A Dark Flash Normal Camera

Zhihao Xia[1*] Jason Lawrence[2] Supreeth Achar[2]
[1]Washington University in St. Louis [2]Google Research

Abstract

Casual photography is often performed in uncontrolled lighting that can result in low quality images and degrade the performance of downstream processing. We consider the problem of estimating surface normal and reflectance maps of scenes depicting people despite these conditions by supplementing the available visible illumination with a single near infrared (NIR) light source and camera, a so-called "dark flash image". Our method takes as input a single color image captured under arbitrary visible lighting and a single dark flash image captured under controlled front-lit NIR lighting at the same viewpoint, and computes a normal map, a diffuse albedo map, and a specular intensity map of the scene. Since ground truth normal and reflectance maps of faces are difficult to capture, we propose a novel training technique that combines information from two readily available and complementary sources: a stereo depth signal and photometric shading cues. We evaluate our method over a range of subjects and lighting conditions and describe two applications: optimizing stereo geometry and filling the shadows in an image.

1. Introduction

In casual mobile photography, images are often captured under poor lighting conditions. Controlling the visible lighting or supplementing it with a flash is often too difficult or too disruptive to be practical. On the other hand, the near infrared (NIR) lighting in a scene can be much more easily controlled and is invisible to the user. In this paper, we demonstrate how a single "dark flash" NIR image and a single visible image taken under uncontrolled lighting can be used to recover high quality maps of the surface normals, diffuse albedos, and specular intensities in the scene. We collectively refer to the albedo and specular intensity estimates as a "reflectance map". Exposing these signals within a photography pipeline opens up a range of applications from refining independent depth estimates to digitally manipulating the lighting in the scene. Although our method is applicable to many types of objects, we focus on faces - the most common photography subject at the short ranges over which active illumination is effective.

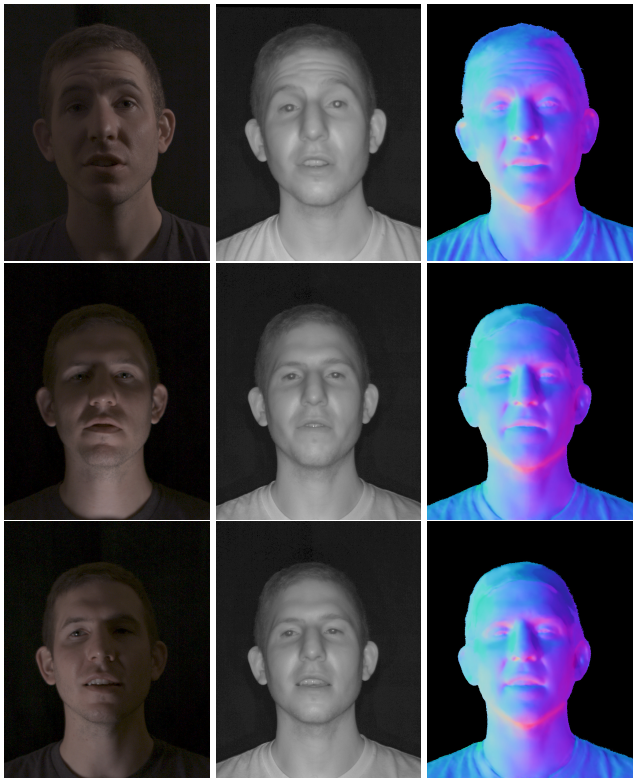


Figure 1: Estimating surface geometry from a single RGB image is challenging. We augment this input with a single NIR "dark flash" image captured at the same time, and present a network that can estimate high quality normal maps and reflectance maps (not shown) under a wide range of visible lighting conditions.

Our use of controlled NIR lighting provides a number of benefits. First, the ambient NIR light in a scene is usually weak or completely absent in indoor environments and is significantly attenuated by atmospheric absorption outdoors, which means it is often practical to control this aspect of a scene. Second, it results in a more tractable estimation problem in contrast to single-image "shape from

*Work done while Zhihao Xia was an intern at Google.

shading" and intrinsic image decomposition techniques that must simultaneously reason about shape, material properties, and lighting. Third, it provides a stable source of information about the shape and appearance of the scene even under very challenging visible lighting. By locating the NIR light source near the camera, this setup minimizes shadows in the scene while producing specular highlights along surfaces that are nearly perpendicular to the viewing direction, giving a useful cue for determining surface orientations.

We present a deep neural network that takes as input one RGB image captured under uncontrolled visible lighting and one monochrome NIR image captured from the same viewpoint, but under controlled lighting provided by a single source located near the camera. The network generates a surface normal and reflectance estimate (diffuse albedo + specular intensity) at each pixel. We train this network by combining two imperfect but complementary cues: a stereo depth map that provides a reliable estimate of the low-frequency components of the scene's 3d shape along with photometric cues that convey higher-frequency geometric details. These measurements are far easier to obtain than ground truth geometry and appearance measurements. We explicitly model the specular reflectance of human skin in a photometric loss term that guides our training along with a prior on the albedo map that favors piecewise constant variation [3].

We compare our technique to a baseline learning approach that uses only a single RGB image as input and state-of-the-art methods for single image intrinsic image decomposition [27] and relighting [21]. We are able to produce overall more stable and more accurate outputs even in very challenging visible light conditions. We also present two applications of integrating our technique in a mobile photography pipeline. In all, this paper makes the following contributions:

- A new network architecture for estimating dense normal and reflectance maps from a single RGB+NIR image pair.
- A new training strategy that combines two independent and complementary signals: one from stereo triangulation and the other from photometric cues in RGB and NIR, along with a hardware setup for collecting this data. Notably, our training is guided by a physically-based image formation model that reproduces both diffuse and surface reflectance.
- We demonstrate two applications of our method in a modern photography pipeline: optimizing depths computed by an independent stereo technique and reducing shadows in an image post-capture.

2. Related Work

Intrinsic imaging and shape from shading. Decomposing a single image into its underlying shape and reflectance is a classical under-constrained problem in computer vision [4, 15]. One class of methods employ hand-designed priors, learned from relatively small datasets [1, 3] or NIR imagery [11], to disambiguate these components. Learning-based methods have been proposed more recently that train convolutional neural networks to perform this task using rendered datasets [28, 18], sparse human annotations [6], or multi-view images under different lighting conditions [40]. Whereas some learning approaches function as "black boxes" [28], others incorporate a physically-based image formation model [5, 27, 18, 31]. Similiar to ours, other approaches explore network inputs beyond a single RGB image, including an additional visible flash image and a depth map [24] or a single NIR image [39].

A number of methods are specifically designed to work on images of faces. This includes 3D morphable models [7], which are commonly used as a prior on reflectance and geometry in learning-based approaches [32, 29, 27]. Sanyal et al. [26] estimate the shape of a face within a single image in the form of blending weights over a parametric face model. Similar to our approach, other techniques estimate dense normal or displacement maps [41] including for faces partially hidden by occluders [33, 13]. However these methods do not attempt to disentangle reflectance data from shading.

In contrast to these prior techniques, we propose a neural network that takes a single front-lit NIR image in addition to a color input image, enabling our technique to perform well even in very challenging visible light conditions. Our training process is also novel in the way that it combines two independent and complimentary signals.

Fusing depth and normals. Depth estimated from methods like stereo triangulation and normals estimated from shading cues are complementary measurements for shape recovery. Nehab et al. [20] describe a technique that seeks to combine the more accurate low-frequency information provided by direct depth measurement techniques with the higher-frequency geometric details provided by photometric measurements. We use their technique to evaluate how our approach could be used to improve a stereo pipeline (Section 5.2). More recent work poses this as an optimization problem that seeks a surface that best agrees with these different signals [2, 12, 38, 9, 19]. While our method does not use any depth information at inference time, our training method is similar to these approaches in that we also combine a stereo and photometric loss term.

Face relighting. Most single image face relighting methods include some representation of shape and reflectance as intermediate components. Our network architecture

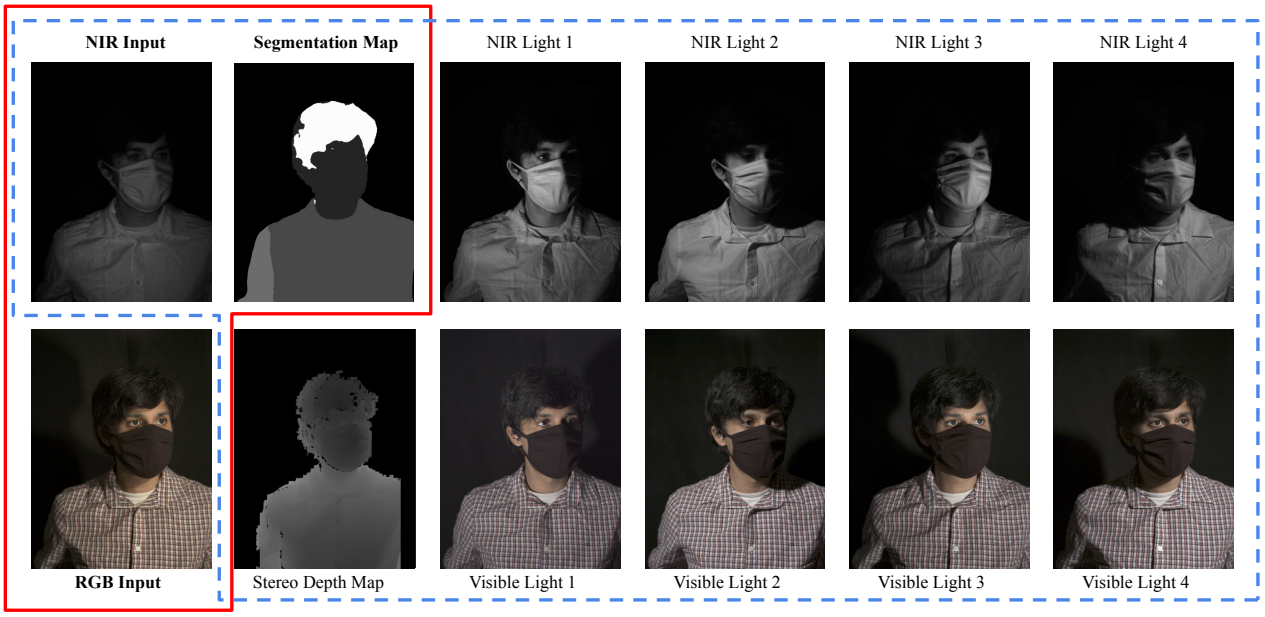


Figure 2: Our network learns to estimate shape and reflectance from a single front-lit NIR image, a single RGB image under arbitrary lighting, and a semantic segmentation map computed from the RGB image (inputs are enclosed by the red line). During training we also use a stereo depth map and replace the RGB image under arbitrary lighting with 4 RGB+NIR image pairs captured under calibrated point lights (the training inputs are inside the blue dashed line).

(Section 3) is similar to the one proposed by Nestmeyer et al. [21] for simulating lighting changes in a single image assumed to have been captured under a single directional light. Zhou et al. [42] present a dataset of relit portrait images generated using single-image normal and illumination estimates and a Lambertian reflectance model. Although surface geometry is fundamental to relighting, it is also possible to train an end-to-end network that does not explicitly reason about shape [30]. We similarly use multiple images of a scene captured under varying controlled lighting to train our network in order to enable a much simpler set of inputs for inference.

Combining infrared and color imagery. A NIR (and/or ultraviolet) dark flash image can be used to denoise a color image captured in low visible light conditions [17], or serve as a guide for correcting motion blur [37]. Techniques have also been developed that employ controlled NIR lighting to simulate better visible lighting in real-time video communication systems [34, 14]. We see these as compelling potential applications of this work.

3. Network Design and Training

Our goal is to estimate a normal map and a reflectance map from a single RGB image and a front-lit "dark flash" NIR image. We train a deep neural network to perform this task. As an auxiliary input, we use a 6-class semantic segmentation map computed from the RGB image (background, head, hair, body, upper arm and lower arm) [10]. We found this segmentation map was a useful cue for helping the network reason about shape and reflectance. An example set of inputs are shown in Figure 2 (red line).

Our training procedure is driven in part by a physically-based image formation model that connects the outputs of our network to images of a scene taken under known point lighting. This image formation model combines a standard Lambertian diffuse term with the Blinn-Phong BRDF [8], which has been used to model the specular reflectance of human skin [35]. Specifically, we introduce a reflectance function f that gives the ratio of reflected light to incident light for a particular unit-length light vector $\mathbf{l}$, view vector $\mathbf{v}$, surface normal $\mathbf{n}$, four-channel (RGB+NIR) albedo $\boldsymbol{\alpha}$, scalar specular intensity ρ, and specular exponent m:

$$f(\mathbf{l}, \mathbf{v}, \mathbf{n}) = \boldsymbol{\alpha} + \rho \frac{m+2}{2\pi} (\mathbf{n} \cdot \mathbf{h})^m, \tag{1}$$

where $\mathbf{h} = (\mathbf{n} + \mathbf{l})/\|\mathbf{n} + \mathbf{l}\|$. The observed intensity at a pixel due to a point light is given by

$$I(\cdot) = f(\mathbf{l}, \mathbf{v}, \mathbf{n})(\mathbf{n} \cdot \mathbf{l})L, \tag{2}$$

the product of the reflectance, cosine term, and light intensity L. We do not observe the reflected intensity from enough unique light directions at each pixel to estimate all

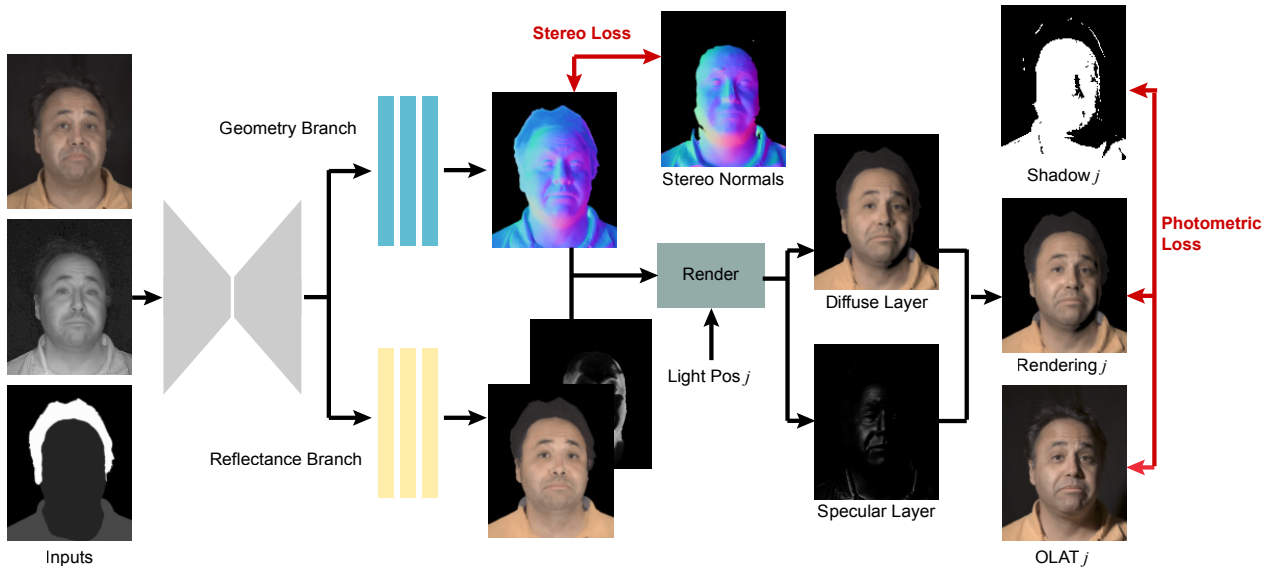


Figure 3: Illustration of our network and training strategy. We estimate network weights that minimize a photometric loss, computed between images rendered from our network outputs and ground truth images captured under known lighting, and a stereo loss, driven by differences between the output normals and those estimated using an independent stereo technique.

of the parameters in Equation 1. We therefore fix the specular exponent to $m = 30$ based on prior measurements of human skin [35] and our own observations, and estimate only $\mathbf{n}$, $\boldsymbol{\alpha}$, and ρ. The geometric quantities $\mathbf{l}$ and $\mathbf{v}$, and light intensity L are determined by the calibration procedures described in Section 4.

Illustrated in Figure 3, we use a standard UNet with skip connections [25]. The encoder and decoder each consist of 5 blocks with 3 convolutional layers per block. The bottleneck has 256 channels. The output of this UNet is forwarded to two separate networks: a geometry branch that predicts a normal map $\tilde{\mathbf{n}}$, and a reflectance branch that predicts an albedo map $\tilde{\boldsymbol{\alpha}}$ and log-scale specular intensity map, $log(\tilde{\rho})$. Both branches have 3 convolutional layers with 32 channels and one final output layer.

We do not rely on ground truth normals or reflectance data to supervise training. Instead we combine a stereo loss and a photometric loss derived from data that is far easier to obtain: four one-light-at-a-time (OLAT) images in both RGB and NIR of the same subject, in the same exact pose, illuminated by a set of calibrated lights activated individually in rapid succession, and a stereo depth map (blue dashed line in Figure 2). These images are only used at training time.

A stereo loss encourages our estimated normals $\tilde{\mathbf{n}}$ to agree with the gradients of the stereo depth map $\mathbf{n}_s$. The gradients are computed by applying a 5x5 Prewitt operator on stereo depth maps that are smoothed with RGB-guided bilateral filtering. Similar to [39], our stereo loss combines a L1 vector loss and angular loss:

$$\mathcal{L}_s(\tilde{\mathbf{n}}) = \|\tilde{\mathbf{n}} - \mathbf{n}_s\|_1 - (\tilde{\mathbf{n}} \cdot \mathbf{n}_s). \tag{3}$$

A photometric loss is computed between each of the OLAT images and an image rendered according to Equation 2 and our network outputs for the corresponding lighting condition:

$$\mathcal{L}_p^j(\tilde{\mathbf{n}}, \tilde{\boldsymbol{\alpha}}, \tilde{\rho}) = \left\| S_j \odot \left(I(\mathbf{l}_j, \mathbf{v}, \tilde{\mathbf{n}}, \tilde{\boldsymbol{\alpha}}, \tilde{\rho}) - \hat{I}_j \right) \right\|_1, \tag{4}$$

where $\hat{I}_j$ is the per-pixel color observed in the j^{th} OLAT image, and S_j is a binary shadow map, computed by ray-casting using the stereo depth and calibrated light position (Section 4). We also apply a prior to the albedo map that encourages piecewise constant variation [3]:

$$\mathcal{L}_c(\tilde{\boldsymbol{\alpha}}) = \sum_i \sum_{j \in \mathcal{N}(i)} \|\tilde{\boldsymbol{\alpha}}_i - \tilde{\boldsymbol{\alpha}}_j\|_1, \tag{5}$$

where $\mathcal{N}(i)$ is the 5×5 neighborhood centered at pixel i. We apply this prior only to clothing pixels, those labeled as either body or arms in the segmentation mask. We found that other regions in the scene did not benefit from this regularization.

Our total loss function is a weighted sum of these terms:

$$\mathcal{L}(\tilde{\mathbf{n}}, \tilde{\boldsymbol{\alpha}}, \tilde{\rho}) = \mathcal{L}_s(\tilde{\mathbf{n}}) + \lambda_p \sum_j \mathcal{L}_p^j(\tilde{\mathbf{n}}, \tilde{\boldsymbol{\alpha}}, \tilde{\rho}) + \lambda_c \mathcal{L}_c(\tilde{\boldsymbol{\alpha}}). \tag{6}$$

We set the weight λ_p to 10 and λ_c to 50 based on the validation dataset.

Data Augmentation and Training. To improve the robustness of our network, we apply a series of data augmentations to our captured OLATs to simulate a variety of different visible light conditions. Specifically, our training uses a combination of: evenly-lit RGB inputs obtained by adding together all of the OLAT images; inputs with strong shadows by selecting exactly one of the OLAT images; a mixture of two lights with different temperatures by applying randomly chosen color vectors to two randomly chosen OLAT images; low-light environments by adding Gaussian noise to a single OLAT; and saturated exposures by scaling and clipping a single OLAT. We sample evenly from these 5 lighting conditions during training. Further details on how these lighting conditions are simulated are provided in the supplementary material.

We train the network using the Adam optimizer [16] for 30K iterations, with a learning rate of 10^{-3} and a batch size of 8. Training takes 12 hours with 4 Tesla V100 GPUs.

4. Hardware Setup and Data Collection

Shown in Figure 4, our setup combines a 7.0MP RGB camera that operates at 66.67 fps with a stereo pair of 2.8MP NIR cameras that operate at 150 fps. The RGB camera and one of the NIR cameras are co-located using a plate beamsplitter and a light trap. The RGB and NIR cameras have a linear photometric response and we downsample all of the images by a factor of 2 in each dimension and take a central crop that covers the face at a resolution of 960×768.

Visible spectrum lighting is provided by 4 wide-angle LED spotlights placed at the corners of a roughly $1.5m \times 0.8m$ (width x height) rectangle surrounding the cameras located approximately 1.1m from the subject. NIR lighting is provided by 5 NIR spotlights, one adjacent to each of the visible lights, and a flash LED light located near the reference NIR camera to produce the "dark flash" input. These NIR light sources are temporally interleaved with projectors that emit NIR dot speckle patterns to assist stereo matching [22]. A microcontroller orchestrates triggering the lights and cameras to ensure that at any time only one visible light source and one NIR light source is active. All light sources are calibrated for position and intensity and treated geometrically as point light sources. The light intensity term L in Equation 2 accounts for these calibrated colors. Note that the NIR and visible light sources are not colocated and so slightly different values of $\mathbf{l}$ are used in Equation 2 between those two conditions.

The image acquisition rate is limited by the RGB camera's framerate and the total light output, but is fast enough for us to record video sequences of people who are gesturing and moving slowly. We compute optical flow [36] between consecutive frames captured under the same lighting condition to correct for the small amount of scene motion that occurs within a single round of exposures. Since the RGB and reference stereo NIR camera are co-located, we can generate pixel-aligned RGB, NIR, and depth images using scene-independent precomputed image warps.

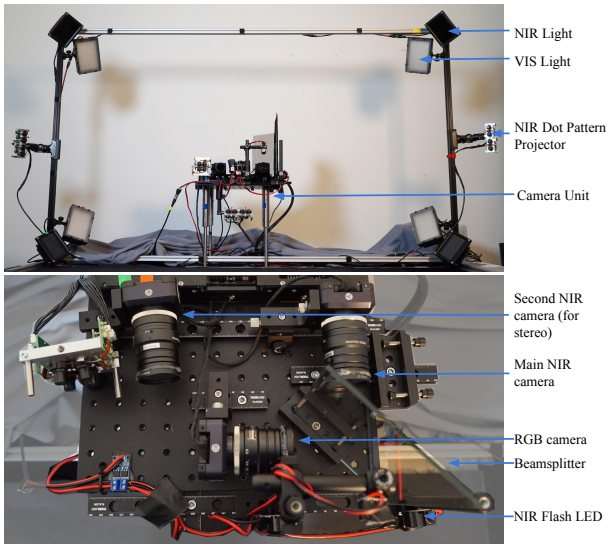


Figure 4: Our hardware setup consists of controllable NIR and visible spectrum light sources, an RGB camera, a stereo pair of NIR cameras, and two NIR dot projectors. One of the NIR cameras and the RGB camera are aligned with a beamsplitter and all of these components are triggered electronically to record the types of images shown in Figure 2.

Each recording in our dataset is 10 seconds long and contains 166 sets of frames. We recorded 9 unique subjects, with between 5 and 10 sessions per subject, for a total of 61 recordings. We used recordings of 6 of the subjects for training and tested on recordings of the other 3.

5. Evaluation

To the best of our knowledge, our method is the first technique for estimating surface normals and RGB albedos from an RGB+NIR image. We demonstrate the value of utilizing NIR inputs by comparing our method to two state-of-the-art RGB-only face normal estimation methods [27, 21] as well as an RGB-only variant of our own method. We also perform several other ablation studies to measure the impact of key design decisions. To illustrate the performance of our method in lighting conditions that do not lie in the span of our captured OLAT images, we also show qualitative results (Figure 1) on a real sequence captured while casually moving a handheld light source around the scene. Note that ground-truth normal maps are not available for this sequence. Finally, we present two applications of our technique. All results are expanded to animated image sequences and can be viewed on our project page at darkflashnormalpaper.github.io. None of the sub-

	Well lit	Shadows	Mixed colors	Overexposure	Low light
SfSNet [27]	14.10	18.32	-	-	-
Nestmeyer et al.[21]	14.82	17.52	15.87	21.85	25.56
Ours (No Stereo Loss)	12.80	12.78	12.78	12.82	12.81
Ours (No NIR Photometric Loss)	12.64	12.66	12.64	12.69	12.75
Ours (No Photometric Loss)	12.77	12.77	12.81	12.79	12.77
Ours (No Specular Component)	12.44	12.43	12.44	12.51	12.47
Ours (No RGB Input)	12.54	12.54	12.54	12.54	12.54
Ours (No NIR Input)	13.13	15.19	16.43	19.82	19.39
Ours	**12.08**	**12.06**	**12.06**	**12.14**	**12.10**

Table 1: Mean absolute angular error in degrees of normal maps computed with modified versions of our full network. Results are reported for the five lighting conditions described in Section 5.

RGB Input | Shading (w/o Photometric Loss) | Shading (w/ Photometric Loss) | Albedos (w/o Blinn-Phong) | Albedos (w/ Blinn-Phong)

Figure 5: Impact of the photometric loss term in our training procedure and the Blinn-Phong BRDF in our image formation model, respectively. When trained without photometric loss, our network learns to output the stereo normals, which lack fine-scale details. This has a fairly small effect on the error measures in Table 1, but is perceptually significant as seen in these "n dot l" shading renderings. Our full image formation model, which includes a Blinn-Phong specular term, produces more accurate albedos across the face than using a Lambertian model alone.

jects shown in our results are in our training set.

5.1. Comparisons and Ablation Studies

In our evaluations we consider five different visible lighting conditions: harsh lighting that produces strong cast shadows; a mixture of lights with different color temperatures; saturated/overexposed intensities; low-light conditions that produce noisy inputs; and a "well lit" condition that achieves largely shadow-free and well exposed inputs. Our process for synthesizing these different lighting conditions from the OLAT training images is detailed in our supplemental document. In lieu of ground truth geometry for quantitative assessments, we construct a baseline using the technique of Nehab et al. [20] to refine our stereo depth maps according to normals computed by applying Lambertian photometric stereo to the RGB OLAT training images.

Table 1 reports the mean absolute angular errors in normal maps computed by two state-of-the-art RGB-based face normal estimation methods [27, 21] along with several variants of our network with different loss terms, image formation models, and inputs. Figures 5 and 6 show examples of the perceptual impact of some of these design decisions.

Comparisons to SfSNet [27] and Nestmeyer et al [21]. Details on how we adapt and retrain SfSNet [27] and Nestmeyer et al [21] on our captured dataset, along with qualitative image comparisons, can be found in our supplemental document. As shown in Table 1, our method outperforms both techniques even in the well lit condition and without using the NIR input, which we attribute to our novel training strategy that combines shape information from complementary stereo and photometric signals. More importantly, in challenging lighting conditions, the benefit of our method becomes far more significant as the additional information provided by the NIR input is crucial in these circumstances. Note that SfSNet [27] uses a self-reconstruction loss that we found could not handle inputs with mixed color casts, saturated intensities, or a significant amount of noise and so it fails to produce plausible outputs in these cases (omitted from Table 1).

Loss terms. As expected, using both stereo and pho-

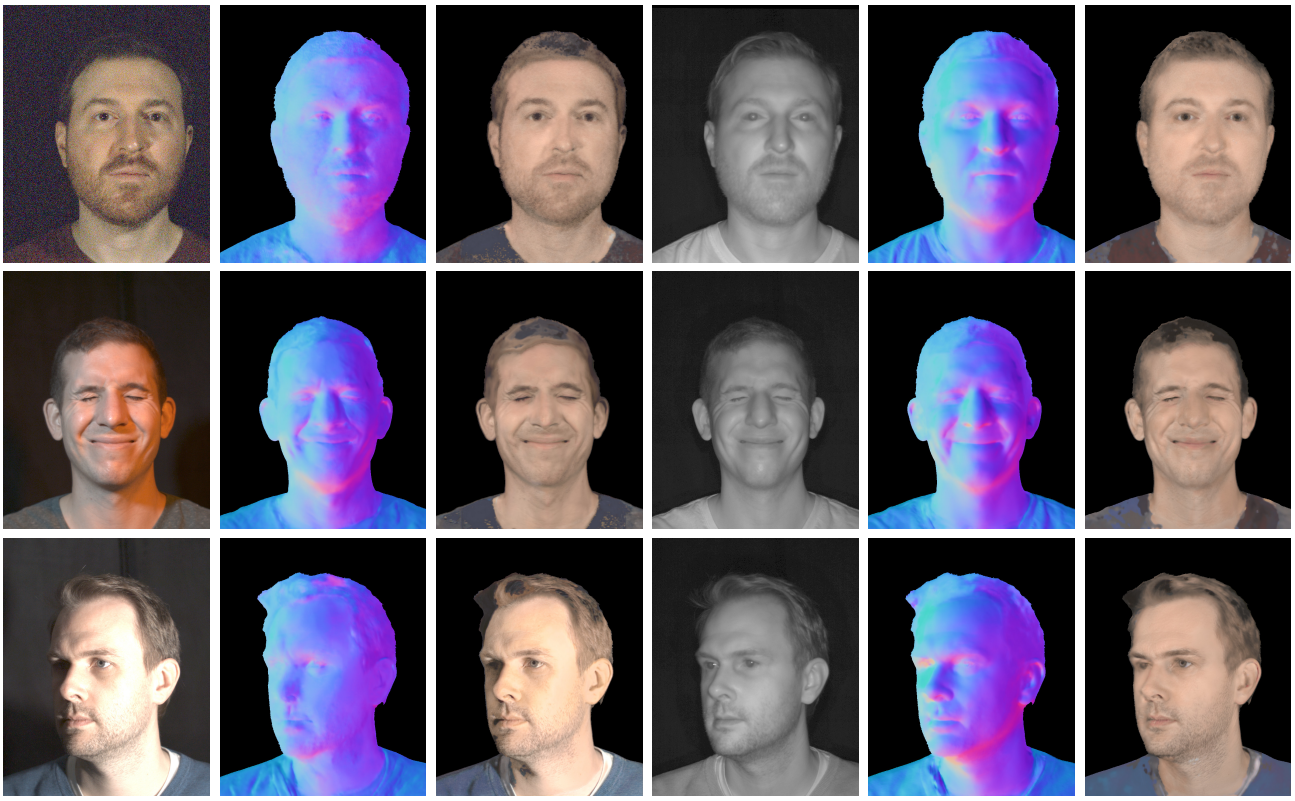
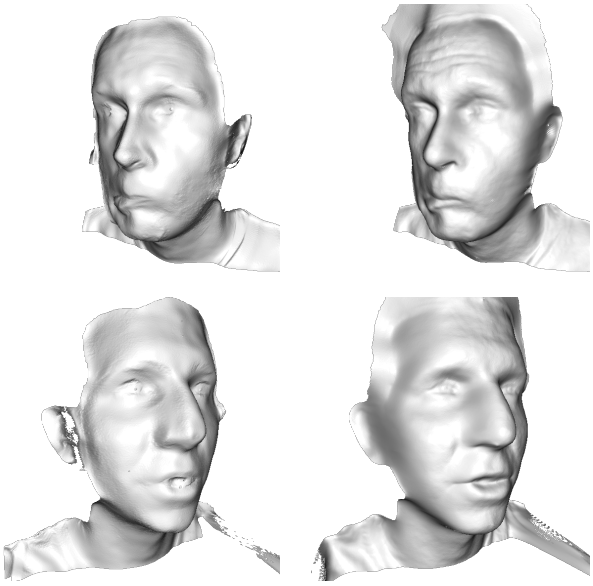


Figure 6: Comparison of our network to a modified version that takes only a single RGB image ("RGB Only") as input. Example results for three common challenging lighting conditions. Top to bottom: low light / noisy inputs; mixed light colors; harsh directional lighting with saturated intensities. The "RGB only" network struggles to produce stable normal and reflectance estimates from these inputs in contrast to our method.

tometric loss terms during training outperforms using either one alone. We consider two types of photometric loss - one computed on only the RGB training images ("No NIR Photometric Loss" in Table 1) and the second computed on both the NIR and RGB training images ("Full Method"). As illustrated in the shading images in Figure 5, including the photometric loss enables estimating fine geometric details that are not captured in the stereo depth maps.

Image formation model. Including the Blinn-Phong BRDF in our image formation model improves the accuracy of the normals and diffuse albedo maps. It results in a modest improvement in the quantitative errors in Table 1, and it produces more uniform diffuse albedo maps with fewer artifacts (Figure 5). We attribute this to the fact that this richer image formation model is better able to explain the observed intensities. We also found that including this BRDF in our model enables reconstructing the glossy appearance of skin (Section 5.3).

Network inputs. Including the NIR input image improves accuracy across the board, especially in poor visible lighting conditions (Table 1). The benefit of the RGB input is comparatively smaller, but making it available to the network enables estimating visible spectrum reflectance data, which is a requirement for many downstream applications such as lighting adjustment (Section 5.3). Figure 6 illustrates the perceptual impact of including the NIR input in different lighting conditions. For these comparisons we modified our network to take only a single RGB image as input ("RGB Only"). The network architecture was otherwise unchanged, and we applied the same training procedure described in Section 3. Note how the performance of this "RGB Only" network significantly degrades in challenging conditions, while our method is far more robust to these conditions due to the more stable NIR input. It's particularly noteworthy how well our method is able to reconstruct plausible diffuse albedos even for highly saturated RGB input images (bottom row of Figure 6).



Smoothed Stereo Refined Stereo (Ours)

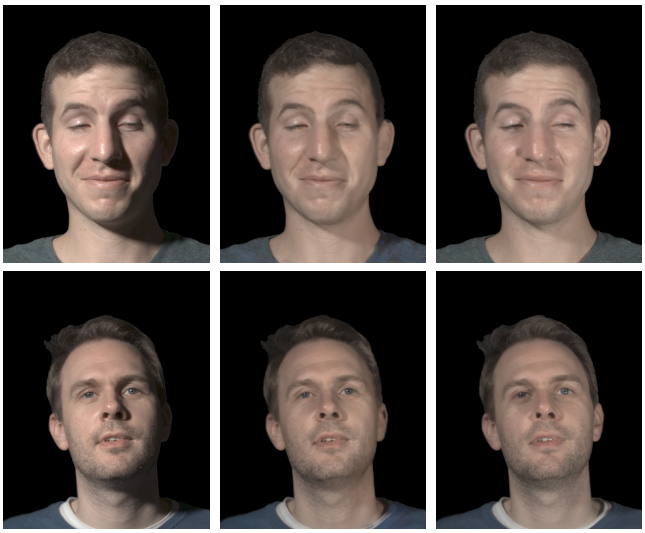
Figure 7: Stereo methods often struggle to recover fine-scale surface details. *Left:* Applying a guided bilateral filter to raw stereo depths yields a smoother surface but with distorted features (e.g. the nose is reduced and skin wrinkles are missing). *Right:* We use the method of Nehab et al. [20] to compute a refined surface according to normals estimated with our method. Note how details are better preserved around the eyes, nose, and mouth, along with fine wrinkles and creases.

5.2. Application: Stereo Refinement

Stereo methods excel at measuring coarse geometry, but often struggle to recover fine-scale surface details. This can be overcome by refining stereo depths according to accurate high-resolution normals typically estimated with a photometric approach [20]. We evaluate using the normals produced by our method to refine depth measurements produced by an NIR space-time stereo algorithm [22] (Figure 7). In comparison to using a standard bilateral filter to smooth the stereo depths, refining them using our normals gives much higher quality reconstructions, most notably around the mouth, nose, and eyes and better recovery of fine wrinkles and creases in the skin. As our method works with a single NIR image it would be straightforward to integrate it into many existing stereo pipelines.

5.3. Application: Lighting Adjustment

We also explored using our approach to digitally improve the lighting in a portrait. Specifically, we evaluated adding a virtual fill light to brighten shadowed parts of the face (Figure 8). We used normal and reflectance maps estimated by our method to render the contribution of a virtual point light located within view of the shadowed region, and then combined this with the original RGB image. Our model enables a convincing effect, even producing realistic specular highlights along the nasolabial folds and the tip of the nose.



RGB Input Relit Ground Truth

Figure 8: Our method can be used to simulate adding lights to a scene to fill in shadows.

6. Conclusion

We have presented a dark flash normal camera that is capable of estimating high-quality normal and reflectance maps from a single RGB+NIR input image that can be recorded in a single exposure without distracting the subject. A key benefit of our method over prior work is its robustness. It performs well even in challenging lighting conditions that are commonly encountered in casual photography such as harsh shadows, saturated pixels, and in very low light environments.

Our approach assumes a single light located near the camera is the only source of NIR light in the scene. Although this is a safe assumption in many indoor environments, it is not always true, especially outdoors. It may be possible to suppress some ambient light through the use of a flash/no-flash image pairs [23].

Our method could be integrated into existing smartphone camera hardware designs and software pipelines to enable a range of applications from boosting the performance of an auxiliary depth camera to enabling face relighting in still images and streaming video. Future work also includes improving our method's performance on hair and clothing and its temporal stability, potentially by allowing the network to consider consecutive frames.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. In *Proc. CVPR*, 2012. 2

[2] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. *Proc. CVPR*, 2013. 2

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 2, 4

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

[5] Anil S. Baslamisli, Hoang-An Le, and Theo Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *Proc. CVPR*, 2018. 2

[6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 2

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 2

[8] James F Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, 1977. 3

[9] Xu Cao, Michael Waechter, Boxin Shi, Ye Gao, Bo Zheng, and Yasuyuki Matsushita. Stereoscopic flash and no-flash photography for shape and albedo recovery. In *Proc. CVPR*, 2020. 2

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[11] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *Proc. ICCV*, 2019. 2

[12] Gyeongmin Choe, Jaesik Park, Yu-Wing Tai, and In Kweon. Exploiting shading cues in kinect ir images for geometry refinement. In *Proc. CVPR*, 2014. 2

[13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proc. CVPR Workshops*, 2019. 2

[14] Prabath Gunawardane, Tom Malzbender, Ramin Samadani, Alan McReynolds, Dan Gelb, and James Davis. Invisible light: Using infrared for video conference relighting. In *Proc. ICIP*, 2010. 3

[15] Berthold K. P. Horn. *Obtaining Shape from Shading Information*, page 123–171. MIT Press, Cambridge, MA, USA, 1989. 2

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015. 5

[17] Dilip Krishnan and Rob Fergus. Dark flash photography. *ACM Transactions on Graphics (TOG)*, 28(3):96, 2009. 3

[18] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2

[19] Zhe Liang, Chao Xu, Jing Hu, Yushi Li, and Zhaopeng Meng. Better together: shading cues and multi-view stereo for reconstruction depth optimization. *IEEE Access*, 8:112348–112356, 2020. 2

[20] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Transactions on Graphics (TOG)*, 24(3):536–543, 2005. 2, 6, 8

[21] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. Learning physics-guided face relighting under directional light. In *Proc. CVPR*, 2020. 2, 3, 5, 6

[22] Harris Nover, Supreeth Achar, and Dan Goldman. ESPReSSo: Efficient slanted patchmatch for real-time spacetime stereo. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018. 5, 8

[23] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics (TOG)*, 23(3):664–672, 2004. 8

[24] Di Qiu, Jin Zeng, Zhanghan Ke, Wenxiu Sun, and Chengxi Yang. Towards geometry guided neural relighting with flash photography. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 2

[25] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 4

[26] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proc. CVPR*, 2019. 2

[27] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Proc. CVPR*, 2018. 2, 5, 6

[28] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *Proc. CVPR*, pages 1685–1694, 2017. 2

[29] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Proc. CVPR*, 2017. 2

[30] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):79–1, 2019. 3

[31] Tatsunori Taniai and Takanori Maehara. Neural inverse rendering for general reflectance photometric stereo. In *Proc. ICML*, 2018. 2

[32] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Au-

toencoder for Unsupervised Monocular Reconstruction. In *Proc. ICCV*, 2017. 2

[33] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proc. CVPR*, 2018. 2

[34] Oliver Wang, James Davis, Erika Chuang, Ian Rickard, Krystle De Mesa, and Chirag Dave. Video relighting using infrared illumination. In *Computer Graphics Forum*, volume 27, pages 271–279, 2008. 3

[35] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (ToG)*, 25(3):1013–1024, 2006. 3, 4

[36] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, Cen Rao, and Michael Isnardi. Bilateral filtering-based optical flow estimation with occlusion detection. In *Proc. ECCV*, 2006. 5

[37] Hiroki Yamashita, Daisuke Sugimura, and Takayuki Hamamoto. Rgb-nir imaging with exposure bracketing for joint denoising and deblurring of low-light color images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 3

[38] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv preprint arXiv:1711.03665*, 2017. 2

[39] Youngjin Yoon, Gyeongmin Choe, Namil Kim, Joon-Young Lee, and In So Kweon. Fine-scale surface normal estimation using a single NIR image. In *Proc. ECCV*, 2016. 2, 4

[40] Ye Yu and William A. P. Smith. InverseRenderNet: Learning single image inverse rendering. In *Proc. CVPR*, 2019. 2

[41] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proc. ICCV*, 2019. 2

[42] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single-image portrait relighting. In *Proc. ICCV*, 2019. 3