A Well-aligned Dataset for Learning Image Signal Processing on Smartphones from a High-end Camera

Yazhou Xing
HKUST

Changlin Li
HKUST

Xuaner Zhang
Adobe Inc.

Qifeng Chen
HKUST

Figure 1: Reconstructed RGB images using our proposed model. Compared with iPhone 6S ISP, our result can better utilize the recorded information in raw data and reconstruct visually appealing results even under backlit scenes.

KEYWORDS

Image signal processing, Image perceptual quality, conditional convolutional networks

ACM Reference Format:
Yazhou Xing, Changlin Li, Xuaner Zhang, and Qifeng Chen. 2022. A Well-aligned Dataset for Learning Image Signal Processing on Smartphones from a High-end Camera. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '22 Posters), August 07-11, 2022.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3532719.3543252

1 INTRODUCTION

Not every camera is equipped with an excellent image signal processing (ISP) pipeline that converts raw sensor data into color images. The main objective of an ISP is to produce a visually appealing image that is also faithful to the scene being captured. Conventional ISP is composed of a sequence of modules including white balance, demosaicking, denoising, tone mapping, and so on. It is labor-intensive and challenging to design an ISP pipeline with many independent modules, and thus the ISP on most mobile phones is sub-optimal, even for the highly-rated ones such as iPhone. In this paper, we present a novel learning-based model that replaces built-in ISP and synthesizes images that match the image quality from high-end professional cameras. Our approach does not rely on the sub-optimal built-in ISP at all but instead utilizes a fully convolutional network with content-aware conditional convolutions to act as ISP. To train the deep learning model, we collect a large-scale dataset with raw and RGB data pairs captured by two popular smartphones and one high-end camera. Our dataset complements the existing Raw-to-RGB ISP dataset [Ignatov et al. 2020] with with more types of smartphone images. Our model takes the raw sensor data from a smartphone as input and generates an RGB image that is optimized to reach the image quality coming from the high-end camera ISP. Experimental results show that our presented model produces perceptually better images than the popular smartphones do when using the same sensor data.

2 OUR ALIGNED DATASET

A big challenge of training a data-driven ISP model is the lack of raw sensor data and desired high-quality images as ground truth. We collect a dataset that contains raw sensor data captured by smartphones with small sensors, and RGB images as the target image from a high-end camera (Nikon Z6) with high-quality in-camera ISP. Image alignment is challenging when using data pairs captured from different devices for training. We find only global alignment like homography transformation can hardly achieve pixel-wise accuracy thus we integrate local alignment to achieve sub-pixel accuracy. First, we estimate a homography transformation using SIFT features. Because of different FOV and the constraint of homography transformation only handling co-planar scenes, this initial alignment is not pixel-wise accurate. We use it as initialization for

SIGGRAPH '22 Posters, August 07-11, 2022, Vancouver, BC, Canada

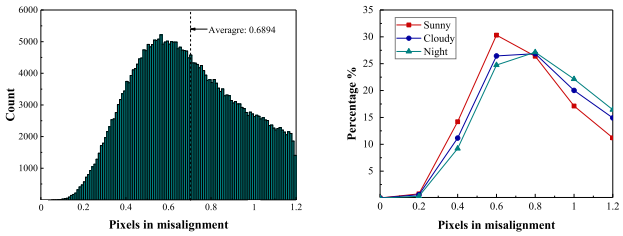
ACM ISBN 978-1-4503-9361-4/22/08.
https://doi.org/10.1145/3532719.3543252

SID [Chen et al. 2018] DPED [Ignatov et al. 2017] PAC [Su et al. 2019] iPhone 6S Ours

Figure 2: We compare our results with state-of-the-art image enhancement models as well as in-camera ISP.

iPhone 6S Nikon Z6 Camera Rig

Figure 3: Sample data triplet and the camera rig used to capture our dataset.



(a) Misalignment distribution (b) Dependence on illumination

Figure 4: Misalignment analysis. In our dataset, most patches have misalignment to 0.4 ~ 0.7 pixels. The same misalignment analysis on different illuminations is consistent with overall misalignment distribution, as seen in (b).

subsequent patch alignment steps. Second, we split Nikon RGB into non-overlapping patches and search for the best matching patch in smartphone RGB, using the normalized cross-correlation (NCC) index as the matching metric. For additional dataset filtering, we apply the PWC-Net to estimate the average pixel flow shift to filter out poorly-aligned patches. We reject patches with misalignment greater than 1.2 pixels. The remaining image patches contain an average pixel misalignment of 0.6 pixels. A detailed analysis is illustrated in Figure 4. In total, we obtain 333K image patch pairs from 1270 iPhone-Z6 pairs and 1154 Mi-Z6 pairs for training and testing.[1]

3 METHOD

Given a raw image X captured by a small sensor camera, our goal is to render a high-quality RGB image $\hat{Y}$. The high dynamic range of raw sensor images imposes a great challenge using a conventional CNN architecture that relies on spatially invariant convolutions, which are considered antithetical to localize edge discontinuities. Thus, directly applying standard convolution can cause apparent artifacts such as halos, which has been identified in previous non-learning-based image filtering methods [Guarnieri et al. 2011; Paris et al. 2011]. We propose a novel edge-aware conditional convolutional network architecture based on the kernel prediction method [Bako et al. 2017].

Given an n-pixel input image $X = (X_1, X_2, \ldots, X_n), X_i \in \mathbb{R}^c$, the output $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n), \hat{Y}_i \in \mathbb{R}^{c'}$ can be obtained by convolution operation with per-pixel kernel W:

$$\hat{Y}_i = \sum_{j \in \delta_i} W^i[i-j] X_j + b, \tag{1}$$

where W^i denotes the conditional convolution kernel at position i, δ_i is the neighbourhood window centered at position i. Note that we formulate images as one-dimension vectors for notation clarity. In general, W^i should be a function of input content:

$$W^i = f(X, \delta_i). \tag{2}$$

We leverage the approximation power of neural networks to estimate the function f. We predict these conditional kernels as the output of the network and then convolve the predicted kernels with input features.

4 EXPERIMENT RESULTS

We present preliminary results in Figures 1 and 2. Under good lighting conditions, our reconstructed results have more vivid colors and can better handle noise and preserve fine details without generating artifacts. Under backlit scenes, our method utilizes the dynamic range of raw sensor data effectively such that both over- and under-exposure regions can be rendered clearly.

REFERENCES

Steve Bako, Thijs Vogels, Brian McWilliams, Mark Meyer, Jan Novák, Alex Harvill, Pradeep Sen, Tony Derose, and Fabrice Rousselle. 2017. Kernel-predicting convolutional networks for denoising Monte Carlo renderings. *ACM Trans. Graph. (TOG)* 36, 4 (2017), 97–1.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to See in the Dark. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabriele Guarnieri, Stefano Marsi, and Giovanni Ramponi. 2011. High Dynamic Range Image Display With Halo and Clipping Prevention. *IEEE Trans. Image Processing (TIP)* 36, 4 (2011), 1–12.

Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. 2017. DSLR-Quality Photos on Mobile Devices with Deep Convolutional Networks. In *International Conference on Computer Vision (ICCV)*.

Andrey Ignatov, Luc Van Gool, and Radu Timofte. 2020. Replacing Mobile Camera ISP with a Single Deep Learning Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Sylvain Paris, Samuel W Hasinoff, and Jan Kautz. 2011. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Trans. Graph. (TOG)* 30, 4 (2011), 68.

Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. 2019. Pixel-Adaptive Convolutional Neural Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[1] Our dataset can be downloaded from here.