Computational Flash Photography through Intrinsics - Supplementary Material

Extended Flash and Ambient Illuminations Dataset

Sepideh Sarajian Maralan Chris Careaga Yağız Aksoy

Simon Fraser University

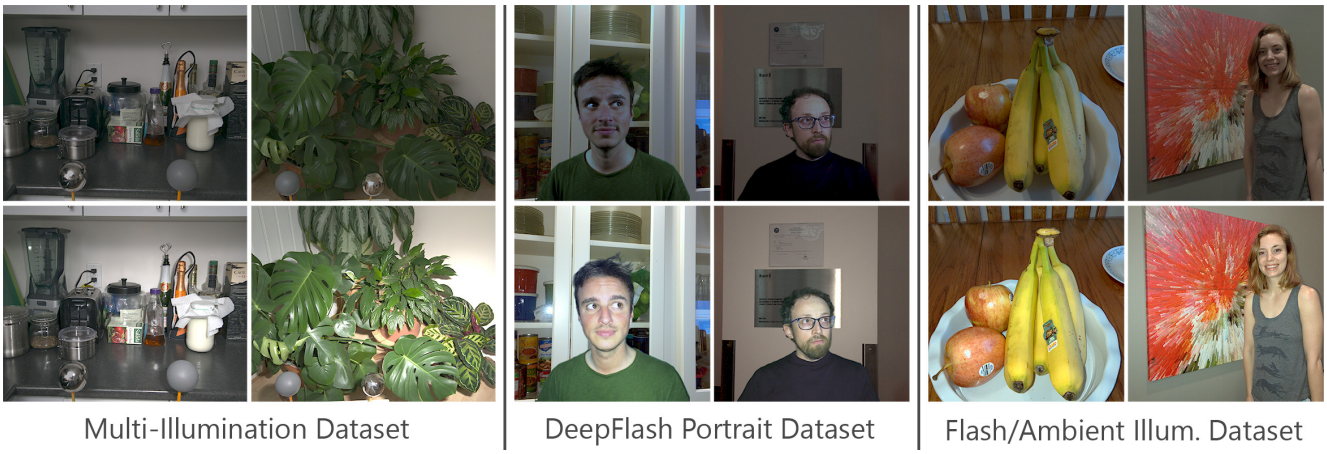

Figure 1. We combine and extend existing flash/no-flash datasets to create a diverse set of real-world flash/ambient pairs suitable for training deep networks. Our dataset is constructed from three existing datasets: The Multi-Illumination Dataset (MID), The Flash and Ambient Illuminations Dataset (FAID), and the Deep Flash Portrait Dataset (DPD). We propose a pipeline for compositing portraits from DPD onto backgrounds from FAID, and discuss multiple considerations for normalizing and augmenting flash/no-flash data.

We address the challenge of obtaining large-scale datasets that are appropriate for training deep networks to perform flash-related tasks in real-world scenarios. Although several small-scale datasets have been suggested in prior studies, none of them offer sufficient data to generalize to images captured in natural settings. To overcome this limitation, we combine and extend three existing datasets: The Multi-Illumination Dataset (MID), The Flash and Ambient Illuminations Dataset (FAID), and the Deep Flash Portrait Dataset (DPD). We propose a method for harmonizing and compositing portrait images from DPD onto plausible backgrounds from FAID. Additionally, we explain our procedure for performing brightness normalization to ensure consistent illumination intensity across datasets. Finally, we discuss strategies for data augmentation, including randomization of backgrounds and ambient color temperature.

1. Flash and Ambient Illuminations Dataset [2]

The Flash and Ambient Illuminations Dataset (FAID) [2] contains more than 2700 flash/no-flash pairs at the resolution of 1440×1080 as linear images taken with Apple iPhone 6s and 7. The photographs are organized into five classes: people, shelves, toys, plants, rooms, and objects. The image pairs are captured by casual mobile photographers recruited by Amazon's Mechanical Turk platform. Two images are captured in succession, with and without flash. The photographs are taken with a half- to one-second delay. Therefore there are small misalignments between the image pairs. Homography alignment is performed on each pair, and the successful ones are selected. Then, the no-flash (ambient) image is subtracted from the flash image in linear space to obtain the flash-only illumination.

The illuminations are in raw format, and their EXIF information is available in .mat files. The illuminations are first converted to XYZ color space using the camera calibration matrix. X approximates the red and green part of color, Y represents the brightness of the color, and Z corresponds to the blue and yellow part. The illuminations in XYZ are then converted to linear RGB space. The calibration illuminant white reference point used for this conversion is also available in the EXIF files. The illuminations are either in CIE standard illuminant A or CIE standard illuminant D65. The "A" illuminant corresponds to a temperature of 2856 K and represents typical, domestic, tungsten-filament lighting,

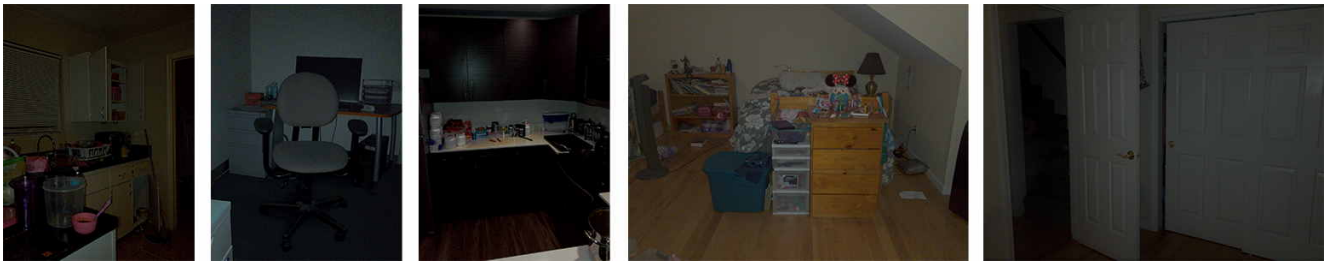

Figure 2. Photos with a dim flash illumination are not included in the database. This occurs when subjects are too far away from the camera

while the "D65" illuminant is correlated to color temperature of 6504K and simulates average daylight [8]. We map images in the "A" illuminant to "D65" illuminant by chromaticity adaptation. Chromaticity adaptation transforms a source color into a destination color by their reference white points. We utilized the Bradford method the newest and considerably the best method for chromaticity adaption.

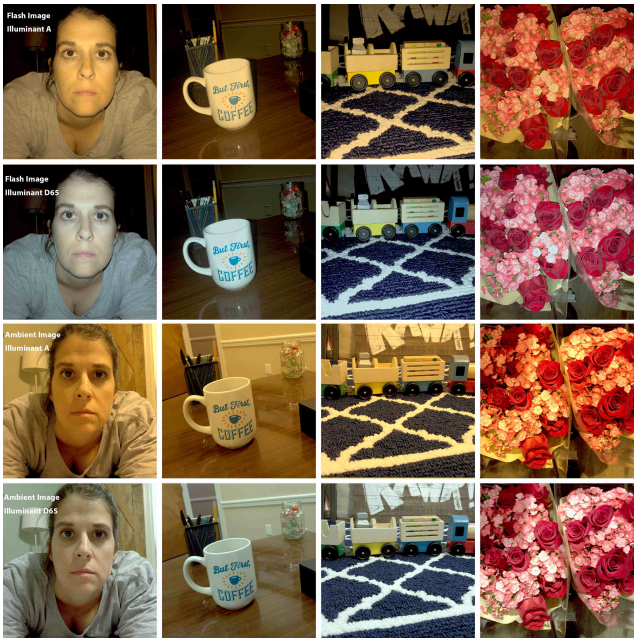

Figure 3. The images in illuminant A in FAID [2] are converted to illuminant D65 through chromaticity adaptation.

We eliminated nearly 700 images that had insufficient flash illumination or were too noisy. The 10% brightest flash illumination pixels are compared to a threshold. Those that do not meet the threshold are omitted. Several examples of removed images are depicted in Figure 2. Around 90% of each category is selected randomly for the training, and the other 10% is used for testing. There are multiple images from each person in the people category. Therefore the train and test set division is done manually in order to make sure each individual only appears in one of the sets.

2. Deep Flash Portrait Dataset [4]

DeepFlash (DPD) [4] presents a dataset containing 495 photograph pairs of 101 individuals. The flash/no-flash images are captured at a resolution of 3120×4160 by a Nexus 6 camera in front of a green screen. Four Lupoled 560 lamps are used as ambient illumination. For each pair, two images are captured, one with the lamps, and another with only flash from the camera. Here the problem of misalignment also occurs, so an affine alignment is performed for each pair. The final images are cropped to 512×512. Sixty-five pairs of images lacked the consent of the subjects and were removed from the dataset. Additionally, we eliminated 184 pairs due to misalignment caused by alterations in the pose or facial expression. Ninety percent of the images are used for training, while the remaining ten percent are kept for the test set. Again, we manually select them to ensure that no individual appears in both sets. Many of the flash photographs have the red-eye effect. This effect occurs in photographs of people and animals whose eyes appear red because of flash reflection when the flash is very close to the camera lens in dark ambient light. We removed these red eyes manually by using Adobe Photoshop. Our goal is to generate and decompose flash illumination in everyday photographs of real-world environments. To achieve this, we substitute the green screen background with indoor images. We utilize 135 image pairs from the room category of the FAID dataset as backgrounds. We ensure that the chosen backgrounds look realistic when composited. The pipeline for creating the new compositions is as follows:

Step 1: Segment foreground and background We use human portrait segmentation [3] that creates a binary map of a given portrait image. The method uses an encoder-decoder network pre-trained on a dataset of 1597 portrait images [10] cropped to 512×512 with their corresponding binary maps. This model consists of an encoder based on MobileNetV2 [9] and a decoder that utilizes transposed convolutions and upsampling layers.

Step 2: Generate trimaps of the foreground The binary maps from the last step are used to create trimaps of the por-

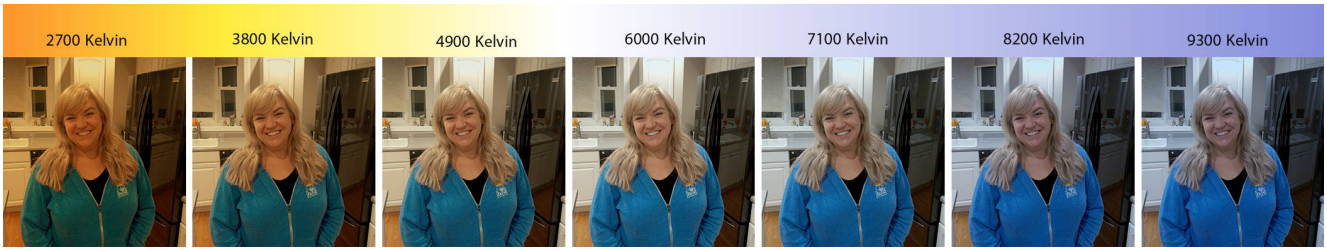

Figure 4. Chromatic adjustment allows for a shift in the white-balanced image's temperature. The temperature goes from warm temperature of 2700 Kelvin to cold temperature of 9300 Kelvin.

traits. Trimaps are three-level maps that segment the image into three regions: definite foreground, definite background and an unknown region. The trimaps are generated and then refined manually. An example trimap is shown in Figure 7.

Step 3: Estimate foreground matte In order to create realistic composites, we utilize alpha matting. With a trimap, the image matting problem is simplified to estimating the alpha values for pixels in unknown regions based on the known foreground and background pixels. We utilize Information-flow Matting [1] to compute the alpha matte. This method uses a variety of affinity definitions so that the unknown regions will receive information from the known regions. Information-flow matting is performed on the ambient illuminations, and then color estimation is done separately by [6] for both flash and ambient illuminations. An example matte is shown in Figure 7.

Step 4: Match white-balance between FAID and DPD To create realistic images, the illumination color of the foreground portrait must match that of the background scene. First, we match the flash colors in the foreground and background illumination. We take advantage of the fact that the flash illumination in a dataset remains constant due to the consistent camera setup. We observe that the FAID flash color is a cooler color than the portrait dataset flash color. We manually estimate a kernel $R \in \mathbb{R}^3$ of the relation of foreground flash to background flash in the three channels of RGB. We multiply this ratio by the flash portrait images:

$$\hat{I}_{f,fg} = R \times I_{f,fg} \tag{1}$$

Where $I_{f,fg}$, and $\hat{I}_{f,fg}$ denote the original and adjusted foreground flash illumination. The result of this procedure is shown in Figure 5.

Step 5: Estimate and match ambient illumination color Unlike the flash illumination, the ambient illumination color is not consistent across different scenes. We can not calculate the ambient illuminations directly but we can calculate the ratio of colors in flash/no-flash backgrounds. This ratio

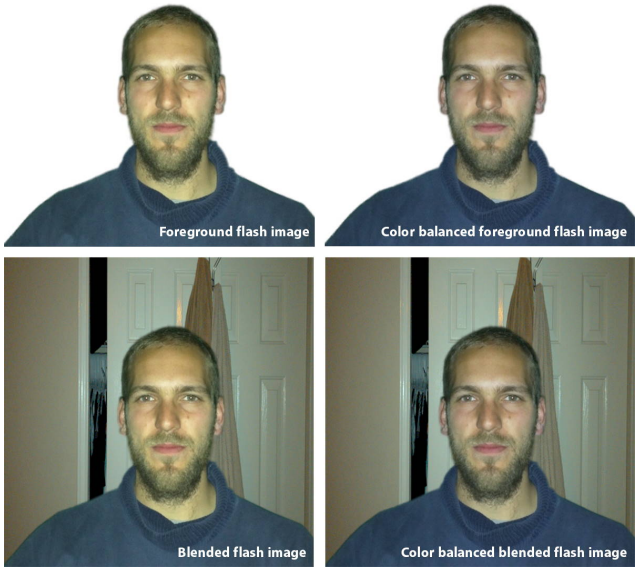

Figure 5. The flash color of the people in the DeepFlash Portrait dataset [4] is matched with the flash color of background images.

should be the same for both the foreground and background pair. Thus, we have to change the color of the ambient illumination foreground to match that of the background. We assume that the scene is Lambertian and the colorful shading S can be decomposed into a shading color and a grayscale shading map. The image can be represented as:

$$I = A \times \lambda \times C \tag{2}$$

Where A denotes the albedo, λ denotes the grayscale shading and C is the color of illumination. From this equation, we can observe that the ratio of flash/no-flash pairs in foreground and background will be 3.

$$\frac{I_f}{I_a} = \frac{\lambda_f \times C_f}{\lambda_a \times C_a} \tag{3}$$

Where I_f, λ_f and C_f correspond to the flash image and I_a, λ_a and C_a correspond to the ambient image. This ratio

has the shading of the scene in it. In order to estimate the ratio of colors and cancel out the grayscale shading, we normalize the images with respect to their grayscale intensities.

$$I_{f,norm} = I_f / I_{f,gs}$$
$$I_{a,norm} = I_a / I_{a,gs} \quad (4)$$

Where the subscript $norm$ denotes the normalized image with respect to the grayscale image and the subscript gs denotes to the grayscale image. This normalized ratio approximately gives us the ratio of flash to ambient.

$$R_{norm} = \frac{I_{f,norm}}{I_{a,norm}} = \frac{C_f}{C_a} \quad (5)$$

This ratio is estimated for each color channel and the foreground ambient image is multiplied by the median of each channel to get the same color as the flash foreground image.

$$I_{a,fg,wb} = I_{a,fg} \times med(R_{fg,norm}) \quad (6)$$

Where $I_{a,fg,wb}$ denotes the ambient foreground white balanced image, $I_{a,fg}$ is the ambient foreground image and $med(R_{fg,norm})$ is the median of normalized ratio. Because we have the same color of flash for the foreground and background, we only have to multiply the adjusted ambient image in the foreground with the ratio of median illuminations in the background.

$$I_{a,fg,cb} = I_{a,fg,wb} \times \frac{med(I_{a,bg})}{med(I_{f,bg})} \quad (7)$$

Where $I_{a,fg,cb}$ denotes the final color-balanced ambient foreground images, $I_{a,fg,wb}$ is the ambient foreground white balanced image and $\frac{med(I_{a,bg})}{med(I_{f,bg})}$ is the ratio of ambient median to flash median in background. This gives us the final version of the ambient portrait image to be composited with the background image. The blended image with and without the color balancing are shown in Figure 6.

Step 6: Composite the background and foreground To augment our dataset, each portrait is resized and composited in the center of twenty different background images. The dataset creation and processing is shown in Figure 7.

3. The Multi-Illumination Dataset [7]

The Multi-Illumination Dataset (MID) [7] contains 1016 indoor scenes taken under 25 fixed lighting directions. The photographs are taken with a DSLR Sony α6500 camera and a Sony HVL-F60M flash-light. Each scene is taken in an indoor environment containing different materials and objects. During the capturing process, the flash is rotated in 25 different directions. One of the directions is direct flash illumination, in which the light beam intersects the field of view of the camera. Twenty of the remaining lighting directions were selected as indirect ambient illuminations, allowing for ambient illumination augmentation. A selection of the ambient illuminations with the flash illumination for a given scene is shown in Figure 9.

Figure 6. The ambient color of the foregrounds from the DPD [4] is matched with the ambient color of backgrounds from FAID [2] by matching the ratio between the ambient and flash images.

4. Brightness Normalization

The photographs in the datasets contain a variety of ambient illumination and flash intensities. This makes it more challenging to estimate the actual quantity of flash shading. Therefore, we ensure that ambient and flash illumination strengths are consistent across all images to create a more uniform dataset. We utilize two methods for normalizing illumination brightness across our dataset. The first technique makes use of the brightness value (Bv) [5] data found in the EXIF files of the photographs. The scene luminance increases as the Bv increases. Brightness Value is defined in the Additive System of Photographic Exposure (APEX), which provides several factors for expressing the exposure in photographs. The brightness value is defined as follows

$$B_v = \log_2(\frac{A^2}{TNK}) \quad (8)$$

Where A is the f-number, T is the exposure time, K is the reflected-light meter calibration constant and N is approximately 0.3 which is the speed scaling constant. By analyzing the information contained in the EXIF files of the FAID

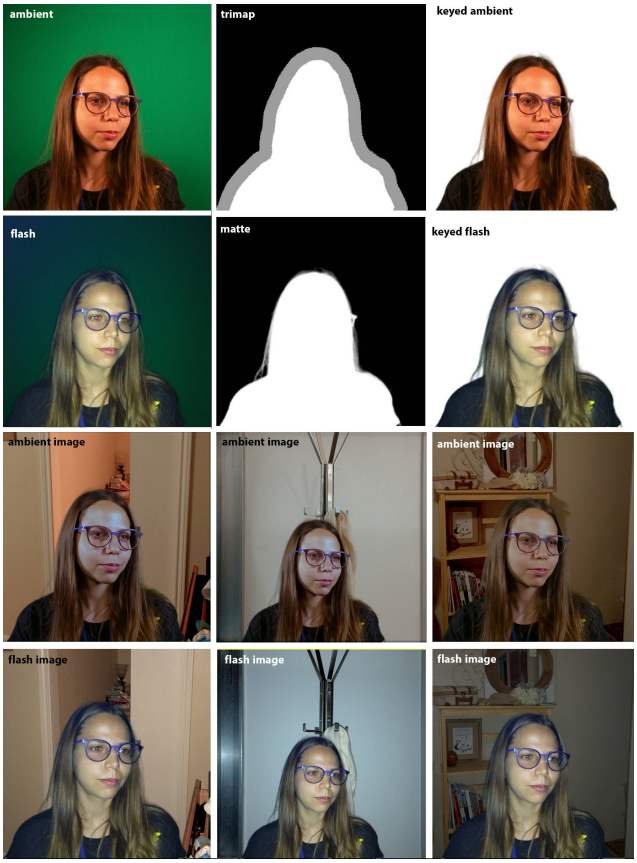

Figure 7. The DeepFlash Portrait Dataset [4] provides photographs taken in front of a green screen. We extract the foreground by generating a trimap and applying standard matting techniques. These extracted foregrounds are color balanced with respect to the background from FAID and composited

dataset, including the shutter speed, brightness value, exposure, and f-number, we observe that brightness value is closely related to scene brightness. This value is a good indicator of the scene's brightness, especially when the camera has utilized it without adjustment to compute aperture and ISO. In some instances, the brightness value may not apply to the entire photograph. We found that we could rely on brightness values for photographs taken with flash illumination in the FAID dataset. We required a flash illumination that was powerful enough to clearly demonstrate the characteristics of the flash illumination. We chose a brightness value that resembles a well-lit flash photograph, which is slightly above the FAID average and normalized all of the flash illuminations by this value. In the ambient illuminations, we decided to calculate the brightness of images as the numbers in the EXIF files were not reliable based on our observations. Therefore, we converted the ambient images to CIELAB (Lab) color space to calculate and normalize the brightness of ambient illumination. The reference white which we used to calculate the Lab color space is CIE Standard illuminant D65. We convert the ambient illumination to Lab and use the L channel to calculate the average brightness of the image without accounting for the 10% of pixels that are the brightest or darkest. This results in the image's overall brightness, without extremely bright or dark outliers. For each image, we define the scale as the ratio of the average brightness of all images to the given ambient image. This scale is multiplied by the L channel to produce the normalized image. Figure 8 shows ambient images being normalized by the L channel.

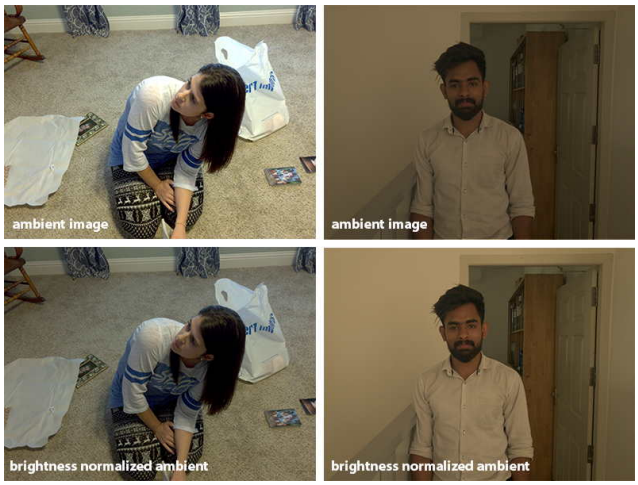

Figure 8. The brightness of ambient illumination is normalised relative to the L channel in CIELAB, resulting in images with approximately the same level of brightness throughout the dataset.

5. White Balance Ambient Illumination

We can assume that the flash light is white, while the ambient light can be any temperature. In order to further augment our dataset, we adjust the ambient light's temperature. We define the flash and ambient illuminations as:

$$\begin{aligned} I_F &= A \cdot S_F, \\ I_A &= A \cdot (c \cdot S_A), \end{aligned} \tag{9}$$

Where S_F is the gray-scale shading, the ambient shading is $S_A \cdot c$, S_A is the gray-scale ambient shading and $c = [r, g, b]$ is the color of ambient shading. We remove this color shift c in the ambient image by white balancing and then adjust the ambient illumination to be colder or warmer with chromatic adaptation. Since the flash illumination is white, we use it to white balance the ambient pair. This procedure is similar to the color balancing done in Section 2. We normalize the flash and ambient illuminations with respect to the brightness and compute the ratio between them. By multiplying the ambient to the median of this ratio, we approximate the white-balanced ambient image which is illustrated for a few scenes in Figure 10.

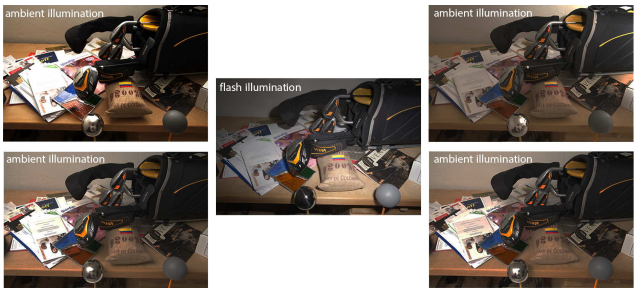

Figure 9. The Multi-Illumination Dataset features a variety flash-illuminated captures for the same scene. Each scene has one capture with direct flash illumination. The other captures utilize a bounce-flash and can therefore be used as indirect ambient illuminations. This provides 20 different ambient illuminations to be used as augmentation

6. Dataset Augmentation

The usual augmentation methods such as random cropping and rotating the image are not meaningful for our task. The flash light is correlated to the scene layout and we do not want to change the direction and strength of the flash illumination. For training purposes, the image is converted to linear RGB space. The display devices use the standard RGB (sRGB) color space however the non-linear gamma correction in sRGB is not suitable for mathematical operations. Additionally, the intrinsic image formulation is defined in the linear RGB space. We also alternate the temperature of ambient illuminations to be warmer or colder temperature in the range of indoor standard color temperatures to further augment our dataset. The chromaticity coordinates of different temperatures are estimated, and the ratio of the source white point and the new white point is calculated and multiplied by each pixel. We estimated the color distribution for the real ambient images in the FAID [2] and observed that the distribution of real-world ambient light color temperatures can be approximated by a Gaussian distribution centered on the warmer side of the range. Some different temperatures are shown in Figure 4. There are 1844 unique pairs within FAID, 985 unique scenes within MID, and 221 unique portrait pairs within DPD. In order to balance the three sub-datasets during training, MID and DPD require more augmentation than the FAID. For DPD we use 20 different backgrounds for the same foreground and in MID we use 20 different ambient illuminations for one flash illumination. To ensure that the dataset is presented to the network for training in a balanced manner, we pick more augmentations from the smaller sub-datasets. We randomly select two different flash-ambient pairs out of the available 20 pairs from MID and four random FAID backgrounds from the available 20 for each portrait in DPD.

Figure 10. The ambient illumination can have various colors while flash illumination is always white. We white balance the ambient illumination with respect to the flash illumination.

References

[1] Yağız Aksoy, Tunç Ozan Aydın, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *Proc. CVPR*, 2017. 3

[2] Yağız Aksoy, Changil Kim, Petr Kellnhofer, Sylvain Paris, Mohamed Elgharib, Marc Pollefeys, and Wojciech Matusik. A dataset of flash and ambient illumination pairs from the crowd. In *Proc. ECCV*, 2018. 1, 2, 4, 6

[3] Sait Aktürk. Human portrait segmentation. https://github.com/saitakturk/portrait_segmentation, 2019. 2

[4] Nicola Capece, Francesco Banterle, Paolo Cignoni, Fabio Ganovelli, Roberto Scopigno, and Ugo Erra. DeepFlash: Turning a flash selfie into a studio portrait. *Signal Processing: Image Communication*, 2019. 2, 3, 4, 5

[5] Douglas A. Kerr. Doug kerr's in-depth description of APEX. http://dougkerr.net/Pumpkin/. 4

[6] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008. 3

[7] Lukas Murmann, Michaël Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proc. ICCV*, 2019. 4

[8] Erik Reinhard, Erum Arif Khan, Ahmet Oğuz Akyüz, and Garrett M. Johnson. *Color Imaging: Fundamentals and Applications*. A. K. Peters, Ltd., 2008. 2

[9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. CVPR*, 2018. 2

[10] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. 2016. 2