Stereoscopic Dark Flash for Low-light Photography

Jian Wang[1], Tianfan Xue[2], Jonathan T. Barron[2], and Jiawen Chen[2]
[1]Carnegie Mellon University, Pittsburgh, PA
[2]Google Research, Mountain View, CA

In this work, we present a camera configuration for acquiring "stereoscopic dark flash" images: a simultaneous stereo pair in which one camera is a conventional RGB sensor, but the other camera is sensitive to near-infrared and near-ultraviolet wavelengths instead of red and blue. When paired with a "dark" flash (i.e., one emitting near-infrared and near-ultraviolet light, but no visible light) this camera allows us to capture a flash/no-flash image pair at the same time, all without disturbing any human subjects or onlookers with a dazzling visible flash. We present a hardware prototype of this camera that approximates an idealized camera, and an imaging procedure that let us acquire dark flash stereo pairs that closely resemble those we would get from that idealized camera. We then present a technique for fusing these stereo pairs, first by performing registration and warping, and then by using recent advances in hyperspectral image fusion and deep learning to produce a final image. Because our camera configuration and our data acquisition process allow us to capture true low-noise long exposure RGB images alongside our dark flash stereo pairs, our learned model can be trained end-to-end to produce a fused image that retains the color and tone of a real RGB image while having the low-noise properties of a flash image.

***Index Terms*—computational photography, low-light imaging, dark flash, stereo**

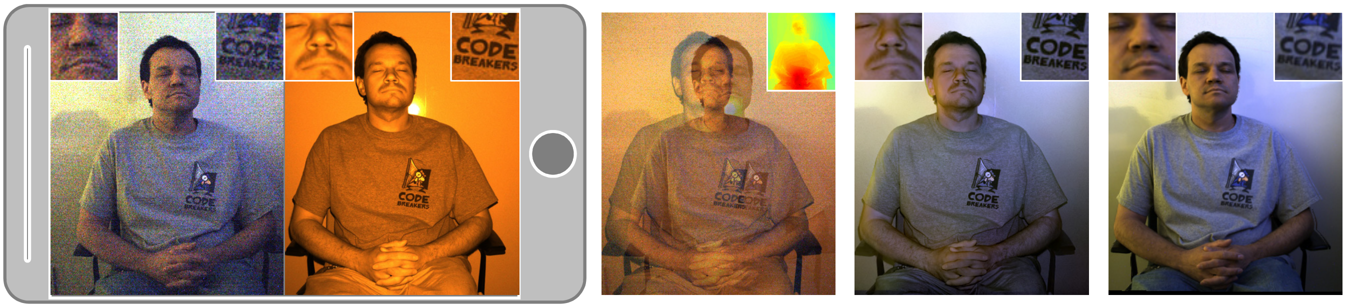
I. Introduction

The rise of mobile computing in the 21st century has caused photography to be a ubiquitous part of the human experience. But the small form factor of mobile phones necessarily limits the aperture size of the cameras that can be built into these devices, which in turn limits the amount of light that these cameras can detect. As a result, images taken by mobile devices in low light environments are often dominated by noise.

This issue can be ameliorated through conventional means, such as increasing the exposure time of the camera or using a flash, but these solutions have necessary drawbacks. Increasing exposure time allows more photons to be captured, but will induce a blur in the resulting photograph if the camera or subject moves — barring the use of a tripod-mounted camera or a still life subject. Using a flash adds light to the scene but fundamentally changes the subject's appearance, often causing photos to look harsh or unnatural. In addition, using a flash may dazzle or otherwise disturb a human subject, or may transgress social norms in some circumstances.

Building on the idea of increasing exposure time, Hasinoff *et al.* [1] approximate a long-exposure image by capturing a burst of short-exposure images, and merging them together to obtain a lower-noise image. But this approach may still fail to reduce noise or eliminate blur in the presence of significant motion or very little light, and at best can only yield an SNR increase that is proportional to the square root of the number of images in the burst. To address the sometimes-unattractive appearance of flash photographs, many researchers have explored capturing "flash" and "no-flash" image pairs and merging them to produce an image with the high SNR of the "flash" image, but with the attractive visual qualities of the "no-flash" image [2], [3]. Though it sometimes produces compelling results, because the flash/no-flash image pairs are taken at different times, this approach may fail in the presence of scene or camera motion. Additionally, a human subject would find it just as bothersome to be photographed by a flash/no-flash camera as they would a conventional flash camera. To address the physiological and sociological problems associated with conventional flash photography, Krishnan and Fergus [4] propose a "Dark Flash" that uses near-infrared (NIR) and near-ultraviolet (NUV) light which are invisible to the human eye. But their approach completely replaces the standard RGB sensor of a camera with another single sensor that conflates red with infrared, and blue with ultraviolet, and so they are entirely dependent on statistical correlations between visible and hyperspectral wavelengths to recover a visible-spectrum image. Though this correlation is strong, it is not deterministic, and so the inferred visible-spectrum image may contain significant artifacts. And because the sole sensor used by this camera configuration conflates visible with hyperspectral wavelengths, it cannot "fall back" to the RGB image as in the case of conventional flash/no-flash photography. Dark flash photography also inherits the vulnerabilities of flash/no-flash photography to camera or scene motion, as the two images of the pair are taken at different times.

Our approach, which we dub "stereoscopic dark flash", is an attempt to address some of the shortcomings of dark flash photography. Instead of using a single RGB camera with its IR/UV filter removed, we instead use two cameras: one standard RGB camera, and a second camera whose red and blue channels are replaced, making them sensitive to NIR and NUV respectively — but insensitive to visible red and visible blue. Like in "dark flash" photography, our flash is limited to only NIR and NUV (see Figure 2 (a)). When taking a photograph, our camera rig fires the NIR-NUV flash and records an image from both the RGB and NIR-G-NUV cameras, all at the same time. Because the two cameras are at different physical locations, we must register the images to each other, which we do using their green channels (both of which are unaffected by the NIR-NUV flash and so appear



(a) Input: left - RGB image (× 5 for visualization), right - dark flash image

(b) Two inputs overlaid

(c) Our output

(d) Long exposure ground truth

Fig. 1. Our stereo camera (a) simultaneously captures a conventional RGB image and an invisible "dark flash" image. We estimate stereo depth (b) and fuse them into a high-quality RGB result (c) that resembles a long-exposure ground truth RGB image (d).

similar). After a per-pixel alignment, we denoise the RGB image using the NIR and NUV channels of the flash image as a guide, which yields a low-noise image with natural RGB colors (see Figure 1).

As we will show, conventional methods for fusing a dark flash image with a noisy RGB image do not produce sufficiently high-quality results. The fused image may contain artifacts due to errors during registration, may have the characteristic harsh lighting of a conventional flash photograph, and may be corrupted by content not present in visible wavelengths. Thankfully, unlike in conventional dark flash photography, the presence of a standard RGB camera in our rig allows us to address this issue by collecting long exposure RGB images, which we use to train a neural network to regress from our naively fused RGB+NIR/NUV images to these ground truth RGB images. The resulting model learns to remove artifacts while retaining color information, and to modify the tonal content of the image to make it look more like a natural RGB image.

The advantages of our proposed setup are many:

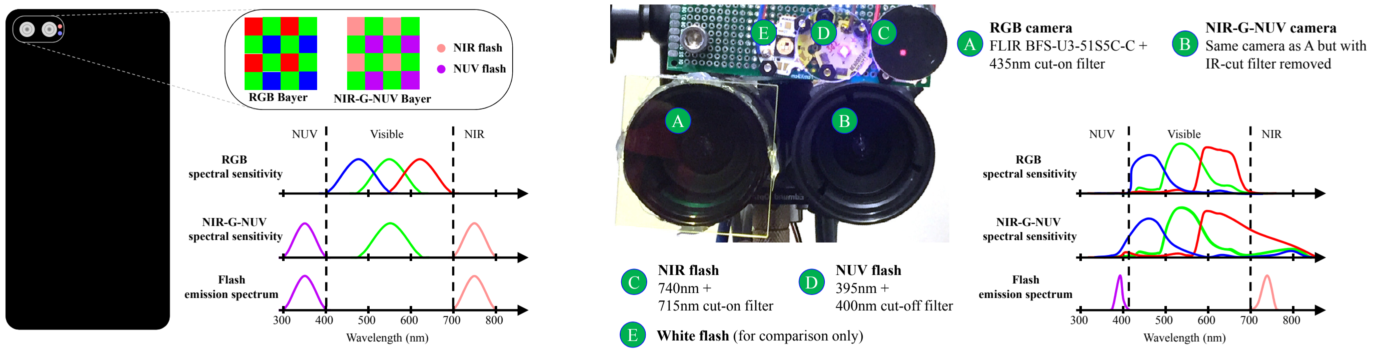
1) Our technique retains all of the value contributed by dark flash photography: low-noise images can be recovered, without having to dazzle or disturb any human subjects or bystanders.
2) Unlike burst photography, flash/no-flash, and dark flash photography, our two images are acquired at the exact same time. This not only allows for a responsive and low-latency user experience, but also means that our system is robust to camera or scene motion.
3) Like in conventional flash/no-flash photography, our rig directly acquires a conventional RGB image. Contrast this with dark flash photography, in which the no-flash image is *not* an RGB image, but instead contains both red and NIR wavelengths in the "red" channel and both blue and NUV wavelengths in the "blue" channel. This means that in well-lit environments in which noise is not an issue, we can produce a high-quality image by simply returning the observed RGB image. Moreover, in the case of some failure either during registration or image fusion, our setup can always degrade gracefully back to the observed RGB image if need be.
4) Unlike dark flash (but similarly to a flash/no-flash or burst photography setup) our setup allows for the collection of long-exposure "ground truth" RGB images, which can then be used for learning.
5) By constructing the spectral response curves of our cameras and flash such that the green channels are identical and have no overlap with our flash, conventional stereo techniques (which, naturally, expect input images to look similar) can be used to recover a reliable depth map. The depth maps we recover may also be useful for other photographic purposes such as background defocus [5].

However, our technique does have some limitations and costs that are worth considering:

1) The intensity of our flash, like any light, necessarily decreases with distance, and so our approach will not provide a benefit in distant scenes.
2) Because we rely on a stereo algorithm to align our input images, our output image may contain some artifacts around occlusions or other image regions where correspondence is difficult to compute.
3) Compared to a standard stereo rig, our setup has slightly less information with which to estimate disparity, as we use only the green channel shared by our two sensors while a standard stereo setup can use all three RGB channels. In practice, we observe that the drop in depth quality when using green instead of RGB appears to be small.
4) Unlike burst, flash/no-flash, or dark flash photography, our approach requires two cameras instead of one, which increases the cost of manufacturing and calibrating such a device.

II. Related Work

Modern camera sensor design has been the focus of decades of research and engineering, much of which is well beyond the scope of this review. Conventional mobile cameras are constructed by placing a Bayer color filter array (CFA) [6] in front of a CMOS sensor that is sensitive to light in the range of 300 to 1000 nanometers. A Bayer filter is composed of repeated "quads" containing four pixels, each of which sits behind a color bandpass filter that eliminates all but red, green, or blue light. After the image is captured, a demosaicking algorithm [7] interpolates an estimate of the two colors at


RGB Bayer
NIR-G-NUV Bayer
NIR flash
NUV flash
NUV
Visible
NIR
RGB spectral sensitivity
NIR-G-NUV spectral sensitivity
Flash emission spectrum
300
400
500
600
700
800
Wavelength (nm)
A
B
C
D
E
RGB camera
FLIR BFS-U3-51S5C-C + 435nm cut-on filter
NIR-G-NUV camera
Same camera as A but with IR-cut filter removed
NIR flash
740nm + 715nm cut-on filter
NUV flash
395nm + 400nm cut-off filter
White flash (for comparison only)

(a) Proposed setup

(b) Hardware prototype

Fig. 2. In (a) we show an idealized version of our proposed imaging setup: a mobile device with two cameras (one RGB, one NIR-G-NUV) and a dark flash, all with ideal spectral characteristics. In (b) we show our prototype camera system and its actual spectral curves.

each pixel that were not directly observed, resulting in a complete color image. Note that, by design, a Bayer filter approach discards approximately two thirds of all incoming light. Though these CFAs are typically designed to filter out all but visible light, they have some leakage in the NIR range that necessitates an NIR cut-off filter, which is usually placed on the lens of the camera. The final spectral sensitivity curves for the sensor pixels of different camera brands are well summarized in [8]. While CMOS sensors have benefits over CCD sensors in terms of power efficiency and readout speed, they require additional space for per-pixel readout and amplifier circuits, and therefore have less area that can be used for observing incident light [9]. These issues, compounded by the aforementioned limits on the size of mobile devices and the commonality of low-light environments, result in a camera whose images are often noisy and therefore benefit from a denoising algorithm or additional information that can be used to reduce noise.

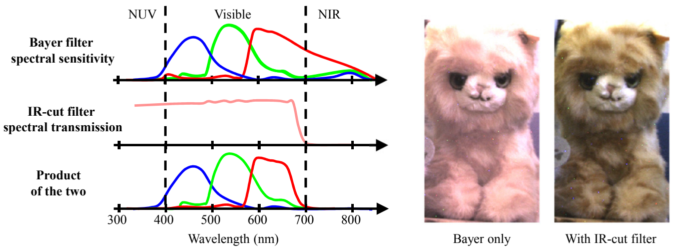
Image denoising has been the subject of significant research, with many techniques using a single image as input, such as BM3D [10], sparse coding [11], low-rank factorization [12], or modern deep learning based methods [13]. These methods are generally computationally expensive, and are necessarily limited in their ability to recover details in the presence of overwhelming noise. Performance can be improved by using a burst of images to denoise a single image [1], [14]–[16], though these approaches require computing a correspondence across images or some technique for being invariant to this correspondence problem, which can be problematic in the presence of significant camera or scene motion.

Instead of acquiring a noisy image and then attempting to remove that noise, one can instead adjust the imaging conditions to capture a less noisy image. As mentioned previously, increasing the exposure time of the camera reduces noise, but results in blurring artifacts in the presence of scene motion or camera shake. This motion blur can be removed through algorithmic means [17], but this “deblurring” problem is itself difficult and underconstrained, and arguably no easier than the denoising problem that is being circumvented. Alternatively, one could reduce noise by increasing the amount of illumination in the scene, through the use of a flash. Flash photographs tend to have an unpleasantly harsh and unnatural appearance, but this can be reduced by merging a flash photograph with a no-flash image [2], [3]. But even if the flash/no-flash problem were solved, many people still find the bright and dazzling white flash of a camera to be annoying or otherwise disruptive. Dark flash photography [4] avoids this problem by using NIR and NUV flashes, and modifying the camera to be NIR/NUV-sensitive by removing the IR/UV-cut filter, though this system has its own drawbacks, as explained previously.

The flash/no-flash imaging strategy presents the question of how to best combine the high fidelity of the flash image with the more pleasing aesthetic qualities of the no-flash image, and this question has received a significant amount of attention. Early approaches [2], [3] use joint bilateral filtering to produce a “detail” layer from the flash image that is then propagated to the no-flash image. Other edge-aware filters, such as the guided filter [18] can be used similarly. Dark flash photography [4] merges its two images using an optimization framework that assumes the gradient of the denoised result should be similar to the gradient of the flash image. Though image gradients are often strongly correlated across different wavelengths, the occasional variations that do occur can cause algorithms that depend on this assumption to fail. To address this issue, Shen et al. [19] explicitly model the structural divergence across wavelengths as a *scale map* — the ratio between the gradient maps of the flash and no-flash images, which they jointly estimate alongside a denoised image. Similarly, mutually guided image filtering [20] also propagates information across disparate wavelengths through a joint estimation process. Though these techniques work well, their hand-engineered nature means that they often propagate gradient information that has undesirable tonal properties from the flash image that would not be present in a long-exposure image. For this reason, we build upon [19] but augment it with a neural network that has been trained to remove the unwanted tonal and color properties of the dark flash.

III. Proposed Setup

Our proposed imaging configuration consists of a conventional RGB camera (cam_1), an NIR-G-NUV camera (cam_2), an NIR flash, and an NUV flash. In Figure 2(a), we illustrate



(a) Spectrum transmission curves

(b) Sample image of a toy lion

Fig. 3. (a) Spectral transmission curves of Bayer filters, low-pass NIR filter, and their product. (b) Sample images with and without the NIR filter.

the envisioned use of our system in a cell phone, alongside the Bayer pattern of the two cameras and the idealized spectral response curves for both sensors' micro-filters and the corresponding flashes. This idealized camera captures a single shot by simultaneously firing both dark flashes while exposing both sensors. The flash is invisible to both the human eye as well as cam_1, but is visible to cam_2 thereby allowing it to record a low-noise flash image in low-light environments. Because the two cameras have different positions, merging the two images requires solving for a per-pixel mapping. We do this by using the green channels of the two images (which, because the green curves of the two cameras are matched, look similar) to estimate a dense stereo depth map, and use this depth map to help merge the two images and produce a single high-quality RGB result.

A. *Our prototype*

We built our prototype camera system using off-the-shelf components that approximates our proposed configuration (Figure 2(b)). In this section, we detail our design choices, the practical limitations caused by our selection of hardware, and how these choices affect our dataset capture strategy. The supplement contains details on hardware components and their specifications.

We chose to match the sensor sizes and lenses of our two cameras, giving us images with the same field of view and resolution. This was done to minimize the difficulty of performing stereo registration, and to prevent artifacts in the final merged image that may result from mismatched FOVs causing the observed areas of the two images to be significantly different. We selected FLIR sensors that are sensitive to RGB, NIR, and NUV light, and we chose lenses that allow NIR through NUV wavelengths to pass through, and that also produce little chromatic aberration when focusing all relevant wavelengths. Our choice of sensors and the diameter of our lenses constrained our baseline to be about 52mm, making stereo registration challenging when subjects are close. For a production smartphone camera, we envision using custom lenses with folded optics to optimize the baseline so as to match the effective range of the flash.

Our left camera (cam_1) is a standard RGB camera, to which we add a UV filter on its lens (which was already equipped with an NIR filter) to ensure that it is only sensitive to visible light. Our right camera (cam_2) is an NIR-G-NUV camera. Because it is difficult to construct a camera that is sensitive to different wavelengths than cam_1 but is otherwise physically identical, we instead build cam_2 by modifying an RGB camera. Recall that Bayer micro-filters are not true band-pass filters, as they transmit significant amounts of light outside the visible spectrum (see Figure 3). This means that if we remove the NIR-cut filter from the lens, we can make an RGB camera sensitive to both NIR and NUV. Unfortunately, this also means that its green channel will receive a non-negligible amount of NIR that is sometimes problematic for our stereo algorithm. To address this issue, our dataset consists exclusively of indoor scenes, which allows us to minimize the amount of ambient IR and UV light that is present during imaging. Still, despite our best efforts, we noticed that images from cam_2 have a slight red tint and are slightly blurrier due to NIR contamination and chromatic aberration.

To demonstrate the feasibility of our algorithms despite whatever practical issues with our hardware we may have, during capture we simulate an ideal shot by acquiring bursts of images for each scene. Our bursts rapidly interleave shots with flash off and flash on. We compute stereo correspondence on the flash-off frames (where the green channels are uncontaminated by NIR), but use this estimated depth map to warp the flash-on frames. This burst capture also let us benchmark our method against burst denoising algorithms as an alternative strategy for producing low-noise RGB images.

Our difficulty in constructing a physical prototype that exactly matches our proposed setup may prompt the reader to question if it is indeed possible to manufacture pure sensor-level R, G, B, NIR, and NUV Bayer filters. Indeed, Spooren *et al.* [21] show that it is possible to construct a compact, low-cost RGB-NIR camera — albeit one that is difficult to procure — using pixel-level monolithic integration of traditional absorption-based RGB color filters with NIR-pass and NIR-cut filters implemented using Fabry-Pérot interference. By using Fabry-Pérot interference filters and mosaicking different filters to select different wavelengths, others have successfully implemented Bayer hyperspectral cameras [22], [23] which exceed our setup's requirements.

IV. Burst Dataset

To facilitate our experiments we collect a dataset of 121 scenes, which will be used for training our model and to benchmark our model's performance against baseline techniques. This dataset collection procedure is designed to be more general and flexible than is needed for this work, in the hopes that a large and rich dataset of this sort may be a useful resource for future research.

Because we acquire bursts of images, and because the problem of image registration is difficult and somewhat out of the scope of this work, our strategy is to minimize differences across images due to anything other than flash (i.e., motion). To this end, all images are captured using a tripod. 30 of our scenes are of static environments, and the remaining 91 scenes contain human subjects. These subjects were told to hold still during acquisition, though our human-subject scenes do generally contain small amounts of motion.

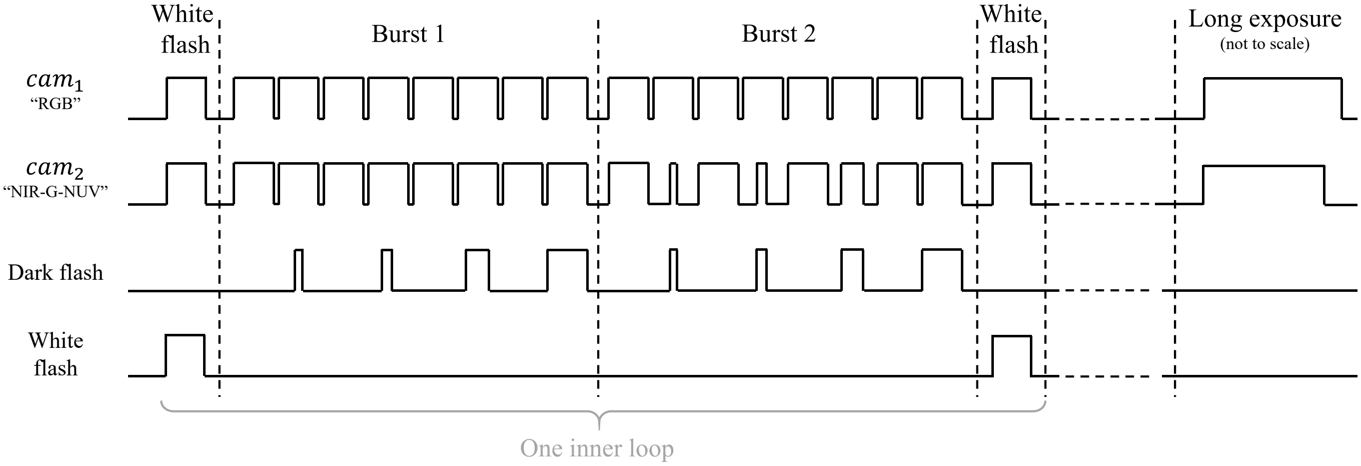


Fig. 4. Our data collection strategy, which includes white flash, two bursts under varying length dark flashes, and two no-flash long exposure images.

For each scene, we first run a simple automatic exposure algorithm (see the supplement) to estimate an appropriate exposure time T as well as gains for the two cameras. We then capture a collection of bursts, where for each burst we vary one property of our acquisition setup (see Figure 4). Our acquisition procedure is described in Algorithm 1.

For *cam*$_1$, we simply capture a uniform burst with exposure time T. For *cam*$_2$, recall that it is an approximation of an ideal NIR-G-NUV sensor. We therefore collect two bursts with different *cam*$_2$ exposure times when the dark flash is on to assess the tradeoff between motion and NIR contamination in stereo registration. Burst 1 maintains a uniform exposure time, thereby ensuring equivalent motion blur in the two images, but *cam*$_2$'s green channel records ambient NIR in addition to that from the flash. In Burst 2, we match *cam*$_2$'s flash-on exposure times with that of the flash. This minimizes NIR contamination at the expense of mismatched motion blur. In practice, we found little difference between the two since our scenes are largely static. We use burst 1 for all of our results.

For each exposure time and flash combination, we bookend the two bursts with a white flash image. Finally, we also capture a long-exposure ground truth image. All images are captured in 16-bit Bayer raw. Figure 5 shows one example of the many images acquired for one scene in our dataset. Note that these bursts are used only for training and evaluation, and that this burst acquisition procedure is *not* necessary for acquiring test-time photographs using our camera rig.

Algorithm 1: Our Burst Collection Procedure

```
for $t \in [T, T/3, T/5, T/7]$ do
    for flash $\in [\mathrm{NIR}, \mathrm{NIR} + \mathrm{NUV}]$ do
        Capture 1st still image with white flash on;
        Capture burst 1;
        Capture burst 2;
        Capture 2nd still with image white flash on;
    end
end
Capture long exposure ground truth;
```

Though our acquisition process is somewhat complicated, the resulting data we acquire has a number of useful properties:

- The long-exposure RGB images can be used as ground truth for training and evaluating models.
- Because we acquire an RGB burst, we can directly compare our results against a standard multi-image denoising technique.
- The interleaved flash/no-flash bursts allows us to compare to existing work in flash/no-flash fusion, both visible and dark.
- Because our scenes are largely static, the correspondence between the two viewpoints is the same across all acquired pairs of images. This let us use the depth map recovered from a no-flash image pair to register a subsequent flash/no-flash image pair.

If our hardware matched the idealized setup in Figure 2(a), we could estimate a per-pixel registration across our two cameras by simply applying a stereo technique to the green channels of our two images. However, because *cam*$_2$'s green channel is in fact quite sensitive to our NIR flash, the green channel of a flash-on image is unsuitable for stereo matching with the green channel of *cam*$_1$. We circumvent this issue by computing our depth maps using the previous pair of frames, where the flash is off. Because images in our bursts are taken in rapid succession, these depth maps tend to hold well across consecutive frames — even when presented with subtle motions in our bursts of human subjects. The problem of trans-modal stereo correspondence is an interesting direction for future research.

After acquiring a burst, we post-process the high-resolution raw data obtained from the two cameras to make the images amenable to joint denoising. First, we demosaic the images using method of [7] and subtract the sensor black level to produce linear images in sensor RGB space. We intentionally do not apply white balance gains, a color correction matrix, tone mapping, or gamma compression in order to disentangle our technique from variations in white balance and lighting calibration. We downsample our linear RGB images to 512×512 and compute stereo registration using the "bilateral flow" algorithm of [24], [25], which produces clean, edge-

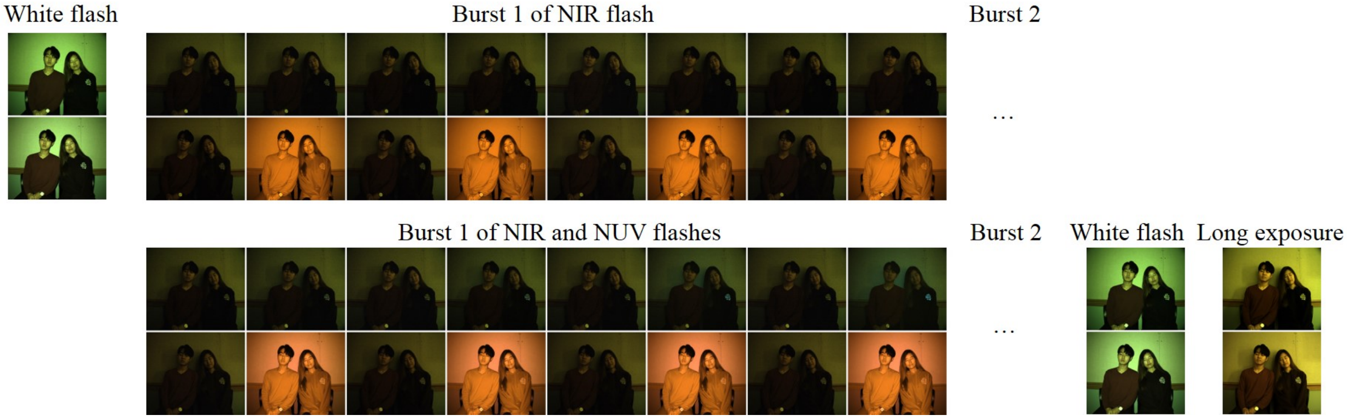


Fig. 5. An example scene from our dataset. We show a series of image pairs where the upper image is from cam_1 (RGB) and the lower image is from cam_2 (NIR-G-NUV). Burst 2 is omitted for space, as it resembles burst 1. The flash images with different exposure time use different analog gains such that they have same level of brightness. These linear images have not been white balanced, hence their green or orange appearance.

aware image alignments that have been demonstrated to work well for computational photography tasks. We use optical flow in this way to circumvent the tedious calibration required by the traditional approach of rectification and stereo depth estimation.

V. Algorithm for Low-light Imaging

This section describes our procedure for fusing a dark flash stereo pair into a high-quality result, as shown in Figures 6 and 7. We first compute a per-pixel registration for the pair using *cam*$_2$ as the base, and then warp *cam*$_1$ accordingly. With this warped image we produce an initial fused result using the scale map algorithm of [19]. Because scale map fusion does not correct for all the tonal and spectral properties of our dark flash image, we then feed the initial fused image along with the warped *cam*$_1$ image to our neural network to produce the final image.

A. Registration

Our imaging configuration requires a per-pixel mapping between the RGB image from *cam*$_1$ and the flash image from *cam*$_2$. Intuitively, we want to preserve the sharp, high-frequency details of the flash image while propagating tonal information from the RGB image, which is lower frequency and more tolerant to error. Therefore, we leave the *cam*$_2$ flash image stationary (the *base*), and compute a flow field that gathers from the *cam*$_1$ RGB image (the *alt*). We employ a variant of the "bilateral flow" algorithm of [24], based on the bilateral solver of [25], to register our stereo pair.

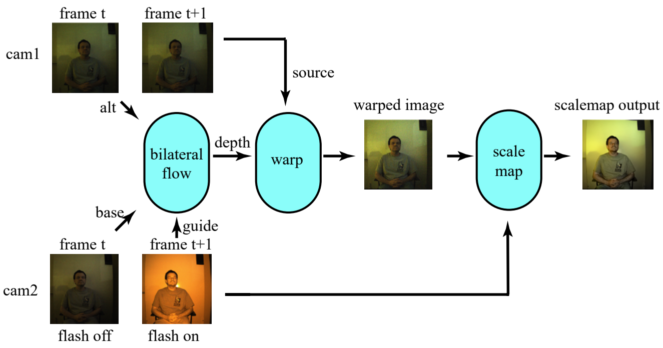


Fig. 6. Our image registration pipeline.

Standard bilateral flow takes as input *base* and *alt*, performs tile matching to compute its data term, and optimizes for a flow field that is smooth while respecting the edges of *base*. We modify standard bilateral flow so that the solution respects the edges of a third *guide* image, and compute tile matching on only the green channels of *base* and *alt*.

If we had access to our ideal camera, the *cam*$_2$ flash image would serve as both *base* and *guide*. To simulate this with the burst dataset captured by our prototype camera, we first compute tile matching on the stereo pair at time t when the flash is off (approximating pure green channels). We then estimate an edge-aware flow field using a bilateral solver, where as the *guide* we use the image produced by *cam*$_2$ at time $t+1$ (when the flash is on). Finally, we use this flow field to warp both the RGB image at $t+1$ and the *cam*$_1$ long-exposure ground truth to the viewpoint of *cam*$_2$. Figure 6 illustrates the data flow.

B. Learned Image Fusion

After registration, for each scene and exposure setting, we have four images from the perspective of *cam*$_2$:

1) A flash image captured by *cam*$_2$.
2) An RGB image warped from *cam*$_1$.
3) A long-exposure RGB image warped from *cam*$_1$.
4) A long-exposure RGB image captured with *cam*$_2$.

The fusion algorithm, given only 1) and 2), needs to generate an image that is close to 4). This is nontrivial, as the flash image 1) looks significantly different from the noisy RGB image 2). Scalemap [19], a representative state-of-the-art algorithm for cross-domain image fusion, can effectively remove much of this noise. However, it also changes the color tone and local contrast of images captured by *cam*$_1$, resulting in unnatural "hazy" images. Skin in particular looks "waxy" (see Figure 8, column 4). But since we collect ground-truth RGB images, we can use modern end-to-end discriminative learning techniques to eliminate these artifacts.

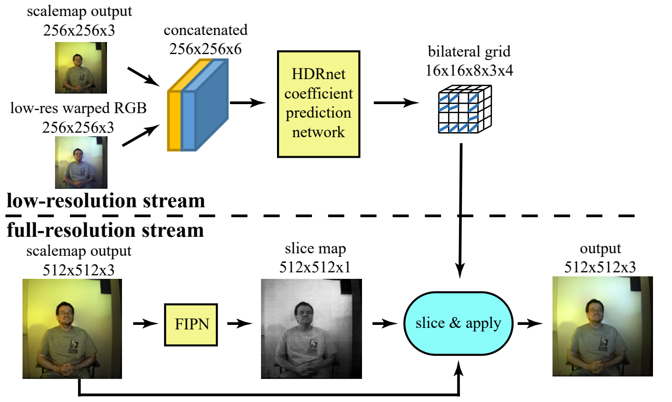


Fig. 7. Our tone correction CNN architecture. Trainable blocks are rendered in yellow and optimized jointly.

One may be tempted to train a general end-to-end model that, from inputs 1) and 2), synthesizes an output resembling 4). However, this places a large burden on the network as it must learn to simultaneously denoise, account for misregistration, and correct for flash shadows in addition to color tone correction, all with a limited dataset. Indeed, our experiments with this approach failed to produce promising results (as shown by the "FIPN" entry in Figure 8). We instead make learning easier by asking the network to only learn color and contrast correction.

Our network takes as input the Scalemap algorithm's output, as these images already have low noise and simply require tonal correction. To ensure that the network exclusively corrects color, and that training does not fail due to stereo misalignment across the input images, for training we use as "ground truth" image 3): the long-exposure RGB image warped from cam_1. This does mean, however, that geometric errors from stereo registration will persist in the output of our model.

Our network is based on "HDRnet" [26], a deep neural network that predicts edge-aware local and global tone correction functions (Figure 7). It consists of a low-resolution stream that predicts an image transformation encoded as an *affine bilateral grid* (a bilateral grid where each cell contains an affine transformation of RGB values), and a full-resolution stream that learns how to best *slice* into the grid and apply the resulting transformation, which is then used to produce the output image.

A straightforward adaptation of HDRnet to our dataset would be to use as input both the flash image 1), the output of Scalemap (fusing the no-flash RGB image 2) with 1)), all concatenated together as a single 6-channel image, and to modify the bilateral grid to contain 3×7 affine transformations accordingly (Figure 7, top). However, we found this to work poorly because HDRnet is unable to express a bilateral grid of affine transformations that removes the shadows cast by the dark flash, in part because HDRnet is designed to "slice" from this grid using only per-pixel luma.

To address this issue, we generalize the model by using another deep network that learns *how to slice* into the bilateral grid (Figure 7, bottom). We replace HDRnet's simple per-pixel network with a significantly more expressive 9-layer fully-convolutional network modeled after the Fast Image Processing Network (FIPN) of [27], to produce a *slice map*. Though the FIPN model is quite general, because we use the output of this model only to slice from a bilateral grid (instead of using it to synthesize a complete image) our HDRnet-like architecture still constrains the output of our complete model to be a local affine transformation of the input image. To accommodate our noisy inputs, we downsample our noisy RGB and Scalemap outputs and concatenate them as input to HDRnet's low-resolution stream. At full-resolution, we predict the slice map using only the Scalemap output. We then slice and apply local 3×4 affine transformations to each pixel. We found that adding the noisy RGB image to the full-resolution stream did not appreciably improve performance. Finally, we replace the luma channel of HDRnet's output with the luma channel from Scalemap, which helps to preserve some details.

After visual inspection, we decided to conduct all experiments only on the $T/5$ subset as it most closely mimics the noise characteristics of modern smartphone cameras. Furthermore, after preliminary experiments, we decided to not use NUV flash and use only NIR. Many materials such as fabric, paper, and glass *fluoresce*–that is, when they absorb UV light, they re-emit that light's energy as visible blue light [28]. Although NUV can clearly help with denoising and many materials such as human skin do not fluoresce, the inconsistencies that fluorescence caused in our dataset made it challenging to achieve good results. The supplement contains one scene depicting fluorescence.

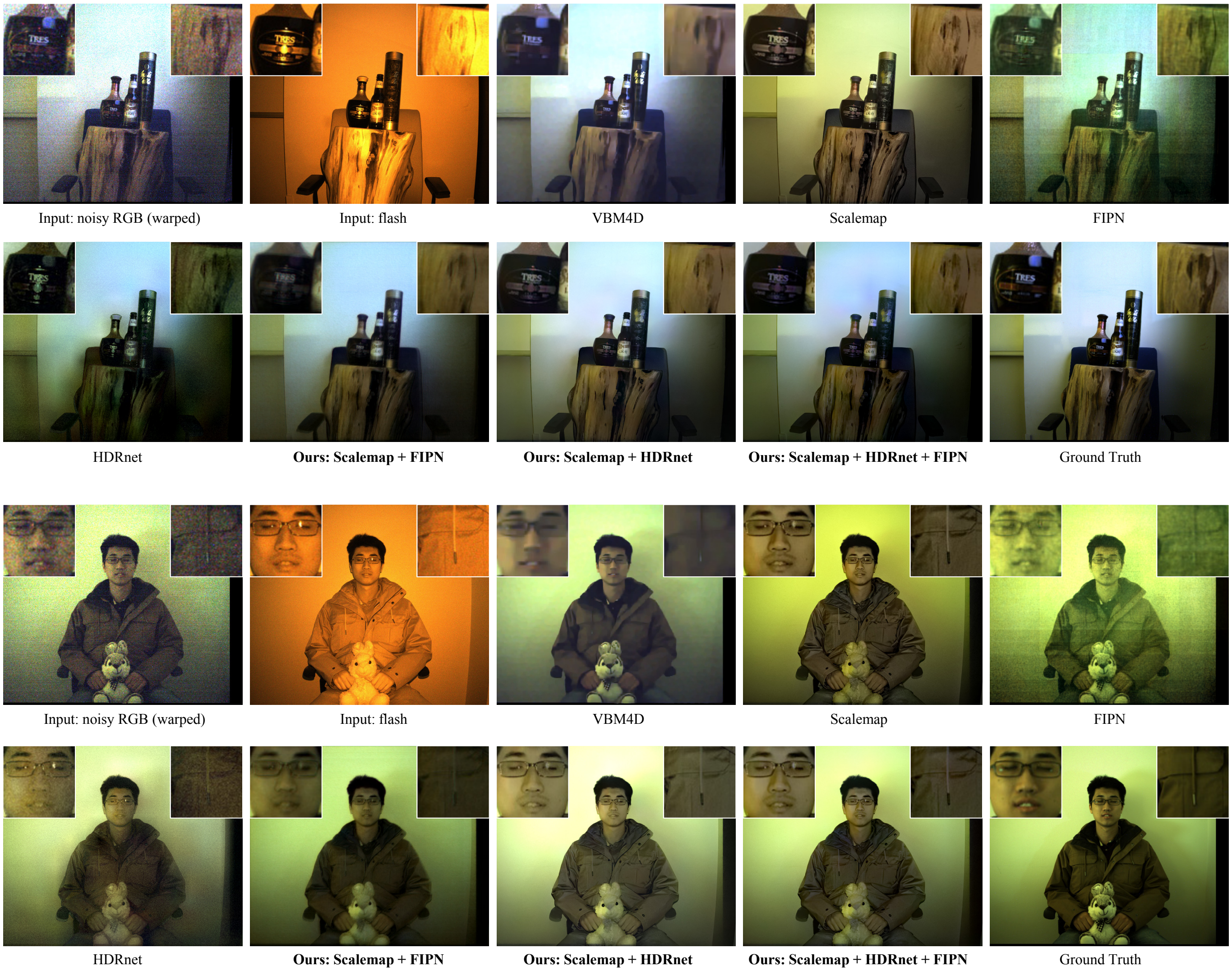
We randomly select 90% of our dataset for training, leaving the rest for testing. Since both HDRnet and FIPN are resolution-independent, we downsample our high-resolution images to 256×256 to accelerate training, but evaluate our results at 512×512. As a baseline, we also train FIPN at 512×512. For all networks, we trained on batches of size 4 using the Adam optimizer [29] with learning rate 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, for 500 epochs. Our model is implemented in TensorFlow.

VI. Results

We evaluate several variants of our method by comparing it to a number of existing baselines:

	Error Metrics		
Method	PSNR	SSIM	Style
Noisy input captured by cam_1	20.37	0.35	4.07
BM3D	22.08	0.79	1.97
VBM4D	22.23	0.79	1.81
Scalemap [19]	20.31	0.68	1.74
HDRnet [26]	22.27	0.58	2.89
FIPN [27]	21.54	0.71	2.32
Ours: Scalemap + FIPN	**27.66**	**0.80**	2.08
Ours: Scalemap + HDRnet	23.16	0.74	1.62
Ours: Scalemap + HDRnet + FIPN	24.47	0.75	**1.43**

TABLE I
Quantitative results under three error metrics, comparing variants of our method against a number of baseline techniques. For PSNR and SSIM, higher is better. For "Style", lower is better.



Input: noisy RGB (warped) Input: flash VBM4D Scalemap FIPN

HDRnet **Ours: Scalemap + FIPN** **Ours: Scalemap + HDRnet** **Ours: Scalemap + HDRnet + FIPN** Ground Truth

Input: noisy RGB (warped) Input: flash VBM4D Scalemap FIPN

HDRnet **Ours: Scalemap + FIPN** **Ours: Scalemap + HDRnet** **Ours: Scalemap + HDRnet + FIPN** Ground Truth

Fig. 8. Our results compared with previous methods and our ground truth on two scenes from our test set. The inputs to our system are a noisy RGB image from cam_1 (visualized here with a digital gain of ×5 for the sake of visualization) and a dark flash image from cam_2. VBM4D denoises a burst of four noisy RGB images. Scalemap [19] fuses the RGB and dark flash images, which removes much of the noise but also results in a color shift and poor local contrast. Direct image synthesis using FIPN reveals significant artifacts from dilated convolution, while unmodified HDRnet can recover the global color tone but fails to improve local contrast over Scalemap. Our neural network restores this detail, resulting in an image substantially close to the ground truth.

- **Input**: we report the error of the noisy RGB image from cam_1 compared to the long exposure image, as a point of reference.
- **BM3D** [30]: a classical single-image denoising algorithm. We use the last no-flash frame from cam_1 (the RGB camera) as input.
- **VBM4D** [31]: a multi-image denoising algorithm. We use all 4 no-flash images from cam_1 (the RGB camera) as input.
- **HDRnet** [26]: a fast and effective neural network model for learning local tone-mapping operators. We directly apply HDRnet to a concatenation of the registered RGB image captured by cam_1 and the flash image captured by cam_2, training the network to approximate the long-exposure RGB image warped from cam_1.
- **FIPN** [27]: a general purpose neural network model for arbitrary imaging transformations. For training, we use the same input/output images as in our HDRnet baseline.
- **Scalemap** [19]: an optimization algorithm designed for fusing RGB and hyperspectral images. We use the warped RGB image captured by cam_1 and the flash image captured by cam_2 as input.

The algorithm we described in Section V-B (which we alternatingly refer to "Scalemap + HDRnet + FIPN", or just "our model") is constructed out of several of our baseline algorithms and trained end-to-end, which we found to produce the most visually pleasing results. We additionally evaluate against two ablations of our model, both of which takes the concatenation of the output by Scalemap and noisy RGB image as input:

- **Scalemap + HDRnet**: Instead of using the output of an FIPN model as the guide map in HDRnet, we use the simple trainable piecewise linear functions originally proposed by Gharbi et al [26].

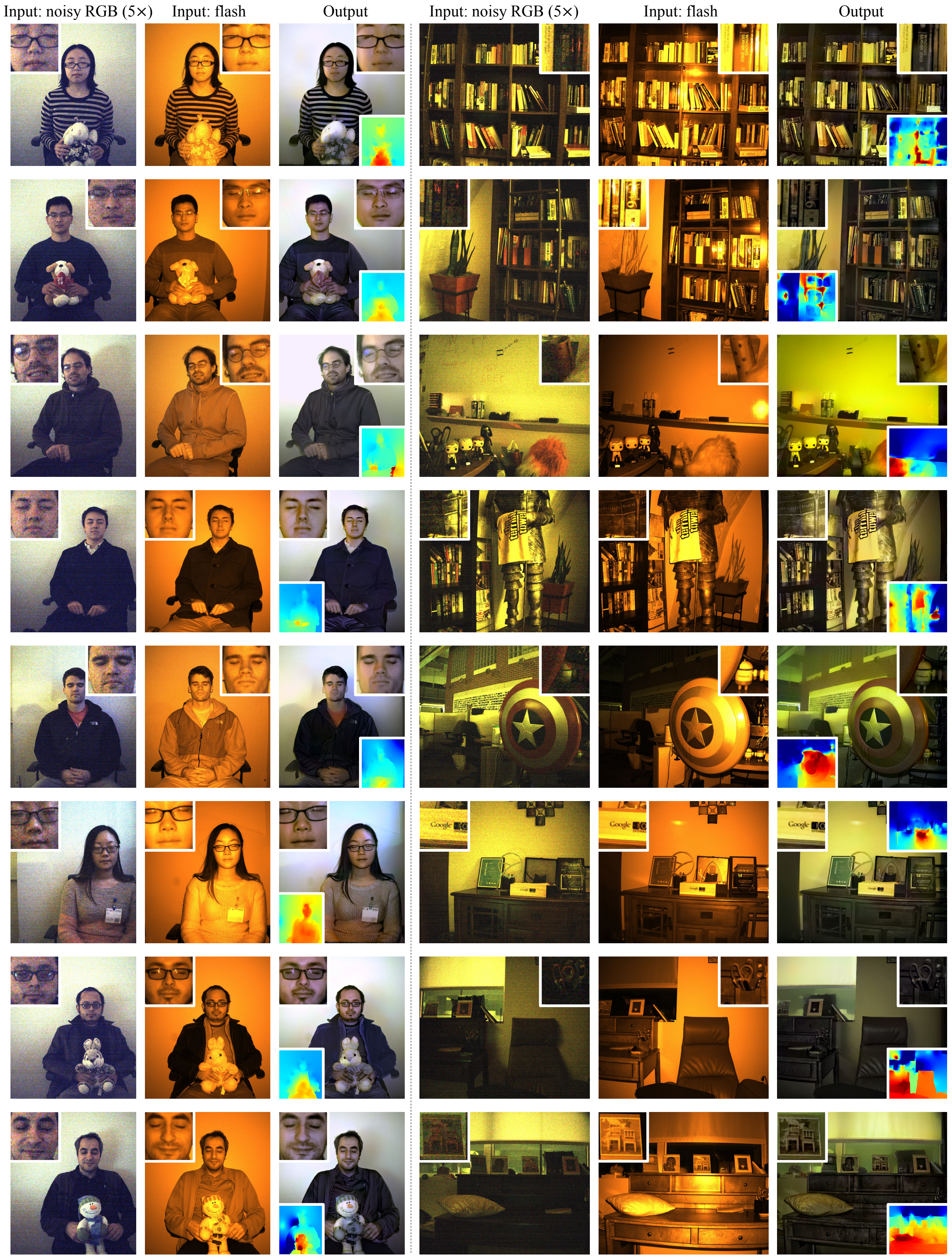


Fig. 9. Inputs and our results on several scenes of both human and still-life subjects. Images are best viewed **zoomed-in** on a computer where noise is visible.

– **Scalemap + FIPN**: Instead of using HDRnet to produce an output image, which restricts the model to only being able to estimate local affine transformations, we train FIPN to directly estimate the output image.

To quantitatively compare the results of our various baselines and model variants to our long-exposure ground truth images, we first rescale outputs to have the same brightness (average RGB value) of the long-exposure ground truth to account for any brightness variation. Then we use three evaluation metrics: PSNR, SSIM [32], and a perceptual error metric (called "Style" in our table). PSNR simply measures any per-pixel differences, while SSIM focuses more on structural differences and is invariant to errors that can be modeled as local shifts or scales. Both of these measures are sensitive to small misalignments in their input images, and because such misalignments are common in our dataset, neither metric is well suited to our task. For this reason, we use the perceptual metric of [33], [34], which is based on *texture similarity* and more forgiving to misalignments. It is commonly used for style transfer applications and is based on the Gram matrix of the feature activations (we used conv2, conv3, and conv4) of a pretrained VGG-16 [35] image classification network.

Table I contains a quantitative evaluation on our test set, and Figure 8 contains a zoomed-in comparison of two examples. The single- and multi-frame techniques for denoising the RGB image(s) without the aid of of the dark flash image tend to generate blurry or oversmoothed output, and as such have lower PSNRs than all other techniques, which do use the dark flash image. Scalemap generates sharp images, but introduces obvious color shifts compared to the ground truth. All three non-learning approaches (BM3D, VBM4D, and Scalemap) have the lowest PSNR in all the approaches, demonstrating the value of learning for this task. Our two learning-based baselines, HDRnet and FIPN, achieve higher PSNR than the non-learning-based approaches, but both of them introduce a significant amount of noise to the output image (Figure 8).

Our "Scalemap + FIPN" ablation achieves the highest PSNR and SSIM. However, upon inspection, we see that its output images are slightly blurred, perhaps due to insufficient training data given the size of the network (see Figure 8). The "Style" metric appears to be sensitive to the artifacts produced by this model and penalizes it accordingly. Both the "Scalemap + HDRnet" ablation and our complete "Scalemap + HDRnet + FIPN" model successfully preserve the color, tone, and contrast of RGB images, while removing most of the high-frequency noise. Our "Scalemap + HDRnet + FIPN" model better preserves local contrast than its "Scalemap + HDRnet" ablation, as can be seen in the texture of the wooden structure in the bottom example of Figure 8. Figure 9 contains additional results on a variety of human and still-life subjects in low-light indoor environments. Our technique produces both a low-noise RGB image as well as a dense edge-aware depth map.

VII. Conclusions

We have presented a design for a stereoscopic dark flash camera, which acquires stereo pairs in which one camera images the complete visible spectrum while the other camera selectively captures some visible and some hyperspectral light. When paired with a hyperspectral flash this camera configuration allows for the acquisition of "dark flash" images even in the presence of motion, thereby allowing for low-noise photography in low-light environments, without disturbing human subjects with a dazzling flash. To this end we have constructed a hardware prototype that approximates our idealized camera configuration and a dataset acquisition procedure that circumvents the shortcomings of our hardware prototype while also capturing ground truth long-exposure images. With the goal of fusing our dark flash stereo pairs into a low-noise and visually pleasing image, we have presented a set of novel deep neural network architectures which we train end-to-end to regress from dark flash stereo pairs to the long exposure RGB images in our dataset. We show that these fused images have the low-noise properties of our dark flash image, while retaining the aesthetically pleasing tonal properties of our noisy no-flash RGB images.

Acknowledgment

The authors would like to thank Nori Kanazawa, Roman Lewkow, Dilip Krishnan and Marc Levoy for their help and constructive discussions. We would also like to thank our volunteers for participating in experiments.

References

[1] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *SIGGRAPH*, 2016.

[2] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs," *ACM TOG*, 2004.

[3] E. Eisemann and F. Durand, "Flash photography enhancement via intrinsic relighting," *ACM TOG*, 2004.

[4] D. Krishnan and R. Fergus, "Dark flash photography," *SIGGRAPH*, 2009.

[5] J. T. Barron, A. Adams, Y. Shih, and C. Hernández, "Fast bilateral-space stereo for synthetic defocus," *CVPR*, 2015.

[6] B. E. Bayer, "Color imaging array," 1976, uS Patent 3,971,065.

[7] H. S. Malvar, L.-w. He, and R. Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," *ICASSP*, 2004.

[8] J. Jiang, D. Liu, J. Gu, and S. Süsstrunk, "What is the space of spectral sensitivity functions for digital color cameras?" *WACV*, 2013.

[9] D. Litwiller, "Ccd vs. cmos," *Photonics spectra*, 2001.

[10] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE TIP*, 2007.

[11] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE TIP*, 2006.

[12] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," *CVPR*, 2014.

[13] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," *CVPR*, 2018.

[14] Z. Liu, L. Yuan, X. Tang, M. Uyttendaele, and J. Sun, "Fast burst images denoising," *SIGGRAPH Asia*, 2014.

[15] K. Dabov, A. Foi, and K. Egiazarian, "Video denoising by sparse 3d transform-domain collaborative filtering," *Eusipco*, 2007.

[16] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajk, D. Reddy, O. Gallo, J. L. abd Wolfgang Heidrich, K. Egiazarian, J. Kautz, and K. Pulli, "Flexisp: A flexible camera image processing framework," *SIGGRAPH Asia*, 2014.

[17] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *SIGGRAPH*, 2006.

[18] K. He, J. Sun, and X. Tang, "Guided image filtering," in *ECCV*, 2010.

[19] X. Shen, Q. Yan, L. Xu, L. Ma, and J. Jia, "Multispectral joint image restoration via optimizing a scale map," *IEEE TPAMI*, 2015.

[20] X. Guo, Y. Li, and J. Ma, "Mutually guided image filtering," *ACM Multimedia*, 2017.

[21] N. Spooren, B. Geelen, K. Tack, A. Lambrechts, M. Jayapala, R. Ginat, Y. David, E. Levi, and Y. Grauer, “Rgb-nir active gated imaging,” *Electro-Optical and Infrared Systems: Technology and Applications XIII*, 2016.

[22] N. Tack, A. Lambrechts, P. Soussan, and L. Haspeslagh, “A compact, high-speed, and low-cost hyperspectral imager,” *Silicon Photonics VII*, 2012.

[23] B. Geelen, C. Blanch, P. Gonzalez, N. Tack, and A. Lambrechts, “A tiny vis-nir snapshot multispectral camera,” *Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VIII*, 2015.

[24] R. Anderson, D. Gallup, J. T. Barron, J. Kontkanen, N. Snavely, C. Hernández, S. Agarwal, and S. M. Seitz, “Jump: Virtual reality video,” *SIGGRAPH Asia*, 2016.

[25] J. T. Barron and B. Poole, “The fast bilateral solver,” *ECCV*, 2016.

[26] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, “Deep bilateral learning for real-time image enhancement,” *SIGGRAPH*, 2017.

[27] Q. Chen, J. Xu, and V. Koltun, “Fast image processing with fully-convolutional networks,” *ICCV*, 2017.

[28] J. Lakowicz, “Principles of fluorescence spectroscopy,” 1999.

[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[30] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Bm3d image denoising with shape-adaptive principal component analysis,” *SPARS*, 2009.

[31] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, “Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms,” *IEEE TIP*, 2012.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE TIP*, 2004.

[33] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[34] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” *ECCV*, 2016.

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv:1409.1556*, 2014.