Computational Flash Photography through Intrinsics

Sepideh Sarajian Maralan   Chris Careaga   Yağız Aksoy
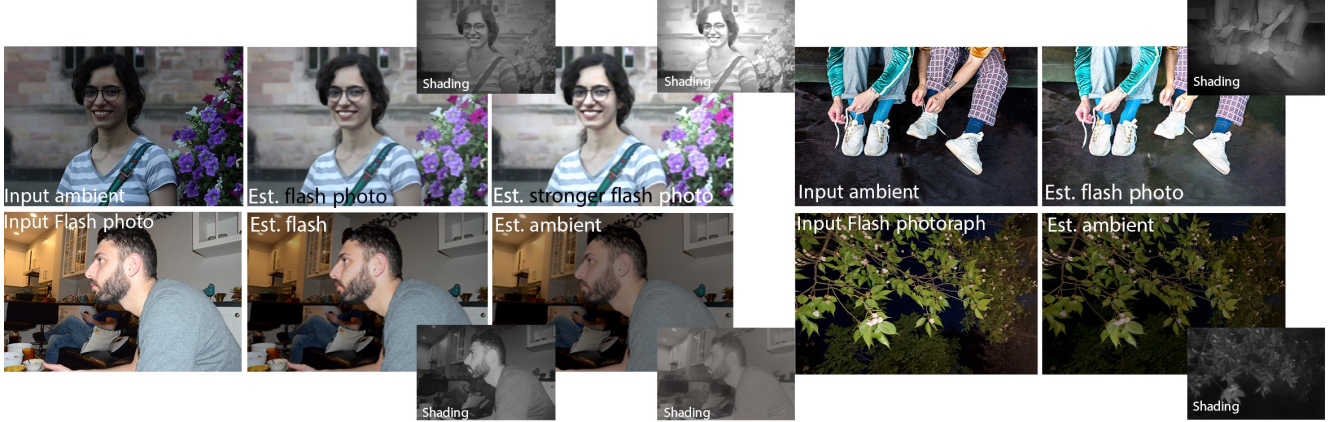
Simon Fraser University

Figure 1. We develop a system to computationally control the flash light in photographs originally taken with or without flash. We formulate the flash photograph formation through image intrinsics, and estimate the flash shading through generation for no-flash photographs (top) or decomposition where we separate the flash from the ambient illumination for flash photographs (bottom).

Abstract

Flash is an essential tool as it often serves as the sole controllable light source in everyday photography. However, the use of flash is a binary decision at the time a photograph is captured with limited control over its characteristics such as strength or color. In this work, we study the computational control of the flash light in photographs taken with or without flash. We present a physically motivated intrinsic formulation for flash photograph formation and develop flash decomposition and generation methods for flash and no-flash photographs, respectively. We demonstrate that our intrinsic formulation outperforms alternatives in the literature and allows us to computationally control flash in in-the-wild images.

1. Introduction

Flash is an essential tool as it often serves as the sole controllable light source in everyday photography available on a wide range of cameras. Although most cameras measure the environment light to adjust the flash strength, whether to use flash or not is a binary decision for the casual photographer without any control over flash characteristics such as strength or color temperature. In addition to hardware constraints, this limitation also stems from the typical use case of flash in situations that do not allow lengthy experi-

mentation with illumination. This makes the computational control of the flash a desirable application in photography.

Using the superposition of illumination principle, a flash photograph can be modeled as the summation of flash and ambient illuminations in linear RGB:

$$P = A + F, \tag{1}$$

where P, A, and F represent the flash photograph, the ambient illumination, and the flash illumination respectively. To computationally control the flash post-capture, the ambient and flash illuminations need to be decomposed. Similarly, to *add* flash light to a no-flash photograph, the flash illumination needs to be generated. In this work, we tackle both flash illumination decomposition and generation problems and show that regardless of the original photograph being captured with or without flash, we can generate and control the flash photograph computationally.

We start by modeling the flash photograph formation through image intrinsics. Intrinsic image decomposition models the image in terms of reflectance and shading:

$$I = R \cdot S, \tag{2}$$

where I, R, and S represent the input image, the reflectance map, and the shading respectively. This representation separates the effects of lighting in the scene from the illumination-invariant surface properties. With lighting

Figure 2. We can model the flash photograph formation in terms of separate shading maps for flash and ambient lights with a shared albedo.

being the only difference between the ambient and flash illuminations, both the decomposition and generation problems are more conveniently modeled on shading. Assuming a mono-color ambient illumination and a single-channel shading representation, we re-write Eq. 1 using Eq. 2:

$$P = R \cdot c_A S_A + R \cdot S_F = R \cdot (c_A S_A + S_F), \quad (3)$$

where c_A is a 3-dimensional vector representing the color temperature of the ambient illumination assuming the flash photograph P is white-balanced for the flash light. Fig. 2 visualizes our representation.

We develop a system for flash decomposition where we estimate c_A, S_A, and S_F from a flash photograph and combine them with the shared reflectance R to compute the final flash and ambient illumination estimations. We show that this physically-inspired modeling of the flash photograph allows us to generate high-quality decomposition results.

Flash decomposition is an under-constrained problem where both A and F are partially observed within the input P. Flash generation, on the other hand, is a more challenging problem that amounts to full scene relighting as, in this scenario, only A is given as input for estimating F. In addition to our intrinsic modeling, we formulate a self-supervision strategy for the generation problem through the cyclic relationship between decomposition and generation. We show that with a combined supervised and self-supervised strategy, we can successfully simulate flash photographs from no-flash input.

2. Related Work

Earlier work on flash photography mainly focused on using flash/no-flash pairs for image enhancement. The pair of photographs, although challenging to capture in real-world setups, have been shown to be useful for reasoning about the ambient light characteristics [7], image enhancement and denoising [9], artifact removal [1, 19], image matting [25], white balancing [15], and saliency estimation [13].

In recent years, multiple works [2, 3, 18] generated datasets of flash/no-flash image pairs. These datasets with innovative methods of image-to-image translation paved the way for the generation or decomposition of flash illumination. [2, 3] and, more recently, [5] tackled the problem of image relighting with flash illumination which is also the focus of our work. Aksoy et al. [2] proposed a baseline network setup for flash illumination decomposition, Capece et

al. [3] presented a system to convert a flash photograph into a studio illumination, while Chavez et al. [5] utilized a conditional adversarial network with attention maps to translate from flash to ambient image. Due to the correlation of flash light with scene depth, depth maps can be used to guide the networks for flash decomposition and subsequently for image relighting [20]. In this work, we model the flash photograph formation through image intrinsics and show that our physically-inspired formulation outperforms the methods formulated as direct image-to-image translation.

Flash generation or decomposition can both be formulated as relighting problems. Image relighting is typically formulated with specific constraints such as portrait relighting or relighting of indoor/outdoor environments. These case-specific relighting methods often can not be directly applied to flash photography. For relighting of indoor and outdoor environments, Helou et al. present VIDIT [10], a synthetic dataset with different varying directions and temperatures. Multiple works are trained using this dataset for one-to-one and any-to-any illumination transfer [6, 11, 14, 30]. Zhu et al. [32] propose a specialized network that utilizes geometry information to perform relighting as a direct image-to-image problem. Similar to our approach, intrinsic decomposition components have been used as an intermediate representation in relighting approaches. Yazdani et al. [30] construct the relit image by estimating the albedo and shading of the target. Wang et al. [28] estimate a shadow-free image generated from multiple illuminations of a single scene. Both methods attempt to predict intrinsic components and relight simultaneously, which makes it a very difficult problem for networks with limited capacity. In our formulation, we utilize intrinsics from an off-the-shelf method and use the intrinsic formulation to simplify the relighting process for the network, allowing us to generate higher-quality results.

3. Flash Illumination Decomposition

We define the illumination decomposition task as estimating the separate flash and ambient illuminations given a flash photograph. In flash photography, most cameras auto-white balance according to the flash light, therefore the flash illumination appears white. Since most environments are under lighting that has a different color than flash, the ambient illumination appears with a color shift from the flash il-
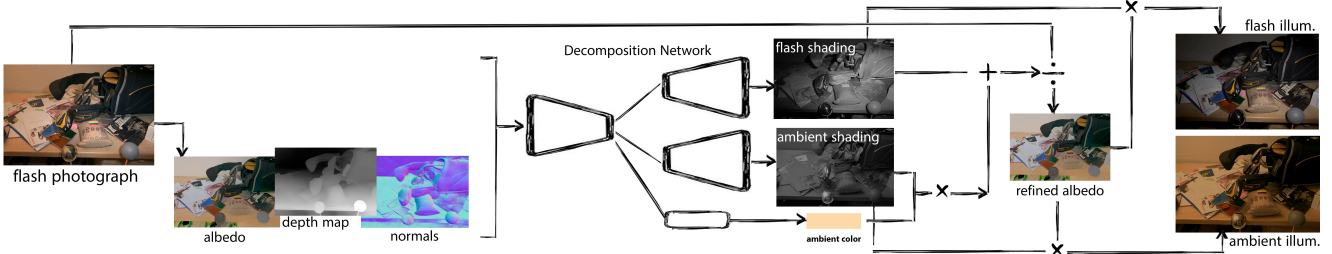
Figure 3. We provide the encoder with the flash photograph, as well as the albedo, surface normals, and depth map obtained through preprocessing. Two decoders and fully connected layers output the estimated flash shading, ambient shading, and ambient temperature. The ambient illumination is the element wise multiplication of the colored ambient shading with the refined albedo.

lumination. Assuming a mono-color ambient illumination, we model the flash photograph formation in the intrinsic domain using Eq. 3, as visualized in Fig. 2.

Following our model, we re-frame the illumination decomposition problem as a shading prediction task. We propose a pipeline to generate shading components that correspond to the two illumination sources from a flash photograph. Our decomposition pipeline consists of a single encoder that feeds into three different decoders. Two decoders generate the single-channel ambient and flash shadings S_A and S_F, while the third uses linear layers to regress the color temperature of the ambient illumination that defines c_A. We present an overview of our pipeline in Fig. 3.

Inputs Since we define the decomposition problem in the shading domain, we also provide the intrinsic decomposition of the input flash photograph in terms of an albedo map. We generate this input using an off-the-shelf intrinsic decomposition method [4]. The input image and this albedo map together provide the information on the combined shading maps of ambient and flash, as these variables are related through Eq. 2. Since this method estimates a single-channel shading for the flash photograph, the color shift of the ambient illumination is baked into this albedo map, as Fig. 4. Since the amount of this color shift is directly related to the strength of the ambient illumination at a given pixel, the input albedo also provides important cues for the relative strength of ambient and flash shadings.

The flash light can be modeled as a point light source close to the camera center, with its direction parallel to the main camera axis. This makes the strength of the flash illumination at a given pixel highly correlated to the distance of that pixel to the camera following the inverse-square law of light. Similarly, it is also correlated with the surface normals in the scene which defines the incident angle of the flash light. We provide this geometric information to our network in the form of monocular depth [12] and surface normal [8] estimations using off-the-shelf methods.

Outputs The three decoders in our system generate estimates representing the scene illumination, $\hat{S}_A$, $\hat{S}_F$, and $\hat{c}_A$. Instead of estimating the 3-dimensional ambient illumination color $\hat{c}_A$ directly, we estimate the color temperature of

Figure 4. We are able to disentangle the color of ambient illumination from the albedo of a flash photograph.

the environment, $\hat{t}_a$, and compute $\hat{c}_a$ from it. This makes the color estimation problem single-dimensional while covering the plausible illumination conditions in natural scenes. Given these variables, using Eq. 3, we can compute the common albedo:

$$\hat{R} = \frac{P}{\hat{c}_A \hat{S}_A + \hat{S}_F}. \tag{4}$$

We then compute the individual flash-only and ambient-only illuminations using the intrinsic model:

$$\hat{F} = \hat{R} \cdot \hat{S}_F, \qquad \hat{A} = \hat{R} \cdot \hat{c}_A \hat{S}_A. \tag{5}$$

Loss functions We define two losses on the estimated shadings as well as the albedo derived from them in Eq. 4 for training our system. While the individual losses on shadings supervise each decoder separately, the loss on the albedo is backpropagated through all three decoders and ensures that the individual components combine well together to reconstruct the scene.

The first loss we define is the standard L_1 loss:

$$\mathcal{L}_{1,X} = \frac{1}{N} \sum_{i=1}^{N} |\hat{X}_i - X_i^*|, \quad X \in \{S_A, S_F, R\}, \tag{6}$$

where X^* represents the ground truth. We also use the multi-scale gradient loss proposed by Li and Snavely [16] on these variables. This loss enforces spatial smoothness as
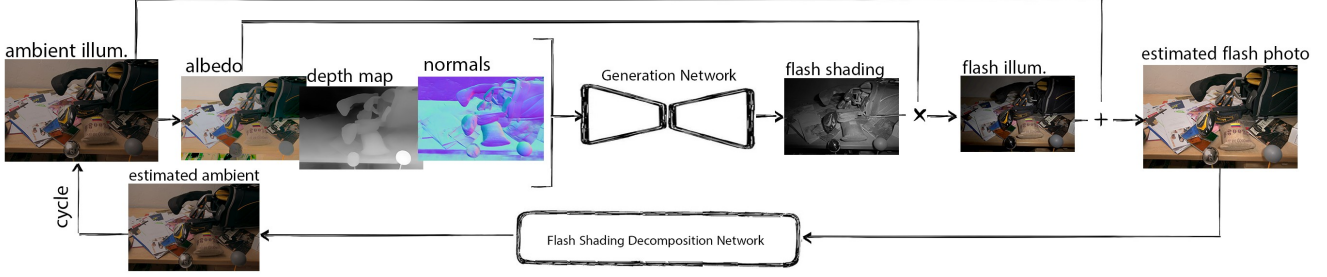
Figure 5. Albedo, normal, and depth maps are estimated from the ambient image and provided to the encoder-decoder network. The network produces the flash shading, which is then used to generate the estimated flash photograph. Self-supervision is also done by cycle consistency with the pre-trained flash shading decomposition network.

well as sharp discontinuities guided by the ground truth:

$$\mathcal{L}_{g,X} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} |\nabla X_{i,l} - \nabla X_{i,l}^*|. \quad (7)$$

We use the standard L_1 loss on the color temperature, $\mathcal{L}_T = |\hat{t}_A - t_A^*|$, and combine all 7 losses to train our decomposition system:

$$\mathcal{L}_D = \mathcal{L}_{1,S_A} + \mathcal{L}_{1,S_F} + \mathcal{L}_{1,R} + \\ 0.5 \left(\mathcal{L}_{g,S_A} + \mathcal{L}_{g,S_F} + \mathcal{L}_{g,R} \right) + \mathcal{L}_T. \quad (8)$$

Network structure   We base our network on the encoder-decoder architecture proposed in [21]. We use the Efficient-Net [27] encoder and a pair of RefineNet [17] decoders for flash and ambient shadings. The third decoder for the ambient color temperature uses global average pooling to flatten the activations from the encoder. The result is fed through 3 linear layers separated by ReLU activations. A final sigmoid activation outputs the color temperature. All components are trained from scratch with no pre-training with the learning rate of 2×10^{-4} for 100 epochs.

4. Flash Illumination Generation

Similar to our decomposition formulation, we model our flash illumination generation method from a no-flash photograph in the intrinsic domain and estimate the single-channel flash shading. We supplement the supervised training with a self-supervised loss by taking advantage of the cyclic relationship between the generation and decomposition tasks. Fig. 5 shows an overview of our pipeline.

Inputs   Similar to our decomposition setup, we estimate the albedo, the monocular depth, and the surface normals from the input no-flash photograph, which is the same as the ambient illumination. The flash illumination depends on the geometry and the albedo and is independent of the ambient shading as Eq. 3 shows. Once these variables are extracted, the original input does not provide additional useful information for the generation task. Hence, we exclude it from the inputs to our generation network.

Outputs   Our generation architecture only estimates the flash shading $\hat{S}_F$. Assuming a no-flash photograph white balanced for the ambient illumination as input A, we can compute the flash illumination using the estimated shading and the input albedo and also generate a flash photograph for any ambient illumination color c_A:

$$\hat{F} = R \cdot \hat{S}_F, \qquad \hat{P} = \hat{F} + c_A A. \quad (9)$$

Self-Supervision through Cycle Consistency   Given the ambient illumination A, our network can generate a flash photograph $\hat{P}$ via Eq. 9. Similarly, given a flash photograph P, the flash decomposition task can generate the ambient illumination $\hat{A}$ via Eq. 5. This exposes a cyclic relationship between the two tasks, where the two networks cascaded should re-create the input image. We exploit this relationship to define an L_1 cyclic consistency loss [31] using a forward pass through decomposition network $\mathcal{D}(\cdot)$:

$$\mathcal{L}_{cyc} = \frac{1}{N} \sum_{i=1}^{N} |\mathcal{D}(\hat{P})_i - A_i|. \quad (10)$$

When the decomposition network is at its initial stages of training, this loss is not very reliable given the unstable $\mathcal{D}(\hat{P})$ in Eq 10. Hence, we first train our decomposition task to completion before starting to train our generation network. Although this cyclic relationship can also be exploited in the other direction for the decomposition task, due to the higher complexity of the generation problem, we observed that it did not improve our decomposition performance.

Losses and Network Structure   In addition to the cyclic consistency loss, we utilize L_1 and the multi-scale gradient losses defined in Eqs. 6 and 7 on the estimated shading $\hat{S}_F$ as well as the corresponding flash illumination $\hat{F}$:

$$\mathcal{L}_G = \mathcal{L}_{1,S_F} + \mathcal{L}_{1,F} + 0.5 \left(\mathcal{L}_{g,S_F} + \mathcal{L}_{g,F} + \mathcal{L}_{cyc} \right). \quad (11)$$

We adopt the same network architecture as our decomposition task described in Sec. 3, with the only difference being the use of a single decoder as only a single shading map is estimated in the generation task.

Figure 6. Our dataset consists of flash/no-flash pairs from FAID [2], MID [18] and DPD [3] composited on background images from FAID.

5. Dataset Preparation and Augmentation

We require matching pairs of flash/no-flash pairs alongside their intrinsic decomposition components to train our networks. The flash decomposition and generation are challenging tasks that would benefit from a large training set. However, the availability of flash/ambient illumination pairs is limited due to the challenging capture process, and the number of input/ground-truth pairs is low when compared to datasets aiming at other relighting tasks such as facial relighting. Hence, we combine the 3 available datasets for flash illumination, namely the Flash and Ambient Illuminations Dataset (FAID) [2], the Deep Flash Portrait Dataset (DPD) [3] and the Multi-Illuminations Dataset (MID) [18] and modify them to meet our needs. The FAID dataset contains pairs of flash/no-flash images, the DPD dataset includes portraits captured with and without flash in front of a green screen, and the MID dataset has 25 lighting directions for each scene. To generate realistic images from the green-screen data in DPD, we apply green-screen keying and add flash/no-flash background pairs to each illumination taken from the Rooms set in FAID. Examples from each dataset are shown in Fig. 6. The DPD is composited onto various room environments to create realistic-looking images, and in MID, the lighting from the front of the camera is selected as the flash illumination, while the other illuminations serve as the no-flash images. The brightness of flash and no-flash illuminations are normalized across the datasets. For the intrinsic components, we use the MID-Intrinsics dataset extension by [4] for MID and use the robust multi-illumination intrinsic decomposition approach [4] to generate the albedo and shading pairs for FAID and DPD. We present a detailed description of our dataset pre-processing pipeline in the supplementary material.

6. Experimental Analysis

To show the effectiveness of our proposed approach, we perform an ablation over the various design choices of our method, as well as comparisons to various prior works. We
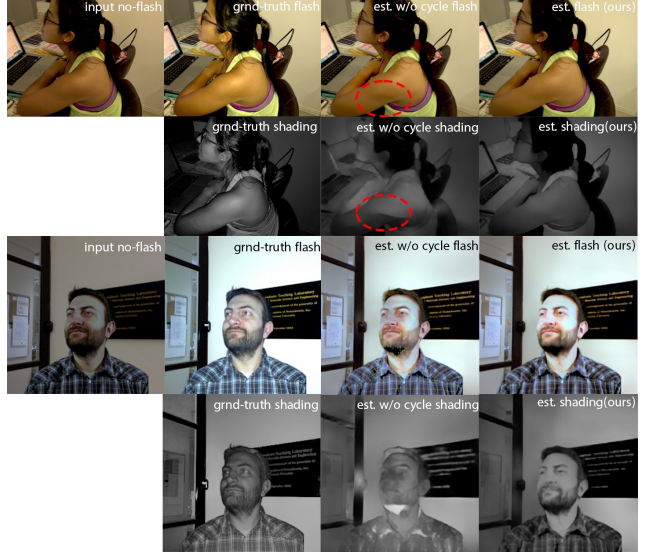

Figure 7. We utilize the cycle consistency loss for generation. The effect of this loss is evident in the shadings depicted by the red circle in the first example, as well as on the man's face in the bottom.
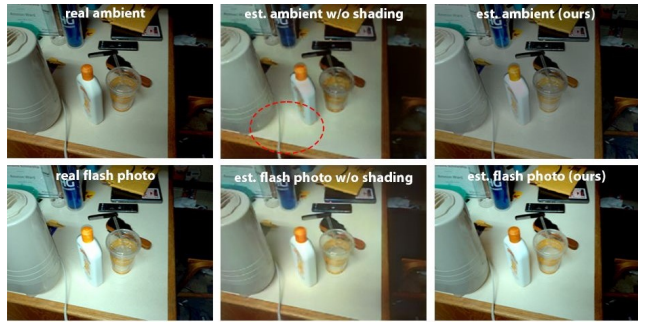

Figure 8. By estimating the shadings instead of estimating the illumination, the decomposition is incomplete in certain areas shown in red circles and the visual quality of estimation degrades.

utilize three common metrics for evaluating performance on both flash decomposition and generation. Namely, structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and Inception Score (IS) [24]. PSNR uses mean-square error to do a pixel-by-pixel comparison between the estimation and ground truth, while SSIM and IS focus on image quality. SSIM utilizes structure, luminance, and contrast to estimate perceived similarity to the ground truth. IS measures high-level image generation quality using an Inception-v3 image classifier [26] network trained on ImageNet [23].

6.1. Ablation Study

The top half of Table 1 shows the quantitative results of our ablation tests for 200 images containing a variety of objects, scenes, and people never seen before by our methods. Each row shows our pipeline with a single aspect missing.

We first examine the efficacy of our proposed intrinsic formulation. We train a version of our pipeline that di-
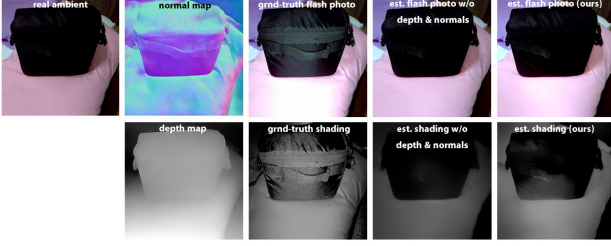
Figure 9. The monocular depth and surface normals provide geometric details that contribute to more accurate flash illumination.

Figure 10. We utilize the multi-scale gradient loss to create sharper edges and smoother gradients in the flash shading estimations.

Table 1. Quantitative performance of different ablation configurations and alternative methods on PSNR, SSIM and IS [24].

Method	Decomposition			Generation		
	SSIM↑	PSNR↑	IS [24]↑	SSIM↑	PSNR↑	IS [24]↑
Ours	**0.8553**	19.6044	**7.2892**	**0.8601**	19.1863	**7.3730**
no albedo/shading	0.774	19.2296	6.0548	0.7352	14.5329	5.9953
no cycle loss	-	-	-	0.8597	19.2155	6.4777
no gradient loss	0.8096	16.2628	6.4617	0.8573	19.0036	6.5127
no geometry	0.8074	14.5056	6.4746	0.8584	19.1301	6.5307
DeepFLASH [3]	0.5744	13.2684	7.1276	-	-	-
FAID network [2]	0.8081	15.6040	7.0183	0.7830	14.1004	6.7445
Pix2pixHD [29]	0.7917	17.486	6.5667	0.6866	12.2330	5.74385
IAN [32]	0.8507	**21.4149**	7.0240	**0.8682**	19.9776	7.2694
OIDDR [30]	0.8547	20.8643	7.2035	0.8579	**20.4645**	7.1679

6.2. Comparison to Prior Work

Although no prior works attempt to generate flash illumination in such diverse input images, to thoroughly evaluate our proposed pipeline, we use our augmented dataset to train five networks from prior works for our generation and decomposition tasks. We train all the networks for 100 epochs and compare them qualitatively using the previously discussed metrics. The bottom half of Table 1 shows the quantitative comparison of these methods on the same test set as the ablation study.

We compare our approach to two image-to-image baseline networks. The baseline network utilized by Aksoy et al. [2] and Pix2PixHD [29] are both based on the original Pix2Pix UNet architecture [22]. These baselines exhibit severe artifacts, exemplifying the inherent difficulty of learning these tasks using a relatively small-scale dataset.

We follow the training scheme proposed by Capece et al. [3] and train their proposed model on our augmented dataset. To train the network, we had to adjust the beta parameters of the Adam optimizer, specifically, we set β_1 equal to 0.5. We failed to train this method for the generation task, in which it simply outputs the input without modification, and hence we exclude it from our evaluation. On the decomposition task, while their method obtains a competitive Inception Score, they perform poorly on low-level metrics measured against the ground truth. This indicates that while their outputs may be visually coherent, they are not able to separate flash and ambient lighting effects.

Zhu et al. [32] propose a direct estimation approach for relighting between two known lighting configurations. They propose multiple architectural improvements specifically designed to achieve better relighting. We find that due to their carefully designed network, they can achieve competitive quantitative results on the test set. However, given their direct approach to relighting, their model still produces noticeable artifacts on difficult scenes such as portrait photographs, as can be seen in the insets of Fig. 12.

Similar to our method, the work of Yazdani et al. [30] utilizes intrinsic components as part of their proposed pipeline. We are therefore able to supervise their model using our generated shading and albedo components. Unlike our

rectly regresses the final relit image for both the generation and decomposition tasks. This causes a significant drop in quantitative performance across all metrics. Without our intrinsic formulation, the network has to reproduce redundant RGB information that is present in the input image, greatly increasing the likelihood of artifacts. Our proposed approach on the other hand disentangles the contributions of reflectance and shading and the network, therefore, has a much easier job of estimating single-channel shading images. Fig. 8 shows some examples from our model trained with and without the intrinsic formulation.

For the generation task, we evaluate our approach without the proposed cycle loss. We observe that while the low-level metrics, PSNR and SSIM, remain the same, the cycle loss significantly degrades the perceptual quality of our generated flash as measured by the Inception Score. Since the cycle loss acts as a complementary form of supervision for the model, we can generate a more plausible flash. Some example outputs from our generation model with and without the cycle loss are shown in Fig. 7.

The multi-scale gradient loss allows us to generate more accurate shading maps at both sharp discontinuities and over smooth image regions. In Fig. 10 we show some examples of our shading estimations, with and without the multi-scale gradient loss for both generation and decomposition.

Naturally, geometry information is very informative for relighting tasks as lighting effects are directly correlated with geometry. We find that removing the depth and surface normal input from our formulation greatly affects performance in both generation and decomposition. Fig. 9 shows an example of our flash generation task with and without geometry information.
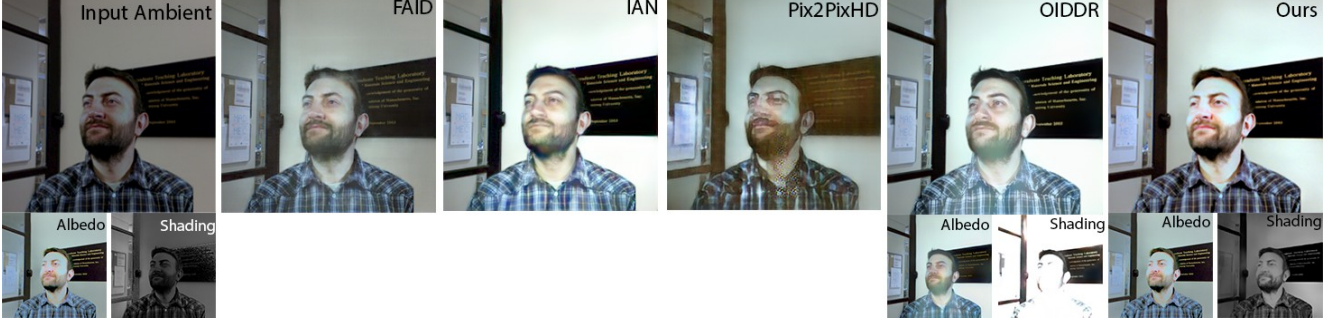
Figure 11. Comparison of generation task with different methods. We also show the generated albedo and shading from the OIDDR [30] which estimations are also done through intrinsics. Please refer to text for detailed discussions.

Figure 12. Comparison of decomposition task with different methods. Please refer to text for detailed discussions.

work, they propose to estimate both the shading and albedo of the relit output. Given that intrinsic decomposition is a notoriously difficult problem, this adds extra complexity to the relighting problem. Our method, on the other hand, utilizes albedo estimations as input, and in the case of the flash generation directly makes use of the albedo to compute the flash illumination. By only estimating grayscale shading, we significantly reduce the complexity of the task and generate stable, artifact-free estimations. Additionally, we can directly compare our predicted intrinsic components to theirs. In the generation task, we observe that their predicted albedo looks very similar to their final output, as seen in Fig. 11. Due to the sigmoid activation they use for shading, their approach fails to properly represent the high magnitude flash shading, causing the flash illumination to leak into the albedo. This indicates that their model is likely not benefiting from the intrinsic modeling to the full extent.

While the baseline approaches are not able to compete with our method quantitatively, the more recent methods IAN [32] and OIDDR [30] obtain competitive quantitative scores. There are multiple factors contributing to the discrepancy between quantitative scores and qualitative observations. Firstly, although our physically-inspired formulation generates more stable outputs, our modeling comes with some constraints. For example, in the generation task, our model relies on the input albedo and grayscale shad-

ing to generate the flash illumination. Oftentimes the mapping from ambient albedo to flash illumination cannot be expressed as a single per-pixel multiplier. This means that while our results are without artifacts, we may not be able to estimate the exact appearance of the ground-truth flash photograph. In the decomposition task, we are estimating the color temperature of the ambient illumination. While this design choice simplifies the shading estimation, a small deviation from the ground-truth color temperature can result in errors across the whole image. Despite these factors, our model still achieves scores that are higher or on par with recent relighting works in terms of SSIM and PSNR. Furthermore, our model outperforms all other competing methods in terms of Inception Score which correlates with the visual quality estimations. Given that we achieve these scores using off-the-shelf networks with no architectural improvements, we attribute our performance to our careful modeling of flash photography.

6.3. Computational Flash Photography

Our decomposition and generation results can be used to edit or simulate the flash light for flash or no-flash photographs. We show in-the-wild examples with simulated flash in Fig. 13. Fig. 14 shows our decomposition with the estimated shadings, while Fig. 15 shows an illumination editing example on an image originally captured on film.
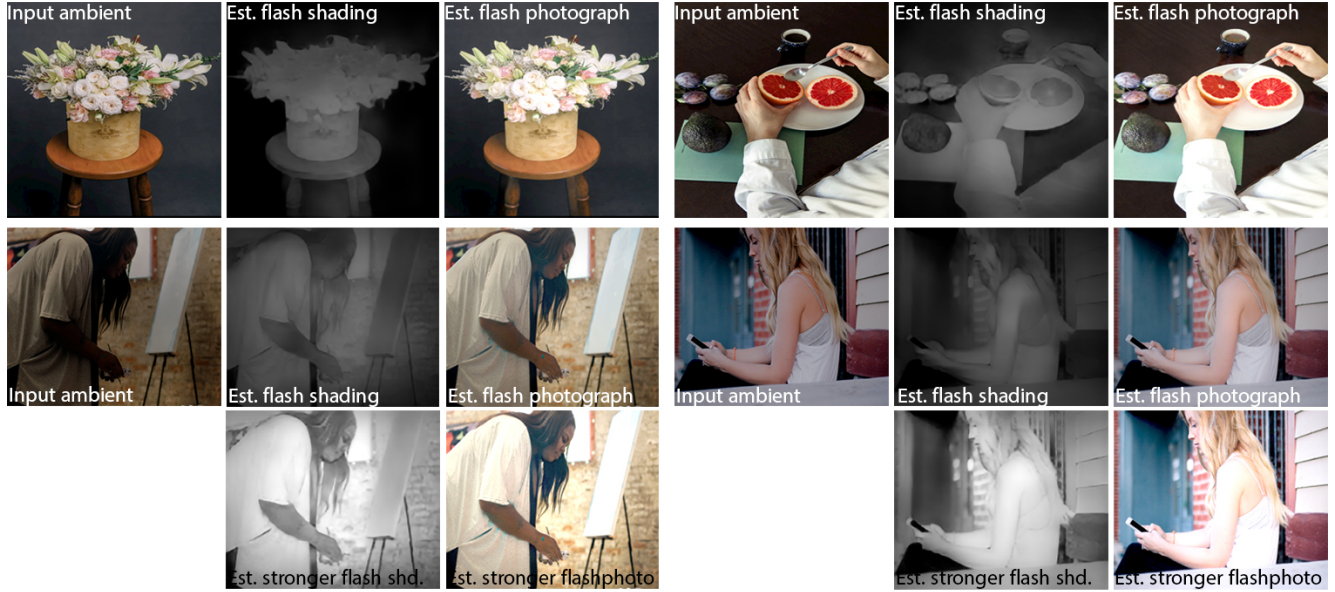
Figure 13. We show examples of our flash generation method on in-the-wild photographs with controllable flash strength.

Figure 14. Our intrinsic shading decomposition is able to decompose a flash photograph into its flash and ambient illuminations.
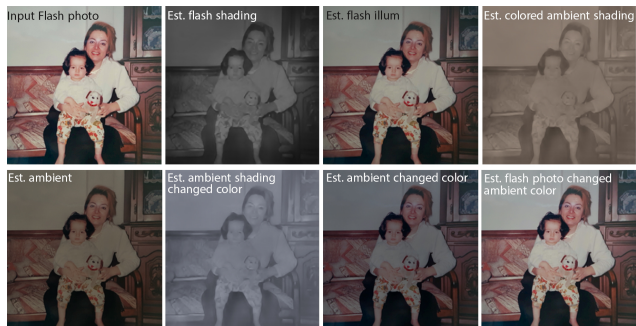

Figure 15. The decomposed ambient and flash illuminations allow us to control the ambient illumination color independent of the flash light.

7. Limitations

Flash generation and decomposition are challenging image relighting problems that require a large amount of training data. Even after combining 3 datasets available for our tasks, the volume of our dataset measures in a few thousand unique flash-ambient pairs. Similar tasks such as portrait relighting are typically achieved using datasets that are a magnitude larger, measuring in tens of thousands. This makes it hard to model challenging areas such as speculari-

ties. Still, our intrinsic formulation helps us generate high-quality results when compared to image-to-image translation under this limited training data. Our generation model takes the albedo, monocular depth, and surface normals as input. While, physically, this is the complete data necessary for flash generation, errors in the estimations of these variables may be carried to our final estimations.

8. Conclusion

In this work, we presented an intrinsic model for flash photograph formation and applied it to two tasks in computational flash photography: illumination decomposition and flash light generation. We show that, despite the limited training data, our physically-inspired formulation allows us to generate realistic results and control the flash light computationally in everyday photography. Our experimental evaluation demonstrates that, compared to image-to-image translation methods commonly utilized for similar tasks in the literature, our intrinsic formulation achieves higher-quality results.

References

[1] Amit Agrawal, Ramesh Raskar, Shree K. Nayar, and Yuanzhen Li. Removing photography artifacts using gradient projection and flash-exposure sampling. *ACM Trans. Graph.*, 2005. 2

[2] Yağız 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 *Proc. ECCV*, 2018. 2, 5, 6

[3] Nicola Capece, Francesco Banterle, Paolo Cignoni, Fabio Ganovelli, Roberto Scopigno, and Ugo Erra. DeepFlash: Turning a flash selfie into a studio portrait. *Signal Processing: Image Communication*, 2019. 2, 5, 6

[4] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition through ordinal shading. *SFU Tech. Rep.*, 2023. 3, 5

[5] José Chávez, Rensso Mora, and Edward Cayllahua-Cahuina. Ambient lighting generation for flash images with guided conditional adversarial networks. In *Proc. VISAPP*, 2020. 2

[6] Sourya Dipta Das, Nisarg A. Shah, Saikat Dutta, and Himanshu Kumar. Dsrn: an efficient deep network for image relighting. In *Proc. ICIP*, 2021. 2

[7] Jeffrey M DiCarlo, Feng Xiao, and Brian A Wandell. Illuminating illumination. In *Proc. Color and Imaging Conference*, 2001. 2

[8] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans. In *Proc. ICCV*, 2021. 3

[9] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph.*, 2004. 2

[10] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. VIDIT: Virtual image dataset for illumination transfer. arXiv:2005.05460 [cs.CV], 2020. 2

[11] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. AIM 2020: Scene relighting and illumination estimation challenge. In *Proc. ECCV Workshops*, 2020. 2

[12] S. Mahdi H. Miangoleh, Sebastian Dille, Mai Long, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proc. CVPR*, 2021. 3

[13] Shengfeng He and Rynson W. H. Lau. Saliency detection with flash and no-flash image pairs. In *Proc. ECCV*, 2014. 2

[14] Majed El Helou, Ruofan Zhou, Sabine Susstrunk, and Radu Timofte. NTIRE 2021 depth guided image relighting challenge. In *Proc. CVPR Workshops*, 2021. 2

[15] Z. Hui, A. C. Sankaranarayanan, K. Sunkavalli, and S. Hadap. White balance under mixed illumination using flash photography. In *Proc. ICCP*, 2016. 2

[16] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*, 2018. 3

[17] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. CVPR*, 2017. 4

[18] Lukas Murmann, Michaël Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proc. ICCV*, 2019. 2, 5

[19] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph.*, 2004. 2

[20] Di Qiu, Jin Zeng, Zhanghan Ke, Wenxiu Sun, and Chengxi Yang. Towards geometry guided neural relighting with flash photography. In *Proc. 3DV*, 2020. 2

[21] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 4

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, 2015. 6

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2014. 5

[24] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proc. NeurIPS*, 2016. 5, 6

[25] Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. Flash matting. *ACM Trans. Graph.*, 2006. 2

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. CVPR*, 2016. 5

[27] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, 2019. 4

[28] Li-Wen Wang, Wan-Chi Siu, Zhi-Song Liu, Chu-Tak Li, and Daniel PK Lun. Deep relighting networks for image light source manipulation. In *Proc. ECCV*, 2020. 2

[29] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. *Proc. CVPR*, 2018. 6

[30] Amirsaeed Yazdani, Tiantong Guo, and Vishal Monga. Physically inspired dense fusion networks for relighting. *Proc. CVPR Workshops*, 2021. 2, 6, 7

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, 2017. 4

[32] Zuo-Liang Zhu, Zhen Li, Rui-Xun Zhang, Chun-Le Guo, and Ming-Ming Cheng. Designing an illumination-aware network for deep image relighting. *IEEE Trans. Image Process.*, 2022. 2, 6, 7