 50% onboard light intensity. These 15 sequences contain 10,153 input images, which CLID expands into 101,530 relit images by sweeping 0–100% intensity in 10% increments, illustrating the efficiency of sparse-to-dense illumination synthesis for robotics. These sparse real captures define the supervision available during training; for a subset, we also capture real 0% and 100% reference runs.

For paired static scenes, we report PSNR and SSIM [20] together with relative luminance deviation.

4.2. Static Fidelity on Paired Scenes

Table 1 shows that CLID reconstructs target appearances with high fidelity across a range of source-to-target intensity changes, achieving an overall PSNR of 27.34 ± 5.53 dB and SSIM of 0.905 ± 0.020. Performance is strongest for moderate or high-intensity transitions that remain close to the training manifold, such as 50→100, where the luminance deviation drops to only $2.5 \pm 2.4\%$. Transitions to the ambient endpoint are more difficult, especially 50→0, because removing the direct source contribution requires recovering fine ambient structure from a brightly illuminated observation. Even so, the ambient-endpoint relighting remains visually coherent, with PSNR above 27 dB for both 50→0 and 100→0 transitions. Together with the qualitative robot-sequence results in Figure 3, this suggests that CLID is accurate enough to expand sparse captures into dense illumination-variation datasets for downstream robotics studies, while highlighting expected failure cases around ambient recovery, specularities, and saturation.

Table 1. Paired-reference fidelity for synthetic illumination generation on custom static scenes. We report mean ± std over all evaluated frames, and additionally break down results by source-to-target intensity transition.

Transition	PSNR ↑	SSIM ↑	Δ Lum. (%) ↓
All (mean ± std)	27.34 ± 5.53	0.905 ± 0.020	11.5 ± 3.8
25 → 50	25.81 ± 6.13	0.924 ± 0.018	11.6 ± 2.7
50 → 100	28.53 ± 3.41	0.920 ± 0.014	2.5 ± 2.4
100 → 50	25.49 ± 7.75	0.927 ± 0.028	9.9 ± 8.0
75 → 25	27.56 ± 8.49	0.927 ± 0.034	8.9 ± 9.5
50 → 0	29.13 ± 4.60	0.878 ± 0.026	21.5 ± 12.6
100 → 0	27.52 ± 4.80	0.856 ± 0.034	14.5 ± 11.9

5. Conclusion

We presented CLID as an efficient synthetic-data generator for controlled illumination variation. By decomposing one observation into an ambient image and a reusable light-dependent term, CLID turns sparse real captures into dense illumination-conditioned image families while staying grounded in the physics of a fixed co-located camera–light rig. The resulting relit data is faithful enough to support illumination-conditioned robotics studies without exhaustively recollecting every scene under every light setting. The main limitations remain ambient-endpoint recovery and specular or saturated regions, but the results still support co-located illumination decomposition as a practical synthetic-data primitive for low-light vision.

References

[1] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988. 2

[2] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 2

[3] Jonathan T. Barron and Jitendra Malik. Shape, Illumination, and Reflectance from Shading, 2020. arXiv:2010.03592 [cs]. 2

[4] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 2

[5] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S. Sukhatme. Interactive Perception: Leveraging Action in Perception and Perception in Action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017. 2

[6] Francesco Crocetti, Raffaele Brilli, Alberto Dionigi, Mario L. Fravolini, Gabriele Costante, and Paolo Valigi. Comparison of DSO and ORB-SLAM3 in Low-Light Environments With Auxiliary Lighting and Deep Learning Based Image Enhancing. *Journal of Field Robotics*, page rob.22595, 2025. 2

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[8] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM transactions on graphics (TOG)*, 23(3):673–678, 2004. 2

[9] Olivier Gamache, Jean-Michel Fortin, Matěj Boxan, François Pomerleau, and Philippe Giguère. Reproducible Evaluation of Camera Auto-Exposure Methods in the Field: Platform, Benchmark, and Lessons Learned. *IEEE Transactions on Field Robotics*, 2:270–287, 2025. 2

[10] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3D reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484, 2016.

[11] Joowan Kim, Younggun Cho, and Ayoung Kim. Exposure Control Using Bayesian Optimization Based on Entropy Weighted Image Gradient. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 857–864, Brisbane, QLD, 2018. IEEE. 2

[12] Sepideh Sarajian Maralan, Chris Careaga, and Yağiz Aksoy. Computational Flash Photography through Intrinsics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16654–16662, Vancouver, BC, Canada, 2023. IEEE. 2, 3

[13] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A Dataset of Multi-Illumination Images in the Wild, 2019. arXiv:1910.08131 [cs]. 2, 3, 4

[14] Takuya Narihira, Michael Maire, and Stella X. Yu. Direct Intrinsics: Learning Albedo-Shading Decomposition by Convolutional Regression, 2015. arXiv:1512.02311 [cs]. 2

[15] Georg Petschnigg, Maneesh Agrawala, Hugues Hoppe, Richard Szeliski, Michael Cohen, and Kentaro Toyama. Digital Photography with Flash and No-Flash Image Pairs. 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[17] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[18] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 2

[19] Yash Turkar, Youngjin Kim, and Karthik Dantu. Active illumination control in low-light environments. In *International Symposium on Experimental Robotics (ISER)*, 2025. To appear. Preprint available at https://arxiv.org/abs/2506.06394. 2

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[21] Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. LumiNet: Latent Intrinsics Meets Diffusion Models for Indoor Scene Relighting, 2024. arXiv:2412.00177 [cs]. 2

[22] Xiao Zhang, William Gao, Seemandhar Jain, Michael Maire, David A. Forsyth, and Anand Bhattad. Latent Intrinsics Emerge from Training to Relight, 2025. arXiv:2405.21074 [cs]. 2