A Dataset of Flash and Ambient Illumination Pairs from the Crowd

Yağız Aksoy[1,2](✉), Changil Kim[1], Petr Kellnhofer[1], Sylvain Paris[3], Mohamed Elgharib[4], Marc Pollefeys[2,5], and Wojciech Matusik[1]

[1] MIT CSAIL, Cambridge, MA, USA
[2] ETH Zürich, Zürich, Switzerland
ya@inf.ethz.ch
[3] Adobe Research, Cambridge, MA, USA
[4] QCRI, Doha, Qatar
[5] Microsoft, Redmond, WA, USA

Abstract. Illumination is a critical element of photography and is essential for many computer vision tasks. Flash light is unique in the sense that it is a widely available tool for easily manipulating the scene illumination. We present a dataset of thousands of ambient and flash illumination pairs to enable studying flash photography and other applications that can benefit from having separate illuminations. Different than the typical use of crowdsourcing in generating computer vision datasets, we make use of the crowd to directly take the photographs that make up our dataset. As a result, our dataset covers a wide variety of scenes captured by many casual photographers. We detail the advantages and challenges of our approach to crowdsourcing as well as the computational effort to generate completely separate flash illuminations from the ambient light in an uncontrolled setup. We present a brief examination of illumination decomposition, a challenging and underconstrained problem in flash photography, to demonstrate the use of our dataset in a data-driven approach.

Keywords: Flash photography · Dataset collection
Crowdsourcing · Illumination decomposition

1 Introduction

Crowdsourcing has been a driving force for computer vision datasets especially with the rise of data-driven approaches. The typical use of crowdsourcing in this field has been obtaining answers to high-level questions about photographs [7] or obtaining ground truth annotations [21] for simple tasks such as segmentation in a scalable and economical manner. However, commonplace strategies that rely on

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01240-3_39) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 644–660, 2018.
https://doi.org/10.1007/978-3-030-01240-3_39

user interaction do not apply to scenarios where complex physical processes are involved, such as flash/no-flash, short/long exposure, high/low dynamic range, or shallow/deep depth of field. With the wide availability and high quality of current mobile cameras, crowdsourcing has a larger potential that includes the collection of photographs directly. With the motivation of scalability and diversity, we tackle the challenge of crowdsourcing a computational photography dataset. We introduce a new approach where the crowd *captures* the images that make up the dataset directly, and illustrate our strategy on the flash/no-flash task.

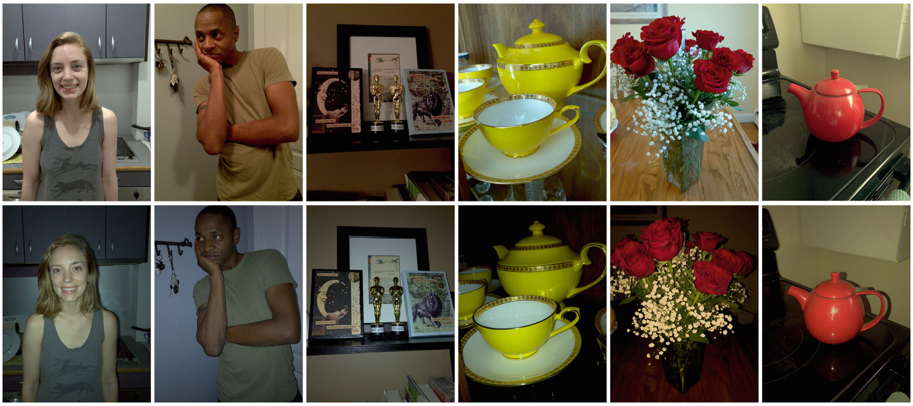

Fig. 1. We introduce a diverse dataset of thousands of photograph pairs with flash-only and ambient-only illuminations, collected via crowdsourcing.

Illumination is one of the most critical aspects of photography. The scene illumination determines the dominant aesthetic aspect of a photograph, as well as control of the visibility of and attention drawn to the objects in the scene. Furthermore, it is an important subject in visual computing and the availability of different illuminations of the same scene allows studying many different aspects of the photograph such as relighting, white balancing and illumination separation. However, capturing the same scenes under different illuminations is challenging, as the illumination is not easily controllable without photographic studio conditions. With its wide availability, flash is the easiest way for a casual photographer to alter the scene illumination. Thus, we focus on collecting a flash/no-flash dataset for demonstrating our crowdsourcing strategy. Similar to the Frankencamera [1], we use burst photography to capture several images in quick succession. This allows us to obtain pairs of nearly aligned images under different conditions; in our case, one is a flash photograph and the other is one only lit by the ambient light sources existing in the scene.

We present a dataset of thousands of images under ambient illumination and matching pairs that capture the same scene under *only* flash illumination. Figure 1 shows several examples of illumination pairs from our dataset. We have crowdsourced the collection of photograph pairs that result in a wide variety of scenes. This would not have been possible under fully controlled studios. We detail our approach, the challenges of crowdsourcing the photograph collection, and the processing pipeline to provide flash and ambient illumination pairs.

We envision that having two separate illuminations can aid in high-level tasks such as semantic segmentation or single-image depth estimation, as such high-level information is illumination-invariant. The dataset with image pairs identical up to illumination can also help with illumination analysis [13] or intrinsic image decomposition. Additionally, as one of the images in each pair is flash illumination, we hope that our dataset will encourage development of automatic image enhancement and lighting manipulation methods for mobile devices such as [8,27] or support computer vision applications similar to [31].

Illumination analysis has been an important problem in visual computing, e.g. the classical research problem of intrinsic image decomposition [4,5,20,23] where an image is decomposed into albedo and shading layers. Our dataset with two separate illuminations enables a new and related problem, single-image illumination decomposition. We present a brief study of illumination decomposition to see our dataset in action, where we train a network to decompose a flash photograph into corresponding ambient and flash illuminations and list the challenges that arise with this underconstrained problem. We show that although it is still an unsolved problem, a network trained with our dataset can generalize to substantially different images.

2 Related Work

Datasets of Separate Illuminations. Capturing the same scene under different illuminations is a challenging task that typically requires specialized setups and controlled environments. He and Lau [11] provide a dataset of 120 flash/no-flash photograph pairs captured with a DSLR camera and a tripod for the application of saliency detection. The dataset includes several objects, which define the salient regions in the image. Hui et al. [12] also provide a small set of 5 flash/no-flash photograph pairs. Murmann et al. [24] present 14 image sets captured using their specialized setup, each set consisting of 4 photographs taken under different flash directions. Krishnan and Fergus [18] also provide flash images taken with their hardware setup for 5 scenes. Our new dataset is significantly larger than the previously available examples, which allows its use for more data-demanding machine learning methods.

Another major difference is that we provide the ambient and flash illuminations separately, while flash photos in most of the previous work are indeed flash-dominant photos including ambient illumination as well. Weyrich et al. [33] provided a large dataset of facial images under different illuminations that were collected using a lighting dome in studio conditions. Separate illuminations have been provided for outdoor scenes with changing daylight and weather conditions [25]. Vonikakis et al. [32] captured 15 scenes under two separate illuminations in studio conditions. In contrast, we have collected our photographs via crowdsourcing in the wild, which allows for a larger dataset with a high variety of scenes.

Crowdsourced Datasets. Crowdsourcing has been an important tool for generating large-scale computer vision datasets. The crowd is typically utilized for tasks like labeling images [7], annotating images for interactive tasks [4,15,17], or drawing detailed object segmentations [21]. These datasets and many others use the crowd to conduct higher-level tasks for a given set of images. In our data collection setup, however, the crowd takes the photographs themselves. The main advantage of this approach is the wide variety of the input images that can be collected. We discuss the challenges that arise with direct data collection via crowdsourcing further in this paper.

Flash Photography. Previous work processing flash photographs mainly focuses on the joint processing of flash and no-flash pairs. Petschnigg et al. [27] and Eisemann and Durand [8] independently proposed the use of a flash photograph to denoise and improve the corresponding no-flash photograph taken in low-light conditions. Agrawal et al. [2] similarly use a flash/no-flash pair to remove the highlights from the flash photograph. In addition to image processing, flash/no-flash pairs have been used to improve image matting [31], automatic object segmentation [30], image deblurring [35], saliency detection [11], and stereo matching [34]. Recently, such pairs have been shown to be useful for white-balancing scenes with multiple ambient illuminations [12], and separation of such distinct light sources [13]. These works point to a wide set of use cases of flash/no-flash image pairs. By providing a large set of flash/ambient illumination pairs, the presented dataset enables further studies in these and other areas, as well as enabling data-driven approaches.

3 A Dataset of Flash and Ambient Illumination Pairs

We introduce a dataset of flash and ambient illumination pairs. Specifically, each pair consists of a photograph with only ambient illumination in a well-lit indoor environment, accompanied by the same scene illuminated *only* with the flash light.

The illuminations are provided as linear images at 1440 × 1080 resolution and 12-bit depth. Utilizing the superposition of light when there are multiple light sources in the scene, the pairs can be used to generate multiple versions of the same scene with varying lighting. For instance, to simulate a regular flash photograph taken in a dark environment, the typical use case of flash, a portion of the ambient illumination can be added to the flash illumination. Figure 3 shows several such variations. The white balance of the two illuminations can also be altered separately to create more alternatives.

Our dataset consists of more than 2700 illumination pairs of a wide variety of scenes. We have divided the dataset into 6 loosely defined categories, and several examples of each category are shown in Fig. 2. Roughly, 12% of the image pairs are in the category *People*, 15% in *Shelves* and *Toys* categories each, 10% in *Plants*, 30% in *Rooms* and the rest in the generic *Objects* category.

Previous work in flash photography presents flash/no-flash photograph pairs taken in dark environments [8,27], and hence the flash photograph contains a

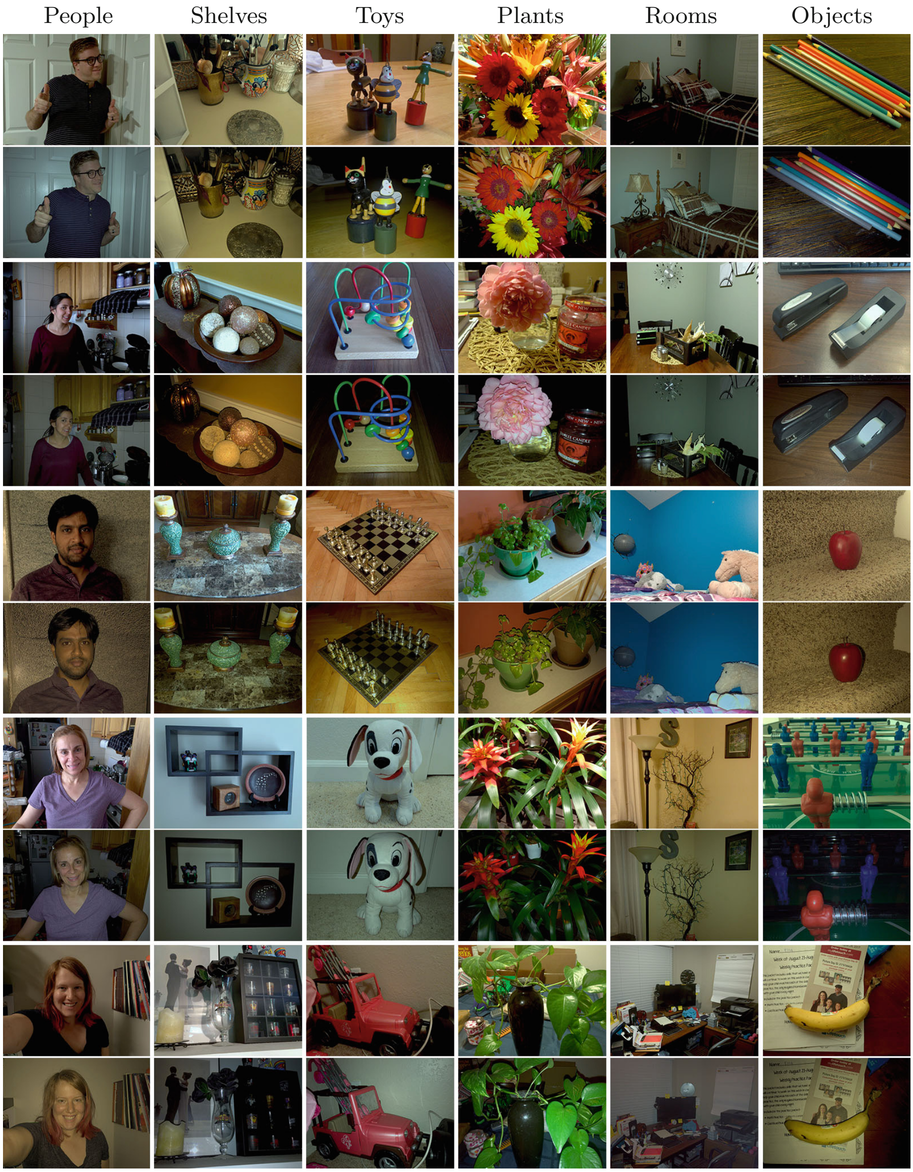

Fig. 2. We present a dataset of flash-only illumination with corresponding ambient illumination. The dataset consists of thousands photograph pairs collected via crowd-sourcing. The wide variety of images cover loosely-defined categories as listed at the top of the figure.

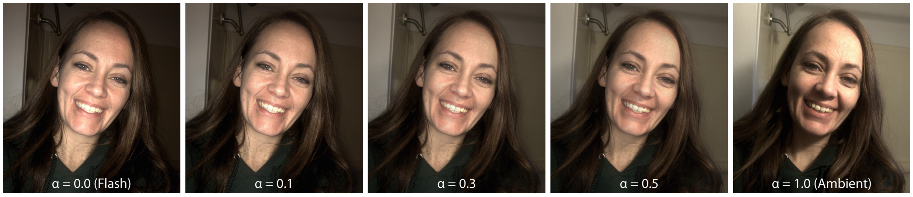

Fig. 3. The flash and ambient illumination pairs can be linearly combined with varying contributions (α) to simulate a flash photograph taken in darker environments.

portion of the ambient light. One important advantage of our dataset is that the flash pair does not contain any ambient illumination, making the two illuminations completely separate.

Ideally, obtaining such separate illumination pairs requires a controlled setup where the ambient illumination can be turned on and off. However, such a controlled setting makes it very challenging to scale the dataset size and limits the variety of the photographs that can be captured. Instead of a controlled setup, we use a dedicated mobile application to capture flash/no-flash photograph pairs and then computationally generate the flash-only illumination. Our setup also enables crowdsourcing of the collection process which in turn increases the variety of the scenes that are included in our dataset. We detail our collection procedure in Sect. 4.

We will open our dataset to the public to facilitate further research. Previous literature in flash photography shows that our dataset can be utilized for studying white balance, enhancement of flash photographs, saliency and more. The availability of such a dataset enables studying these problems in a data-driven manner. In addition, the availability of separate illuminations can be utilized for studying illumination-invariance in a variety of scenarios, as well as opening up the study of new problems such as illumination decomposition. We provide a brief examination of illumination decomposition, i.e. estimating the flash illumination from a single flash photograph, as an example use case of our dataset in Sect. 5.

4 Dataset Collection

We compute flash-only illumination from a pair of photographs, one taken with flash and one without. Using the superposition of different illuminants as seen by the camera, the difference between the flash photograph and the no-flash one contains the information for the flash illuminant under certain conditions. First of all, the raw values from the camera are required to correctly estimate the flash light. The camera parameters such as exposure and white balance must also match for the two photographs. In addition, the photographs, especially the no-flash photograph, should not contain saturated pixels. Finally, the two photographs must be well-aligned.

In order to allow for crowdsourcing, we needed to devise an easy and uncontrolled capture setup. We achieved this with a dedicated mobile application that takes the photograph pair with a single click. The application saves the raw image files and allows the user to upload them to a server. We match the image camera parameters using the photo metadata before computing the flash illumination. Handheld capture of two consecutive photographs inevitably results in small misalignments. We computationally align the two photographs before subtracting the no-flash photograph from the flash one. We detail these procedures and our crowdsourcing framework in the rest of this section.

Illumination Conditions. Unlike the previous work on flash photography where the photographs were typically taken in dark environments, we would like to have sufficient ambient lighting in the environment to generate varying illumination conditions as shown in Fig. 3 as well as to enable a larger set of possible uses for our dataset. In addition, we would like to reliably estimate the flash illumination and hence the flash light should be bright enough relative to the ambient light. This prevents us from taking photographs under daylight, where the flash light is much weaker compared to the sunlight. Hence, we restrict the scenes we capture to be indoor environments with sufficient ambient illumination, and free of directly visible light sources to avoid saturation.

Alignment. The flash and no-flash photographs are sequentially captured by our mobile app with a half to one second delay between the two exposures. This results in a small misalignment that must be corrected. However, accurate and reliable alignment of two images with different illuminations is a challenging task, as the image features that alignment methods rely on can be quite different in the presence of the flash light. One particular challenge is the hard shadows cast by the flash light, which results in strong gradients in one of the images. Hence, we limit our alignment to be rigid and estimate a homography between the two images using two different methods. We then review the two alignments and select the successful one by visual inspection, or remove the pair from the dataset if both methods fail.

The first method we utilize is the dual inverse compositional alignment algorithm (DIC) [3] from the image alignment toolbox [9]. DIC estimates geometric and photometric transformations simultaneously and is known for its robustness against illumination changes. DIC is effective when the ambient illumination is strong but fails in the presence of hard shadows.

We complement DIC by generating a shadow-free gradient map for both photographs and using Lucas-Kanade image registration [22] between these representations. We observed that the edges of the shadows cast by the flash light appear colorless when the white balance matches the flash color. We take advantage of this fact to remove the shadow edges, as shown in Fig. 4. We generate RGB gradient images for the two photographs and convert them to the HSV color space. We then multiply the saturation and value channels in this representation, which effectively diminishes the colorless edges including those of

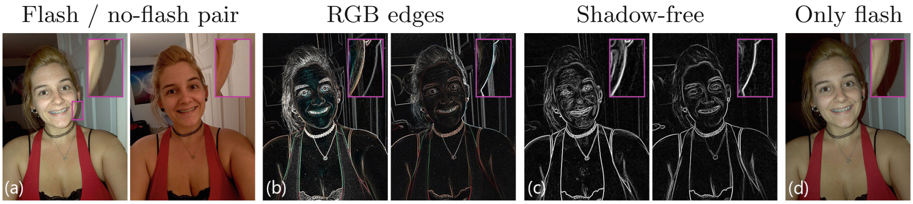

Fig. 4. For an unaligned flash and no-flash photograph pair (a), we compute the shadow-free edge representations (c) from the RGB image gradients (b). Inset shows the edge of the shadow in the flash image (b) disappearing in our representation (c). We estimate the alignment using the shadow-free representations and subtract the no-flash image from flash image after alignment to get the flash-only illumination (d).

flash shadows. Although some naturally colorless edges are lost, we are able to properly align the two photographs using this representation.

Collecting Photograph Pairs via Crowdsourcing. Diversifying images is important to generate a representative and generalizable dataset. This was a major advantage that drove us to crowdsource our dataset collection. We used Amazon's Mechanical Turk platform to recruit a large and diverse group of casual photographers. Using the crowdsourcing terminology, we will refer to the assignments as *human intelligence tasks (HITs)* and the photographers as *workers*. We list the major considerations we had to devise to enable the collection of our dataset in this part.

Framework. One essential component of our crowdsourcing effort was the mobile application we use to capture the dataset. The mobile application, developed for iOS devices, enables many casual photographers to participate in our collection effort from their home. Our HIT definition details our previously listed illumination expectations, provides a link and instructions for our application, and assigns a unique identifier per HIT. The worker is asked to install the application and enter the HIT identifier in the application. After taking the photographs, the worker uploads them to our server via the application. We then match the identifiers from Mechanical Turk and our servers to confirm the uploads. An example HIT definition is provided in the supplementary material.

Scene Categories. After initial trials, we observed that specifying scene categories guides the workers to find suitable scenes and increases the participation and the quality of the photographs. This lead us to define the first five categories shown in Fig. 2. These categories are loosely defined to allow the workers to easily find matching scenes around them, typically in their homes or workplaces. For most categories, we request the workers to take ten pictures per HIT. We give more details for the people category below. Not all workers strictly followed the category definitions, hence, we added the last category *objects* for the photographs that do not fit elsewhere.

People Category. The workers in Mechanical Turk typically work alone. This makes the HITs that require photographing other people much more challenging. In addition, the movement of subjects between the two photographs makes the alignment procedure even harder. Nevertheless, as one of the main use-cases of flash is portrait photography and facial image editing is an important topic in research, it is an important category to cover and, therefore, we defined the respective HITs more carefully.

A unique instruction is that instead of asking for ten photographs of different people, we ask the workers to take *five* photographs of the *same* person from different angles. This makes the worker complete each HIT much more quickly and makes the HIT more attractive. We ask the worker to instruct their subject to be still during capture, but subjects often fail to do so especially due to the flash light. This makes the percentage of portrait photographs we have to discard higher than average. Having five pictures of the same person increases the chance of using at least one of the poses and hence does not waste the worker's effort to recruit subjects. We also ask the worker to explain our dataset collection effort and get an explicit confirmation from the subject to participate in the study, as well as to avoid photographing minors.

Compensation and Noise. Not all the photographs we received were included in our dataset. Some common issues were photographs in very dark or very bright environments, non-static scenes and motion-blurred images. Other than such issues, there were cases of workers not uploading any photographs or uploading the same scenes many times. In such cases, we contacted the workers directly and usually got a positive response. We retained about a third of the images we received from the workers. We set the compensation as one U.S. dollar per HIT.

5 The Dataset in Action: Illumination Decomposition

Previous work in flash photography focuses on improving the low ambient lighting using a matching flash pair [8,27] or combining multiple flash images [24]. While these methods focus on estimating one high-quality image by combining multiple images, data-driven approaches allow tackling more difficult, but also more general problems.

To test our dataset, we present a data-driven approach to illumination decomposition as a baseline for future work. The goal of illumination decomposition is to estimate and separate the ambient and flash illuminations from a single flash photograph. We define the application scenario to include typical flash photographs that are taken in dark environments. We generate the input images by combining the illumination pairs in our dataset to simulate such dark environments as shown in Fig. 3. The ambient and flash illuminations then serve as the ground truth.

Illumination decomposition is an underconstrained problem, even when one of the illuminations is coming from the flash. We present several strategies we found to be useful in tackling this problem and present the challenges that arise.

We use a standard architecture to test our dataset and show that although the decomposition problem is far from being solved, a network trained with our dataset can be helpful in editing legacy photographs.

Network Structure and Implementation Details. We adopt the architecture proposed by Isola et al. [14] for generic image-to-image translation and experiment with several alternatives for the loss function and the estimation on ratio images as detailed in the rest of this section. The generator part of the network utilizes the U-net 256 network scheme [28] with eight convolution-deconvolution layer pairs and skip connections to pass full-resolution information to the next stage. It predicts the ratio image $\hat{r}_a$, which is used to reconstruct the estimated ambient illumination $\hat{I}_a$. We use Adam solver [16] with an initial learning rate of $2 \cdot 10^{-4}$ to train the main network and a lower rate of $2 \cdot 10^{-6}$ to train the discriminator. We decrease the learning rate by a factor of 10 every 30 epochs and we terminate the training after 150 epochs which takes approximately 2 h on our setup. The forward pass including all other processing is fast enough for interactive applications (Fig. 5).

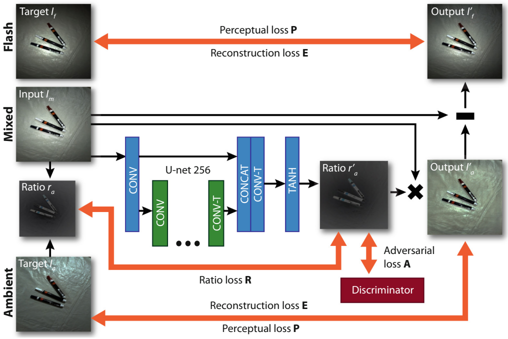

Fig. 5. The network structure with the losses shown in orange. (Color figure online)

Ratio Images. We observed that directly estimating the ambient or flash illuminations results in loss of high-frequency details and periodic noise structures. Trying to correct such artifacts via joint filtering such as the domain transform [10] either does not remove such artifacts, or oversmooths the image. Instead of a direct estimation of the illuminations, we chose to first estimate the ratio image an intermediate representation. Inspired by the use of *ratio images* in facial relighting literature [6,26], we define the output of our network as the ratio between the input image I_m and the ambient illumination I_a. Although the exact definition of how this ratio is computed does not have a substantial effect on the network, we define the ratio to be in range [0, 1] and compute the estimated ambient illumination accordingly:

$$r_a = \frac{2 \cdot (I_a + 1)}{3 \cdot (I_m + 1)} - \frac{1}{3} \qquad\qquad I_a = \frac{3 \cdot (r_a + I_m\, r_a) + I_m - 1}{2} \tag{1}$$

The artifacts mentioned above also appear in the ratio images, but when the domain transform is applied, we are able to remove them without the loss of

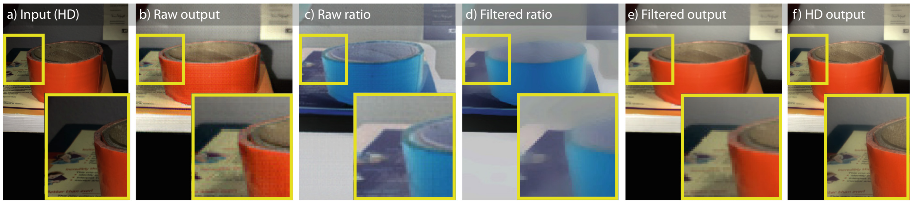

Fig. 6. High resolution input image (a) is downsampled and fed into the network. The ambient illumination image output may contain a residual noise (b) which would stand out in an otherwise smooth ratio image (c). We filter the ratio image while preserving its edges using the input image as a guide (d). The reconstructed image is free of noise (e). Same approach is used to upsample the output to the original high resolution (f).

high-frequency details in the output ambient illumination. Another advantage of using ratio images is the ease of upscaling. The network works on a downscaled version of the image and we upscale the ratio image before the domain filtering. This way, we are able to generate high-resolution outputs without the need of feeding the image at full resolution to the network. Figure 6 shows our workflow.

Dataset Augmentation. At each epoch, we generate the input images by randomly sampling α (Fig. 3) to determine how dark the ambient illumination is compared to the flash light. This makes the learning process more robust against different illumination conditions. We also randomly crop, rotate, flip and scale the images at each epoch before feeding them into the network.

Loss Functions. The original Pix2Pix [14] combines an adversarial loss $\mathbf{A}$ and an L_1 loss $\mathbf{R}$:

$$\mathbf{R} = \|\hat{r}_a - r_a\|_1, \tag{2}$$

which we refer to as the ratio loss, as the output of the network is the ratio image in our approach. In addition to these, we also define an L_2 loss that applies to both the estimated ambient illumination and the flash illumination $\hat{I}_f = I_m - \hat{I}_a$:

$$\mathbf{E} = \|\hat{I}_a - I_a\|_2^2 + \|\hat{I}_f - I_f\|_2^2. \tag{3}$$

We observed that, while the two terms of Eq. 3 are correlated, including the losses on both flash and ambient illuminations leads to better performance.

We have also tested a perceptual loss $\mathbf{P}$ proposed by Sajjadi et al. [29] that preserves the perceived image quality:

$$\mathbf{P} = \|\mathcal{P}(\hat{I}_a, I_a)\|_2^2 + \|\mathcal{P}(\hat{I}_f, I_f)\|_2^2 \tag{4}$$

where $\mathcal{P}$ extracts the features from a pre-trained network [19].

We combine a subset of these losses in our experiments with empirically determined weights $100 \cdot \mathbf{R} + 1000 \cdot \mathbf{E} + 1000 \cdot \mathbf{P} + \mathbf{A}$. We have tested the quantitative results for several combinations of these loss functions and the results are summarized in Table 1. We observed that the use of adversarial loss **A** leads to a strong high-frequency noise. Using the perceptual loss **P**, on the other hand, results in a color shift for some image regions. Removing loss **R** or **E** generally led to a result similar to the baseline **RE** as they are correlated metrics. However, we got best visual as well as quantitative results using both losses.

Table 1. Test errors for the estimated ambient illuminations

α	PSNR (dB)					SSIM				
	RE	R	E	REP	REA	RE	R	E	REP	REA
0.1	**15.220**	14.801	14.708	15.163	14.099	**0.583**	0.516	0.505	0.567	0.334
0.3	**18.046**	17.338	17.288	17.676	16.431	**0.720**	0.638	0.634	0.694	0.427
0.5	**20.261**	19.199	19.333	19.568	18.116	**0.804**	0.707	0.716	0.773	0.495

Qualitative Evaluation. Figure 7 shows several examples that demonstrate the strengths and shortcomings of the presented illumination decomposition method. For instance, the network estimates a uniform illumination for the ambient illumination, and a dark background and bright foreground for the flash illumination (green highlights in the figure). Even when the ambient illumination in the input image is very dim, such as in (1, 2, 5), the estimated ambient illumination is uniform and bright. The highlights from the flash light are typically well-detected by the network, as seen in examples (3, 4).

However, there are several limitations. The flash highlights may bleed into the estimated ambient illumination (2, 5) or the flash (1) and ambient (4) shadows cannot be reliably separated. After analyzing our results, we believe that a more dedicated approach to facial images would be useful. In some images, our network is better at identifying the highlights on the face but misses more subtle ambient lighting details (3). In others, it may fail to generate a satisfactory ambient light, especially if the environment is dark (5). With the wide variety of images, some unusual examples also arise. For example, our network works well in decomposition in (6) but gets confused around the mirror image of the flower.

These examples demonstrate the difficulty of illumination decomposition. We argue that its underconstrained nature underscores the need for data-driven approaches, as they can potentially learn the strong priors for the flash illumination from many examples to better constrain the problem.

Generalization. We present illumination decomposition examples in old photographs that were scanned from film to test if our dataset can train the network for a generalizable decomposition. We apply inverse gamma mapping to linearize the input images before feeding them into the network. In Fig. 8, from a single image, we recreate the photographs with varying ambient illumination we showed in Fig. 3 using the actual flash and ambient illuminations. This way,

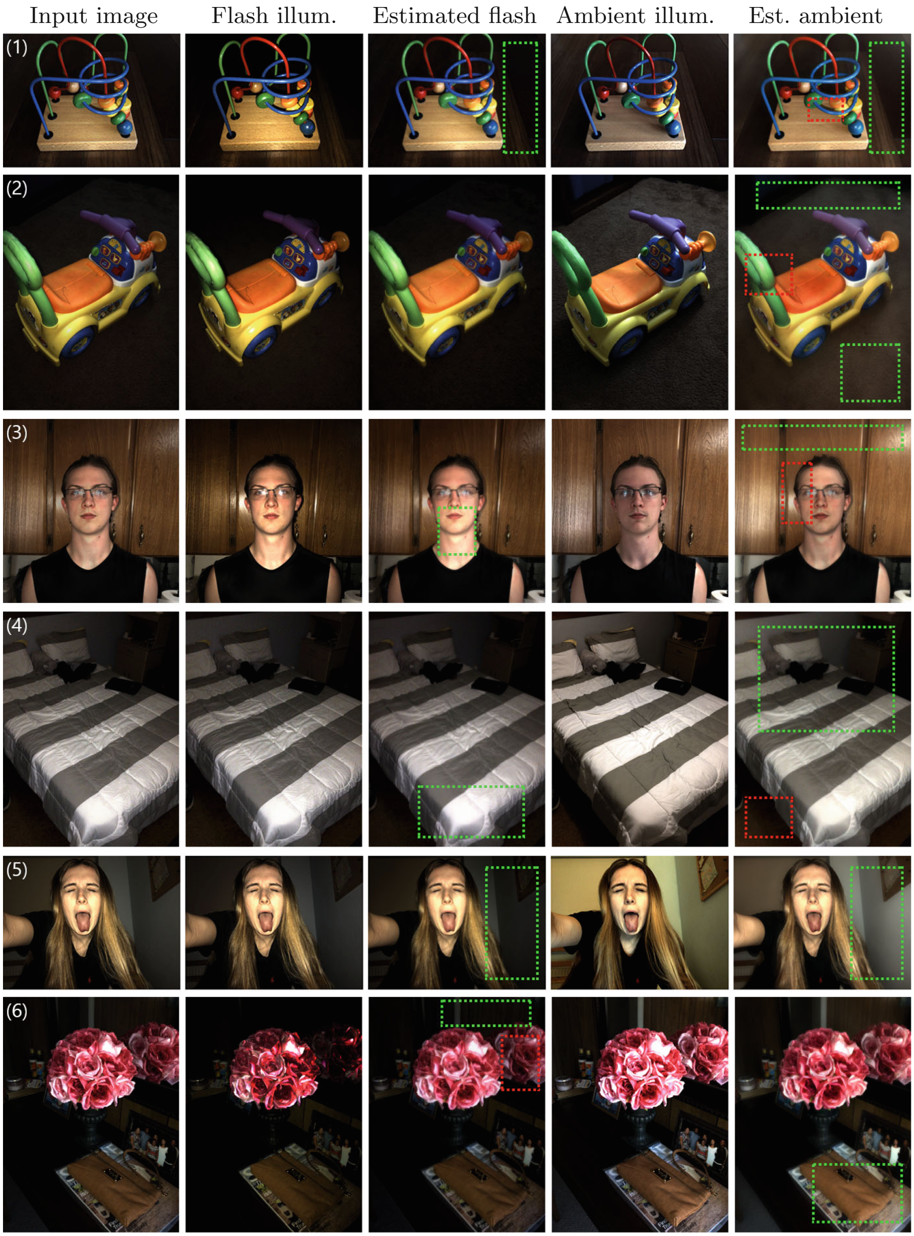

Fig. 7. Several illumination decomposition examples using our dataset. The highlighted areas, red squares demonstrating the limitations, are discussed further in the text. (Color figure online)

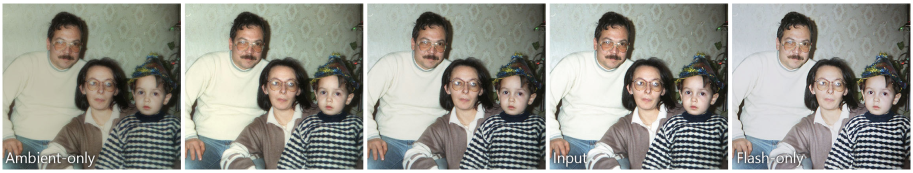

Fig. 8. The estimated ambient and flash illuminations can be used to generate a wide range of possible combination of two illuminations.

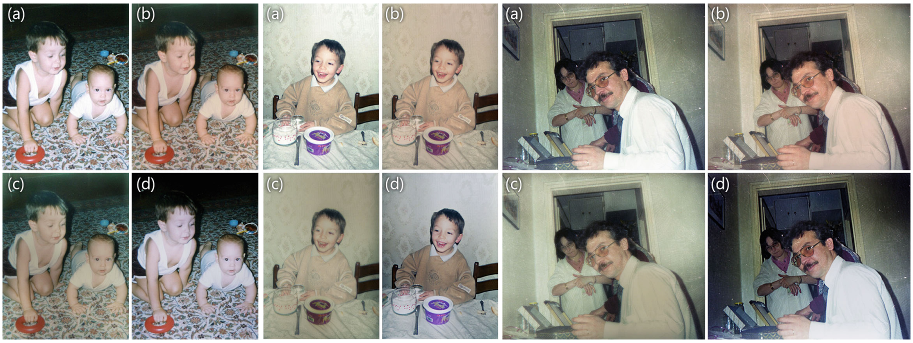

Fig. 9. Input images that were originally taken on film (a) are decomposed into ambient (c) and flash (d) illuminations. The decomposed illuminations can be edited separately and then combined for a more pleasant look (b).

using illumination decomposition, an artist can change the ambient illumination as desired. The decomposed illuminations can also be used to create more pleasant photographs by softening the flash light. Figure 9 shows such examples, where the decomposed flash illuminations are edited to match the color of the ambient illumination, and the ambient illuminations are made stronger to give the photographs a more natural look.

These examples demonstrate that, even with a modest size of several thousands of pairs, by allowing a wide range of augmentations such as varying ambient contribution to the image, our dataset can be used to train a network that can generalize to previously unseen images.

6 Conclusion

We presented a large-scale collection of crowdsourced flash and ambient illumination pairs using smartphone cameras. Our dataset is unique in that it provides complete separation of flash and ambient illuminations in its photo collection unlike previous datasets, and consists of a significantly larger number of photographs. We provide the details of our data collection pipeline, which leverages crowdsourcing and the increasing capabilities of current smartphones, that is designed to be used in unconstrained environments. We demonstrate the use of our dataset in the problem of single-image illumination decomposition and provide considerations for further research in this avenue.

Acknowledgements. We would like to thank Alexandre Kaspar for his support on crowdsourcing, James Minor and Valentin Deschaintre for their feedback on the text, and Michaël Gharbi for our discussions. Y. Aksoy was supported by QCRI-CSAIL Computer Science Research Program at MIT, and C. Kim was supported by Swiss National Science Foundation fellowship P2EZP2 168785.

References

1. Adams, A., et al.: The frankencamera: an experimental platform for computational photography. Commun. ACM **55**(11), 90–98 (2012)
2. Agrawal, A., Raskar, R., Nayar, S.K., Li, Y.: Removing photography artifacts using gradient projection and flash-exposure sampling. ACM Trans. Graph. **24**(3), 828–835 (2005)
3. Bartoli, A.: Groupwise geometric and photometric direct image registration. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2098–2108 (2008)
4. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. Graph. **33**(4) (2014)
5. Bonneel, N., Kovacs, B., Paris, S., Bala, K.: Intrinsic decompositions for image editing. Comput. Graph. Forum **36**(2), 593–609 (2017)
6. Chen, J., Su, G., He, J., Ben, S.: Face image relighting using locally constrained global optimization. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 44–57. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_4
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
8. Eisemann, E., Durand, F.: Flash photography enhancement via intrinsic relighting. ACM Trans. Graph. **23**(3), 673–678 (2004)
9. Evangelidis, G.: IAT: a Matlab toolbox for image alignment (2013). http://www.iatool.net
10. Gastal, E.S.L., Oliveira, M.M.: Domain transform for edge-aware image and video processing. ACM Trans. Graph. **30**(4), 69:1–69:12 (2011)
11. He, S., Lau, R.W.H.: Saliency detection with flash and no-flash image pairs. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 110–124. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_8
12. Hui, Z., Sankaranarayanan, A.C., Sunkavalli, K., Hadap, S.: White balance under mixed illumination using flash photography. In: International Conference on Computational Photography (ICCP) (2016)
13. Hui, Z., Sunkavalli, K., Hadap, S., Sankaranarayanan, A.C.: Illuminant spectra-based source separation using flash photography. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

15. Kaspar, A., Patterson, G., Kim, C., Aksoy, Y., Matusik, W., Elgharib, M.: Crowd-Guided ensembles: how can we choreograph crowd workers for video segmentation? In: ACM CHI Conference on Human Factors in Computing Systems (2018)
16. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2014)
17. Kovacs, B., Bell, S., Snavely, N., Bala, K.: Shading annotations in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
18. Krishnan, D., Fergus, R.: Dark flash photography. ACM Trans. Graph. **28**(3), 96:1–96:11 (2009)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems Conference (NIPS) (2012)
20. Lettry, L., Vanhoey, K., Van Gool, L.: DARN: a deep adversial residual network for intrinsic image decomposition. In: Winter Conference on Applications of Computer Vision (WACV) (2018)
21. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
22. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence (IJCAI) (1981)
23. Meka, A., Zollhöfer, M., Richardt, C., Theobalt, C.: Live intrinsic video. ACM Trans. Graph. **35**(4), 109:1–109:14 (2016)
24. Murmann, L., Davis, A., Kautz, J., Durand, F.: Computational bounce flash for indoor portraits. ACM Trans. Graph. **35**(6), 190:1–190:9 (2016)
25. Narasimhan, S.G., Wang, C., Nayar, S.K.: All the images of an outdoor scene. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 148–162. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47977-5_10
26. Peers, P., Tamura, N., Matusik, W., Debevec, P.: Post-production facial performance relighting using reflectance transfer. ACM Trans. Graph. **26**(3) (2007)
27. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. ACM Trans. Graph. **23**(3), 664–672 (2004)
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Sajjadi, M.S., Schölkopf, B., Hirsch, M.: EnhanceNet: single image super-resolution through automated texture synthesis. In: International Conference on Computer Vision (ICCV) (2017)
30. Sun, J., Sun, J., Kang, S.B., Xu, Z.B., Tang, X., Shum, H.Y.: Flash cut: foreground extraction with flash and no-flash image pairs. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
31. Sun, J., Li, Y., Kang, S.B., Shum, H.Y.: Flash matting. ACM Trans. Graph. **25**(3), 772–778 (2006)
32. Vonikakis, V., Chrysostomou, D., Kouskouridas, R., Gasteratos, A.: Improving the robustness in feature detection by local contrast enhancement. In: International Conference on Imaging Systems and Techniques (IST) (2012)
33. Weyrich, T., et al.: Analysis of human faces using a measurement-based skin reflectance model. ACM Trans. Graph. **25**(3), 1013–1024 (2006)

34. Zhou, C., Troccoli, A., Pulli, K.: Robust stereo with flash and no-flash image pairs. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
35. Zhuo, S., Guo, D., Sim, T.: Robust flash deblurring. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2010)