Seeing in Extra Darkness Using a Deep-Red Flash

Jinhui Xiong[1*] Jian Wang[2*] Wolfgang Heidrich[1] Shree Nayar[2]
[1]KAUST [2]Snap Research

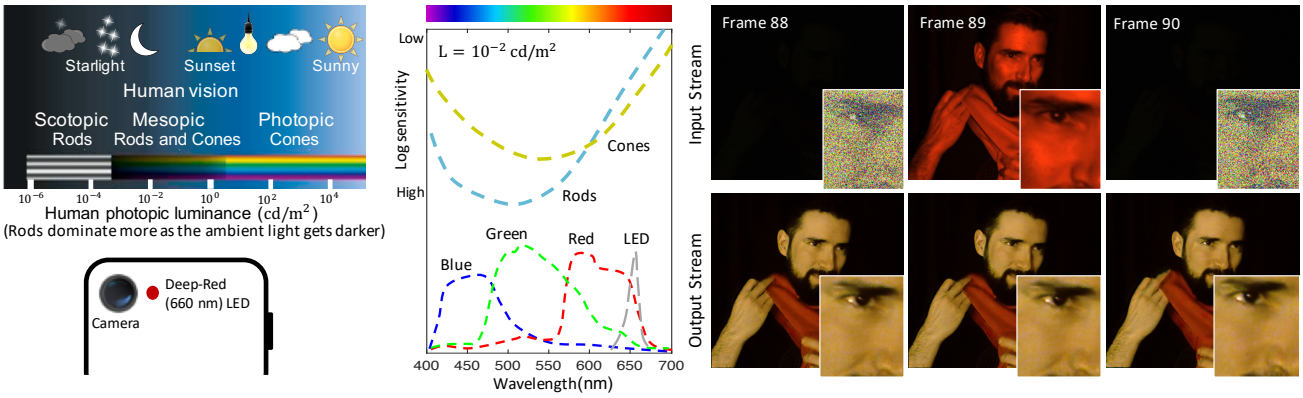


Figure 1: Top left: Human vision uses cones and rods for the perception of light. Photopic vision is associated with cones, occurring at bright-light conditions (over 3 $\mathrm{cd/m^2}$). Scotopic vision is associated with rods, occurring at dim-light conditions (below 10^{-3} $\mathrm{cd/m^2}$). At intermediate light levels, both rods and cones are active, which is called mesopic vision. Bottom left: We propose to use deep-red (e.g. 660 nm) light as flash for low-light imaging in mesopic light levels. This new flash can be introduced into smartphones with a minor adjustment. Middle: The eye spectral sensitivity in a dimly lit environment (0.01 $\mathrm{cd/m^2}$) and the relative responses of R, G and B color channels of the camera we used, as well as the emission spectrum of the red LED flash. Under dim lighting, rod vision dominates, *yet the rods are nearly insensible to deep-red light*. Meanwhile, our LED flash can be sensed by the camera especially in the red and green channels. Right: Inputs to our videography pipeline are a sequence of no-flash and flash frames, and the outputs are denoised and would yield temporally stable videos with no frame rate loss. Top left figure is reproduced from [20]. Data for cones and rods are from [38] and [17].

Abstract

We propose a new flash technique for low-light imaging, using deep-red light as an illuminating source. Our main observation is that in a dim environment, the human eye mainly uses rods for the perception of light, which are not sensitive to wavelengths longer than 620 nm*, yet the camera sensor still has a spectral response. We propose a novel modulation strategy when training a modern CNN model for guided image filtering, fusing a noisy RGB frame and a flash frame. This fusion network is further extended for video reconstruction. We have built a prototype with minor hardware adjustments and tested the new flash technique on a variety of static and dynamic scenes. The experimental results demonstrate that our method produces compelling reconstructions, even in extra dim conditions.*

*denotes equal contribution. Part of the work was done while Jinhui Xiong was an intern in Snap Research.

1. Introduction

Low-light imaging has been a critical capability for smartphone cameras. Existing solutions range from improved sensor design such as back-illuminated sensors [40], to the use of different color filter arrays (e.g. RYYB instead of traditional RGGB Bayer filters [32]). Furthermore, computational photography techniques like burst denoising [16] are widely deployed by many companies.

Flash photography also has a long history and generally provides the best results, especially for very low light levels or scenes with complex motions, for which burst-denoising image alignment might fail. However, flash photography also has several downsides. The intensity falls off quadratically with distance from the flash, making it difficult to shoot well-exposed photos with a large depth range. Moreover, the flash itself is dazzling to human eyes. Especially in very dark environments, a strong white light causes unpleasant light pollution and may destroy the dark adaptation of the human visual system. Using invisible flashes

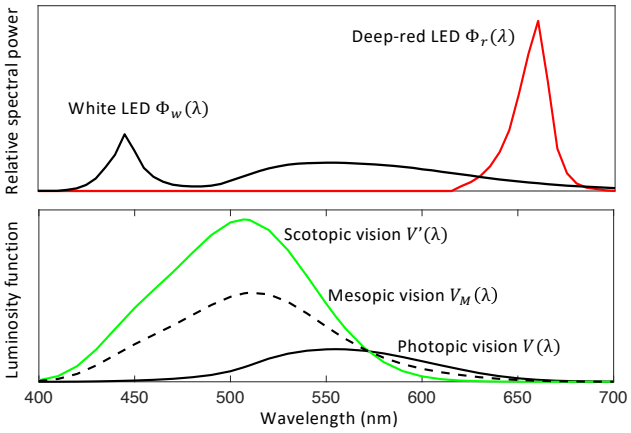
(the Near Infrared (NIR) or the Near Ultraviolet (NUV) flash) [21] can avoid this disturbance. However, RGB cameras do not have sensitivity to the invisible spectrum; thus either an additional NIR and NUV sensible camera or an IR-cut filter switch should be built into mobile devices. The additional camera or mechanical shutter is not favored in the current smartphone design, which is compact and has limited real estate. Moreover, the image structures between RGB and NIR-NUV images could contain significant discrepancies due to wavelength-dependent reflectance. It makes both cross-modal image registration and image fusion challenging, especially in a dim environment where the RGB image is highly corrupted. Using NIR in smartphones is controversial and raises privacy issues as it could see through some clothes (e.g. synthetic fabrics).

1.1. Human Visual System

We explain how the human visual system is affected by a white flash, and propose a new flash to avoid the downsides.

The retina is the part of the human eye responsible for the perception of light and is composed of two basic types of photoreceptors – *cones* and *rods*. The cones function in bright-light conditions and are responsible for the perception of color. Their peak spectral sensitivity is at around 550 nm. The rods become active in dim-light conditions which only provide black-and-white vision. They are most sensitive to bluish-green wavelengths at around 500 nm, and insensitive to long-wavelength light; their sensitivity to a 650 nm light is about 3 orders of magnitude lower to that at 500 nm (see e.g. [15]). A combination of cones and rods forms mesopic vision as illustrated in Fig. 2 middle (the dashed line). Given the photopic and scotopic luminosity functions as $V(\lambda)$ and $V'(\lambda)$, the mesopic luminosity function $V_M(\lambda)$ is a blend of $V'(\lambda)$ and $V(\lambda)$, which can be approximated as $(1-x)V'(\lambda) + xV(\lambda)$, where x is determined by photopic luminance and the wavelength of light.

When the ambient light changes, the human eye will adjust the visual system to adapt to the change in luminescence. This does not, however, happen instantaneously. The transition from day to night vision is called dark adaptation and undergoes a slow process of accommodation as shown in Fig. 3 (an elaborated rendering model can be found in [10]). In a brightly lit room, the human visual system has a good sense of color and high spatial acuity. When the light is off, the human eye is temporarily blind and the visual threshold drops rapidly but stays at a relatively high level; cones reach their greatest sensitivity. After about 9 minutes, the sensitivity of the rods exceeds that of cones; the visual system undergoes a transition from cone vision to rod vision. This point is called *Purkinje shift*, after which the threshold drops again; the human eye achieves night vision with low acuity and nearly no color sense. Rhodopsin, a light-sensitive receptor protein found in rods, enables human vision in extra-dark conditions. The opposite adjustment is called light adaptation, and in contrast, it occurs in a much shorter period. We refer interested readers to [15] for additional details. Rapid light adaptation, which happens when using white flashes, would also cause bleaching in photoreceptors, leading to discomfort or even disability glares [9, 3]. When exposed to light, rhodopsin immediately photobleaches and loses its sensitivity to light, causing flash blindness. Rhodopsin regenerates after bleaching following an exponential time course and takes minutes for a full recovery of rod response [2].



Luminance (cd/m^2)	x (White LED)	x (Red LED)	Brightness gain	or	Signal-to-noise gain
0.005	0.04	0	1/353.47		353.47
0.01	0.13	0	1/338.45		338.45
0.05	0.33	0.22	1/25.22		25.22
0.1	0.42	0.34	1/15.98		15.98

Figure 2: Top: The normalized spectral power distributions of our employed red and white LEDs, denoted as $\Phi_r(\lambda)$ and $\Phi_w(\lambda)$. Middle: Photopic ($V(\lambda)$), mesopic ($V_M(\lambda)$), and scotopic ($V'(\lambda)$) luminosity functions. *The perceived power of light by human mesopic vision* is proportional to the inner product of the luminosity function $V_M(\lambda)$ with the spectral power distribution $\Phi(\lambda)$, written as $\int V_M(\lambda)\Phi(\lambda)d\lambda$. Bottom: Some values of x reported in MOVE mode [13] are listed. *Brightness gain* is computed as the ratio of the luminous flux from the red flash to the white flash, where the total power received by the camera is the same for both flashes. *Signal-to-noise gain* is the ratio of the total received signals by the camera when using the red flash versus using the white flash, which has the same brightness to the human eye. In dim conditions around 0.01 cd/m^2, using white flash is over **two orders of magnitude brighter** than using the red flash as perceived by the human eye when the camera receives the same amount of signals; or camera receives **two orders of magnitude more** signals when they have the same brightness.

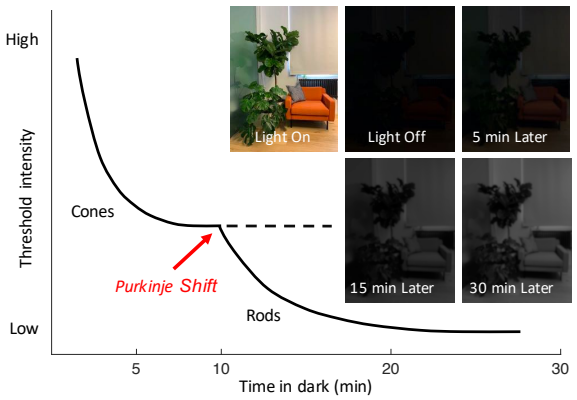


Figure 3: A simulation of dark adaptation. It takes around 30 minutes for the rods to reach a fully functional stage.

1.2. Deep-Red Flash

We address the fundamental issues that have arisen in alternative flash approaches and propose to use deep-red flash for scenarios in which the human eye works in the mesopic regime. We can quantitatively compare the *perceived brightness* using the white flash with the red flash at different mesopic luminance. The numerical results are reported in Fig. 2 bottom. This new method is termed as *mesopic flash photography*. **Compared to using white flashes**, 1) the perceived luminous flux is significantly reduced, which could easily reach one to two orders of magnitude lower, and thus distraction or discomfort glare is significantly less likely, and 2) night vision is also preserved as long-wavelength lights prevent rod cells from bleaching. **Compared to using invisible flashes**, a single RGB camera is capable of capturing both flash and no-flash image pairs, unnecessary to modify the camera hardware or to use an extra camera. Moreover, the deep-red flash is in the range of visible wavelengths, therefore the structural discrepancy is mitigated; image fusion becomes easier and more robust. A conceptual comparison could be found in Fig. 4. In specific, we make the following contributions:

- We propose a novel deep-red flash for low-light imaging based on the mechanism of human visual system.
- We introduce a modulation operation during the training phase for a guided image filtering network, and the network could better exploit the high frequency components in the guide frame.
- We build a prototype and validate the proposed flash method on static and dynamic scenes. Our method produces compelling results even in extra-dim conditions.

2. Related work

Image denoising has long been studied to suppress unsatisfactory observations in captured images. Conventional approaches rely on image priors, for instance Total-Variation [39, 35], low rank structure [12], or self similarity [4, 8]. With the success of applying neural networks in high-level image understanding, they have been increasingly popular for low-level image reconstructions [18, 44].

Low-light imaging (without using a flash) is an increasingly important feature of current smartphone cameras. Fusing a burst of noisy images [27, 16, 26] to obtain a clear image is favored by most cellphone companies due to its reliability, relatively runtime efficiency and robustness. However, adequate illumination levels are still required, otherwise, the registration of the images becomes problematic. Li et al. [24] propose to fuse a simultaneously captured RGB image and a black-and-white monochrome image from a two-camera module. This camera module imitates how cones and rods cooperate in the human visual system. A precise pixel-to-pixel alignment is also required to fuse these two images, which is problematic at very low light levels.

More recent approaches tackle the low-light imaging problem using modern neural network models on either processed images by Image Signal Processors (ISPs) [29] or directly on raw sensor data [6, 5, 19]. Among them, processing on raw data has distinct advantages owing to the reduced quantization error and higher dynamic range. However, just like other software-only methods, these approaches still fail in very dim environments, where the raw data suffers from very a poor signal-to-noise ratio.

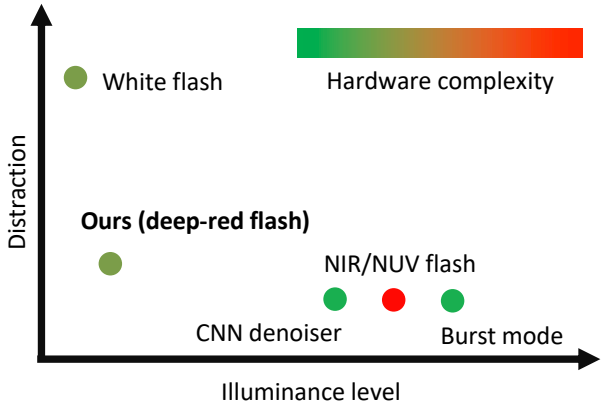


Figure 4: A conceptual comparison for modern low-light imaging approaches. Using deep-red flash can realize low-light imaging in extra dark conditions, and greatly mitigates unwanted distraction with a comparably easy hardware adjustment. Flash and no-flash methods are suited for different low-light situations. Thorough explanations for these methods can be found in Sec. 2.

Flash photography has a long history dating back to the early days of photography, and is still prevalent in modern devices including cellphones. However, flash images de-

stroy the ambience of the scene and are therefore not often desirable. This has lead to the development of **Flash/no-flash techniques** [11, 33, 1], which aim to fuse the detail of flash images with the ambience of noisy environmentally lit images. While this approach improves the image quality aspects of flash photography, the visually disruptive nature of a white flash remains, as detailed in Sec. 1.

To avoid the intrusive burst of white flash, attempts have been made by applying invisible spectrums to the human eye for flash (NIR or NUV or both), known as dark flash. Krishnan et al. [21] formulate an optimization framework by exploiting the gradient in the dark-flashed image as guidance to remove the noise in the RGB image.

However, the NIR image and RGB image cannot be captured at the same camera configurations, which means physically switching the NIR-cut filter on and off and capturing asynchronous image pairs or using an additional camera for NIR and NUV sensing is required [43]. The first option requires highly sophisticated solutions to add the mechanical shutter to smartphone cameras. While both would require precise registration between the dark-flash image and no-flash image, which is challenging due to the intrinsic structural inconsistency, texture loss [7], and it offers no reliable solutions when the RGB image is severely noisy at rather low-light conditions. Although RGB-NIR sensors [28, 42] are available for simultaneous acquisitions of the RGB and NIR signals, the inevitable photon noise and spectral cross talk make it an ill-posed inverse problem to extract true RGB color channels from the contaminated sensor data.

Red light illumination has a long history of being used in dark environments, as it enables people to perceive high spatial frequencies with cone-based vision, while maintaining the dark adaptation of the rods. The aircraft cockpit is illuminated with red light to ensure that pilots can see the instrument panel while maintaining their night vision when it is dark outside. In some animal research laboratories, red light is used to mimic darkness for animals who are insensible to long-wavelength light, at the same time permitting human researchers to continue their work, which would be impossible in the dark [14]. More recently, long-wavelength light has also been utilized in display devices to reduce discomfort glares [30] and has been built into some modern cameras for low-light auto-focus assist.

3. Camera and Flash Module

Fig. 5 represents a prototype system for our proposed mesopic flash videography. A deep-red LED (its spectral power distribution refers to Fig. 2) is placed near the camera to minimize shadow effects. We chose this LED since the rods are not sensitive to it, while the camera sensor still has an adequate spectral response. The LED is triggered by signals from the camera such that it can be synchronized. We employ a Basler acA2040-120uc USB 3.0 camera. The outputs from the camera are 8-bit or 12-bit raw data. An example of processed (white balance, demosaic) no-flash and flash video frames is shown in Fig. 1 upper right.

Camera module	Basler acA2040-120uc
Deep-red LED	Osram Sylvania LZ4-40R208-0000
LED driver	SparkFun COM-13716
Micro-controller	Teensy 3.2
Power	XP Power VER36US240-JA

Figure 5: Left: The hardware prototype. Right: The table lists the hardware components used in our prototype.

4. Mesopic Flash Reconstruction

We first describe how to select the guide signals from red-flash images, and then introduce the model we used to fuse no-flash and flash images. We then generalize the fusion algorithm to no-flash and flash video frames which do not have to be well aligned.

4.1. HDR Guide Signal

We aim to exploit the guide signals from the red-flash images. The red channel of our employed camera is roughly 4 times more sensitive than the green channel and 10 times more sensitive than the blue channel at the wavelength of $660\ \mathrm{nm}$ (also see Fig. 1, center). One straightforward option is to select the red channel signal as the guide signal for reconstruction. The measurable dynamic range, however, is limited as we learn it is easy to arrive at saturation for the reddish objects and still receive low signals for bluish objects in the red channel.

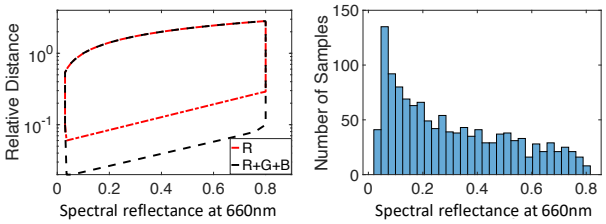


Figure 6: Left: The measurable range when using red channel only and the summation of RGB color channels, defined by the reflectance at a wavelength of $660\ \mathrm{nm}$ and the relative distance to the light source. Right: histogram of reflectance values at $660\ \mathrm{nm}$ for 1269 Munsell color chips.

To analyze the brightness of different real-world materials under the red LED illumination, we analyzed the spectra of 1269 Munsell color chips that represent a majority of natural materials. We considered the reflectance of each chip at $660\ \mathrm{nm}$ in an ideal situation where the light source is constant and the object surface is perpendicular to the light

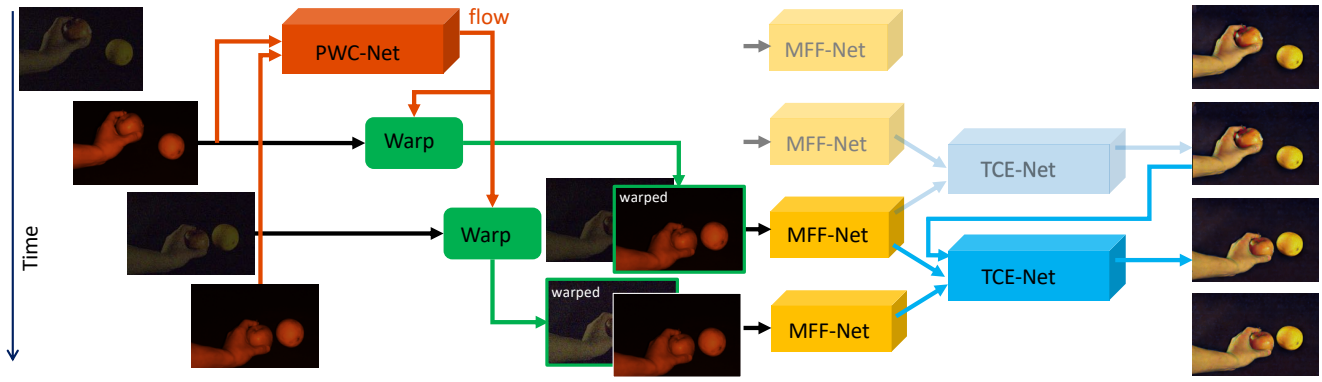


Figure 7: The pipeline for mesopic flash videography. It consists of three parts: intermediate frame synthesis for both red-flash images and no-flash images (PWC-Net + Warp), guided image fusion by our proposed merging method (MFF-Net), and temporal consistency enhancement (TCE-Net).

source. An object is considered visible if the reflected signal is within the camera's measurable range. The measured pixel intensity depends on the reflectance of the materials and the distance to the light source. The left side of Fig. 6 shows that using the red channel may become saturated for close objects (small relative distance) with high spectral reflectance in the red wavelengths. This causes edge information to be lost in the red channel. On the other hand, the edge can still be observed in the green or blue channel due to their relatively lower sensitivity to red light. Simply summing up the three color channels as the guide signal can increase the dynamic range. For natural materials, the right side of Fig. 6 shows the histogram of reflectance values for $660\ \mathrm{nm}$ light, demonstrating that about 80% of materials have a reflectance at least 0.1. This affirms the practical applicability of the employed deep-red flash.

4.2. Mesopic Flash Image Fusion (MFF)

Given a pair of flash and no-flash images, the ultimate goal is to extract edges from the flash image, color from the ambient light image, and blend them to generate a noise-reduced image. A desired fusion process should take the features of image spatial structure into account, meaning that reconstructed pixel intensities within the same image structure will be homogeneous. The U-net architecture [34] has shown excellent performance in a number of image reconstruction and style transfer tasks. Constructed by a pyramid of encoders and decoders such that the reconstructed pixels have a good sense of its neighboring pixels, ideal for our purposes. Therefore, we train a neural network in ResUnet architecture by adding residual blocks to the U-net. The input to the neural network is the concatenation of the no-flash image and its associated guide signal, passing three consecutive encoders, 16 residual blocks, and then three consecutive decoders. The final output is a denoised image in RGB color channels.

4.3. Videography Pipeline

To generalize mesopic flash photography to videography, the input becomes a sequence of flash and no-flash image pairs recorded by the same camera, and the output is the denoised video stream. A simplistic approach is to fuse the image pair and produce one denoised image frame akin to photography. However, this decision generates a frame rate reduced by half and in conjunction with camera shake and moving objects a precise image alignment between the image pairs becomes challenging. We integrate our image fusion network with a state-of-the-art flow computation network (PWC-Net [41]) and temporal consistency enhancement network (TCE-Net [22]). We propose an effective videography pipeline to address robustness issues in image alignment while preserving the original frame rate.

The videography pipeline is shown in Fig. 7. We first compute the backward optical flow between two consecutive guide signals (separated by a no-flash frame). This robust computation takes into account the guide signals which have a relatively high signal-to-noise ratio and are captured under the same illumination conditions. We then forward-warp the preceding guide signal to the no-flash frame using one-half of the estimated optical flow field. We also forward-warp the no-flash frame to the next flash frame using the other half of the optical flow field. By repeating the above procedures on the video stream, all no-flashed frames are paired with the synthesized guide frames, and all guide frames are paired with synthesized no-flash frames. In the subsequent step, we apply MFF-Net to each pair. The output video stream is therefore at the original frame rate. To alleviate flickering artifacts, we further feed the processed frames into the TCE-Net. It takes a video as input and produces one output frame at each time step that is temporally consistent with previously generated outputs under the learned short-term and long-term temporal loss. This

output is then fed into the network as part of the input for the subsequent process.

5. Experiments

5.1. Training

Modulator for guide signal. Given a set of clear training RGB images (we use NYU v2 dataset [31] for our training purpose; training on other datasets exhibits similar performance, and please see the Supplement for additional results). A straightforward way to generate training data is to synthetically introduce camera noise to the RGB images and treat the summation of RGB channels as the guide signal, concatenating synthetically generated low-light images and guide signals as input. The original RGB images are then regarded as ground truth. Under this training configuration, the network fails to perform the desired guided image fusion as it will directly transfer the intensity information in the guide signal to the red, green and blue channels of the generated output.

However, we only desire the transfer of the edge information in the guide frame. We, therefore, introduce a specific modulator for this purpose, which is a low-frequency sinusoidal function with random period and amplitude (see Fig. 8). The modulator can be expressed as:

$$f(x,y) = \alpha \cdot \sin(\frac{2\pi}{T}\sqrt{(x-\bar{x})^2 + (y-\bar{y})^2}) + \beta,$$

where $x \in \{1, 2, ...W\}$ and $y \in \{1, 2, ..., H\}$ (W and H are the width and height of images, respectively.). α is the amplitude, T is the period, $\bar{x}$ and $\bar{y}$ are the phase shift, and β is the vertical shift. All parameters are randomly chosen.

Figure 8: Applying a modulation function on the guide frame of one of the training images.

Applying this modulation function to the guide signal decorrelates the intensity in the guide and output images, while still retaining the edge correlation. This helps the network learn to exploit edge information instead of relying directly on the guide signal's intensity.

Loss functions. The perceptual loss from a pre-trained VGG16 network [37] has been widely adopted in image reconstruction tasks due to its ability to recover finer details and produce sharper outputs compared to per-pixel losses. It measures the high-level perceptual differences between the outputs and the ground truth labels, fitting well with human visual perception. However, perceptual loss alone results in degraded color fidelity. A weighted combination of the perceptual loss and ℓ_2 (MSE) loss delivers both visually pleasing and color-accurate results.

Ablation studies. To validate the proposed modulation strategy and the selection of the loss functions, we perform an ablation study. We use a set of low-light images and corresponding red-flash images for validation. Images captured under strong ambient light served as ground truth. Table 1 reports the reconstruction accuracy with different settings. With the employment of the modulation strategy, our network could achieve significantly improved image restoration results and generalize well to real captured data. Please see the Supplement for qualitative comparisons.

Table 1: Quantitative results on different training settings.

Modulation	L2 loss	VGG loss	PSNR/SSIM
×	×	✓	23.34/0.59
×	✓	✓	23.82/0.59
✓	×	✓	26.73/0.71
✓	✓	✓	**26.89/0.72**

5.2. Image Fusion

When to use the flash? We first compare the proposed flash method against state-of-the-art (SOTA) no-flash approach SID [6], which trained an end-to-end denoising neural network dedicated to low-light imaging. Fig. 9 shows the visual comparisons, where the illuminance level is around 0.1 lux (please see Supplement for an explanation of lux and cd/m^2). Our output closely matches the reference image concerning the image details and color. In comparison, SID fails to produce satisfactory images under such a severe noise level. *In extra dim environments, which are beyond the capability of no-flash methods, it is essential to use flashes to capture interested scene details.* A flash can easily boost the light level by $100 - 1000\times$.

Comparison to other image fusion methods. We further compare our reconstruction results to SOTA learning-based joint image filtering network DJF [25], as well as SOTA model-based image fusion approach Scale Map [36]. The inputs to all methods are the same, a guide image and a noisy image. We test the algorithms on data with different bit-depths. Low-light images are usually corrupted by Poisson noise, readout noise (named as Gaussian noise), and quantization noise. When the bit-depth is low, quantization noise tends to dominate for small pixel values. This quantization noise is mitigated by high-bit images, whose noise will be Gaussian-dominated as a large analog gain is favored when capturing low-light images. Fig. 10 indicates

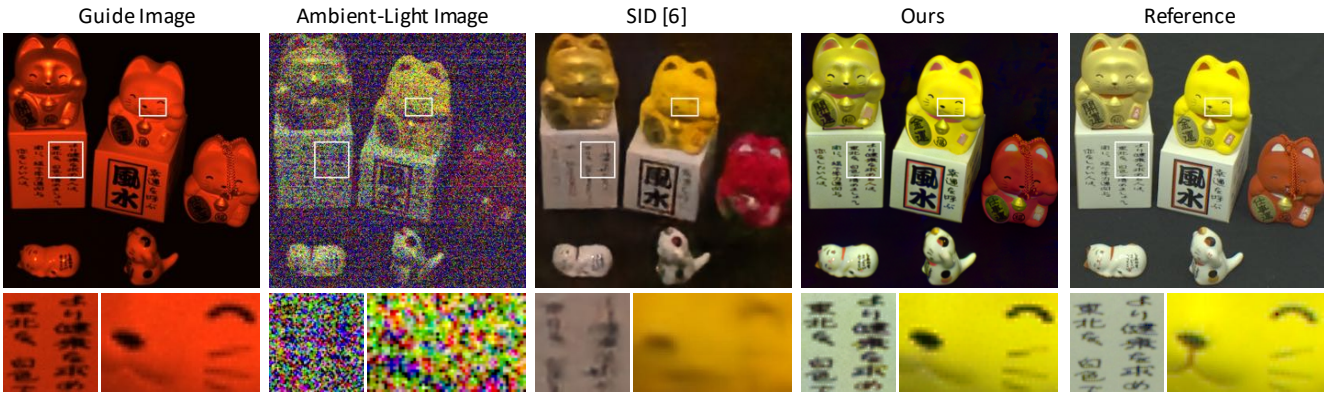


Figure 9: Visual comparisons between SID [6] and ours on 12-bit raw input. Without guide information, the output from SID is highly blurred and image details are significantly lost. Using a flash could preserve fine details even when camera captures poor ambient-light signals. The image taken at good illuminance is shown as a reference. We recommend a zoom-in view.

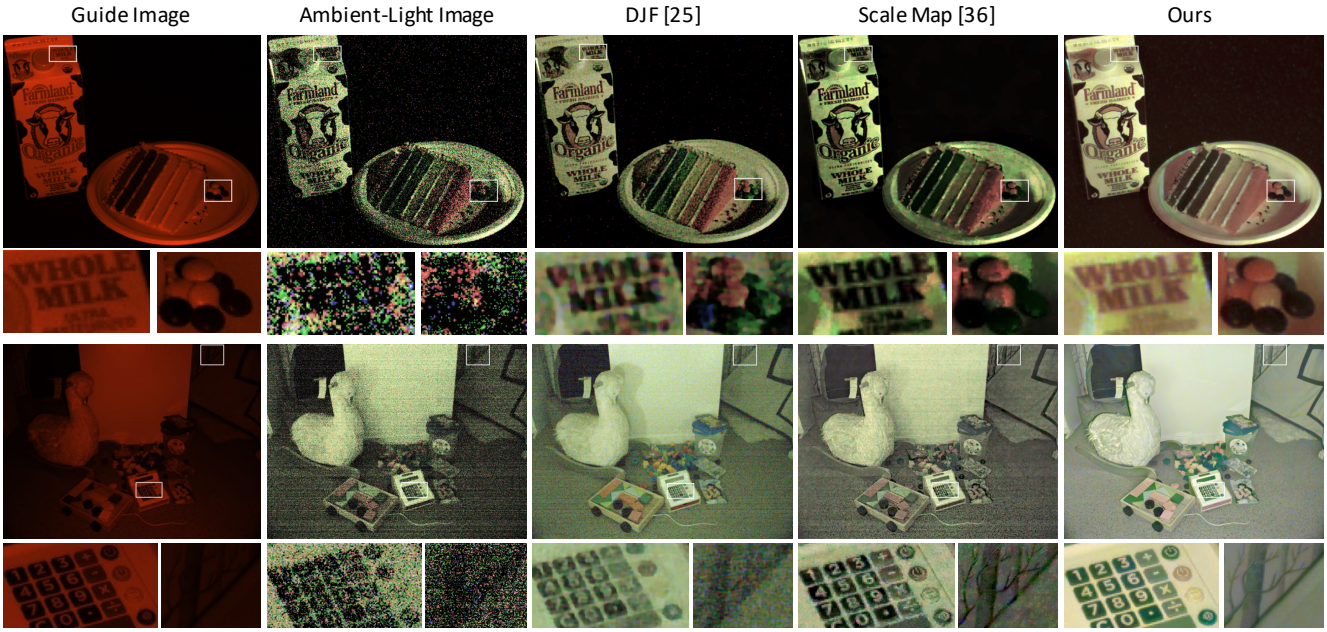


Figure 10: The visual results for the recovered low-light images from a 8-bit image (first row) and a 12-bit image (second row), with their associated guide images. DJF [25] and Scale Map [36] fail to deliver good recovery when the ambient-light images are badly corrupted. Our results exhibit remarkable image reconstructions at various noise models (quantization-noise dominated in the first row and Gaussian-noise dominated in the second row). The second example demonstrates falloff of the flash, whereas our model exploits the edges not the brightness from the flash frame when realizing image fusion.

that DJF fails to retrieve fine image details under both severe noise models. Scale Map recovers slightly more image details, while still fails to produce compelling results in these rather challenging situations, whereas our method produces clean outputs with fine image features recovered.

Quantitative comparisons. We also report the quantitative results in terms of PSNR, SSIM and VGG loss in Table 2, comparing our method with SID and Scale Map. Raw images are captured either in 8 bits or in 12 bits. Working with high bit-depth images shows significant advantages in low-light imaging. At illumination levels of 0.1 lux or less, the majority of the image pixels are 0 when using 8-bit raw data; using 12-bit raw data can still realize adequate recovery. At various low-light conditions and bit-depths, our method consistently exhibits superior performance.

5.3. Video Reconstruction

Fig. 1 right and Fig. 11 show the reconstructed consecutive frames in the video stream for dynamic scenes. The even-number frames are the ambient light images and

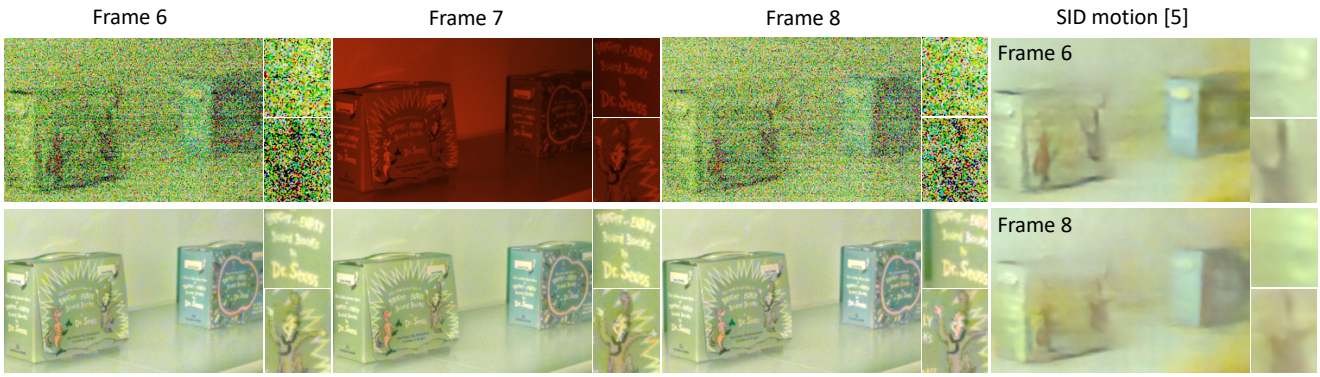


Figure 11: Reconstructed video frames for dynamic scenes. We include the denoised results from SOTA no-flash video reconstruction method SID motion [5] for reference, where image details are significantly lost.

Table 2: Quantitative results for SID [6], Scale Map [36] and our method. The comparisons are made on different bit-depth raw images and various low-light illuminance levels. VGG loss is measured by comparing high-level image features extracted by pre-trained VGG16 network [37].

	Lux	Input PSNR/SSIM/VGG	SID PSNR/SSIM/VGG	Scale Map PSNR/SSIM/VGG	Ours PSNR/SSIM/VGG
8 bits	0.05	–	–	–	–
	0.1	–	–	–	–
	0.2	8.9/0.05/112.4	18.2/0.56/12.0	18.6/0.55/12.4	**22.8/0.63/9.3**
	0.4	10.9/0.08/67.3	20.4/0.60/10.9	21.7/0.61/10.0	**24.4/0.70/7.6**
12 bits	0.05	7.7/0.03/149.8	16.9/0.53/13.8	17.7/0.56/14.4	**20.9/0.63/11.1**
	0.1	9.9/0.06/85.1	19.2/0.58/10.7	20.6/0.59/11.2	**23.8/0.67/9.0**
	0.2	11.5/0.09/60.6	21.8/0.61/9.5	23.0/0.64/9.8	**25.6/0.71/7.9**
	0.4	13.7/0.17/35.7	24.5/0.65/8.2	25.6/0.68/7.6	**28.2/0.74/6.6**

the odd-number frames are the red-flash images. The outputs are noise-reduced frames without frame rate loss. Our proposed pipeline can effectively transfer the ambient-light color information and red-flash image features across neighboring frames, and produces compelling video reconstructions. Like no-flash photography, SOTA no-flash video denoising approaches cannot recover sufficient image details when the camera sensor captures poor signals.

The frequency at which flashes are steady to the human eye is known as the flicker fusion threshold. The threshold is determined by a few factors, such as the wavelength of light, luminance, and degree of dark adaptation [23]. The red-flash video could be recorded at frame rates ranging from 40 fps to 80 fps, which follows the change in ambient brightness and will not cause any discomfort flickering to the human visual system. Our videography pipeline could achieve 84 fps on 1024×768 video frames using Nvidia Tesla V100 GPU. With the employment of the temporal enhanced network, an averaged 15% reduction in terms of temporal warping error [22] is achieved. Temporal flickering artifacts are mitigated without loss of image sharpness.

Limitations. Intermediate frame synthesis relies on an accurate estimation of the motion fields. It is known that the computation of the optical flow is prone to error in the presence of large motions, occlusion and dramatic illumination change. The reconstructed video frames tend to produce color artifacts on fast moving objects. Enhanced temporal consistency or an integrated video fusion strategy could be studied in the future.

We also find that the proposed network shows a slightly degraded recovery for the color information in the blue channel, compared to that in the red and green channels. This is due to the fact that the camera sensor has a relatively lower spectral response in the blue channel. Moreover, auto white balance in low-light conditions is challenging as the sensor receives highly corrupted signals, and this is not considered by the proposed network. Enhancing the color restoration and integrating auto white balance into the network could be an interesting avenue for future work.

6. Conclusions

We have presented a novel deep-red flash technique for low-light imaging. This new flash method overcomes a number of limitations that have arisen from alternative flash-based approaches. We propose a modulation strategy to train a network to exploit the high-frequency features in the guide frames when performing image fusion. We have conducted the experiments on a variety of scenes and both the photography and videography results reveal remarkable performance under low-light conditions.

The utilization of a flash is an easy way to boost the ambient light levels and capture interesting scenes in extra dim conditions when reaching the limit of camera sensor sensitivity. In reality, no-flash and flash methods complement different low-light scenarios, like smartphones are equipped with Night mode and white flashes. By being less distracting, socially friendly, and easy to assemble, our proposed deep-red flash has the potential to change low-light imaging on smartphones.

References

[1] Amit Agrawal, Ramesh Raskar, Shree K Nayar, and Yuanzhen Li. Removing photography artifacts using gradient projection and flash-exposure sampling. *ACM Transactions on Graphics (TOG)*, 2005.

[2] Mathew Alpern. Effect of a bright light flash on dark adaptation of human rods. *Nature*, 230(5293), 1971.

[3] Tariq M Aslam, David Haider, and Ian J Murray. Principles of disability glare measurement: an ophthalmological perspective. *Acta Ophthalmologica Scandinavica*, 85(4), 2007.

[4] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, 2005.

[5] Chen Chen, Qifeng Chen, Minh Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, 2019.

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018.

[7] Gyeongmin Choe, Srinivasa G Narasimhan, and In So Kweon. Simultaneous estimation of near ir brdf and fine-scale surface geometry. In *CVPR*, 2016.

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8), 2007.

[9] Van den Berg. On the relation between glare and straylight. *Documenta Ophthalmologica*, 78(3-4), 1991.

[10] David E. Jacobs, Orazio Gallo, Emily A. Cooper, Kari Pulli, and Marc Levoy. Simulating the visual experience of very bright and very dark scenes. *ACM Transactions on Graphics (TOG)*, 2015.

[11] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM transactions on graphics (TOG)*, 2004.

[12] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12), 2006.

[13] Marjukka Eloholma and Liisa Halonen. New model for mesopic photometry and its application to road lighting. *Leukos*, 2006.

[14] Kathryn M Emmer, Kathryn LG Russart, II Walker, H William, Randy J Nelson, and A Courtney DeVries. Effects of light at night on laboratory animals and research outcomes. *Behavioral neuroscience*, 2018.

[15] James Ferwerda. Fundamentals of spatial vision. *Applications of visual perception in computer graphics*, 140, 1998.

[16] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 2016.

[17] Robert Francis Hess, RF Hess, Lindsay Theodore Sharpe, and K Nordby. *Night vision: Basic, clinical and applied aspects*. Cambridge University Press, 1990.

[18] Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. In *NIPS*, 2009.

[19] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*, 2019.

[20] Almut Kelber. Vision: Rods see in bright light. *Current Biology*, 28(8), 2018.

[21] Dilip Krishnan and Rob Fergus. Dark flash photography. *ACM Transactions on Graphics (TOG)*, 2009.

[22] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018.

[23] Carney Landis. Determinants of the critical flicker-fusion threshold. *Physiological Reviews*, 34(2):259–286, 1954.

[24] Muxingzi Li, Peihan Tu, and Wolfgang Heidrich. Robust joint image reconstruction from color and monochrome cameras. In *BMVC*, 2019.

[25] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[26] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *ACM Transactions on Graphics (TOG)*, 2019.

[27] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Transactions on Graphics (TOG)*, 2014.

[28] Yue M Lu, Clément Fredembach, Martin Vetterli, and Sabine Süsstrunk. Designing color filter arrays for the joint capture of visible and near-infrared images. In *ICIP*, 2009.

[29] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, 2018.

[30] Rafał Mantiuk, Allan G Rempel, and Wolfgang Heidrich. Display considerations for night and low-illumination viewing. In *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, 2009.

[31] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[32] Mike O'Connor. Tested: Huawei p30 pro. *Australian Photography*, 2019.

[33] 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)*, 2004.

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

[35] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4), 1992.

[36] Xiaoyong Shen, Qiong Yan, Li Xu, Lizhuang Ma, and Jiaya Jia. Multispectral joint image restoration via optimizing a scale map. *IEEE transactions on pattern analysis and machine intelligence*, 2015.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[38] Andrew Stockman and Lindsay T Sharpe. The spectral sensitivities of the middle-and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision research*, 40(13), 2000.

[39] David Strong and Tony Chan. Edge-preserving and scale-dependent properties of total variation regularization. *Inverse problems*, 19(6), 2003.

[40] Shunichi Sukegawa, Taku Umebayashi, Tsutomu Nakajima, Hiroshi Kawanobe, Ken Koseki, Isao Hirota, Tsutomu Haruta, Masanori Kasai, Koji Fukumoto, Toshifumi Wakano, et al. A 1/4-inch 8mpixel back-illuminated stacked cmos image sensor. In *International Solid-State Circuits Conference Digest of Technical Papers*, 2013.

[41] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[42] Huixuan Tang, Xiaopeng Zhang, Shaojie Zhuo, Feng Chen, Kiriakos N Kutulakos, and Liang Shen. High resolution photography with an rgb-infrared camera. In *ICCP*, 2015.

[43] Jian Wang, Tianfan Xue, Jonathan T Barron, and Jiawen Chen. Stereoscopic dark flash for low-light photography. In *ICCP*, 2019.

[44] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *NIPS*, 2012.