End-to-End Single-Frame Image Signal Processing for High Dynamic Range Scenes

Khanh Quoc Dinh and Kwang Pyo Choi
Samsung Research, Samsung Electronics
{kq.dinh, kp5.choi}@samsung.com

Abstract

This paper considers photography of high dynamic range scenes containing mixtures of shadows and highlights on mobile phones. Multi-frame merging constructs a high-quality image at the cost of capturing multiple frames of the same scene. Contrarily, end-to-end optimized image signal processing (E2EISP) produces an enhanced image from a single-frame Bayer array. This paper combines the merits of the two approaches by using labels of high-quality multi-frame merged images to train E2EISP with a novel neural network architecture composed of a multi-head mixture of brightness enhancement for accurately processing shadows/highlights and a multi-head mixture of image processing featured camera settings of white balance and color correction for a proper color generation. We also proposed a combination of supervised, unsupervised, and generative adversarial losses for brightness, edge, and detail enhancement. Experimental results show that the proposed single-frame ISP produces enhanced images and outperforms state-of-the-art methods.

1. Introduction

Photography on mobile phones has become essential for daily life, however, their limits on cost and size lead to multiple technical challenges including taking images of high dynamic range (HDR) scenes, such as outdoor landscapes on sunny days. The conventional image signal processing (ISP) from a single frame suffers from noisy shadows and saturated highlights if the image is taken with a long exposure. Contrarily, a short exposure causes much-dimmed shadows despite satisfying highlights. The common practice is to capture multiple frames of a scene, with either the same or different exposures, and then merge them into either an HDR or enhanced image [19, 12, 29]. As multiple frames convey more information about the scene, shadows can be brightened with more details while keeping proper highlights. Unfortunately, this approach requires much power and computation for multi-frame capturing, which is luxurious on mobile phones.

This paper proposed another approach of replacing the whole ISP with a neural network, named E2EISP, to transform a single Bayer array paired with its camera settings at capturing time (i.e., white balance and color correction matrices) of an HDR scene to a visually delightful image. For end-to-end optimization, we assume available labels of high-quality enhanced images of HDR scenes that are generated with proper multi-frame merging and image enhancement processes. This assumption is not very strict as many mobile phones already have their own multi-frame HDR imaging engines [4, 1, 12]. We design our E2EISP with main blocks of 1) brightness enhancement in Bayer array domain to adjust each Bayer element in terms of the stop of exposure, which is the mixture of multiple exposure candidates learned from the Bayer input, 2) demosaiced-feature extraction to extract deep feature essential for a pleasing output, and 3) color processing incorporating camera settings to be flexible to various capturing settings (conceptually following conventional ISP, i.e., white balancing, color correction, and gamma correction), where we generate multiple image-processed candidates and regress to final high-quality output. Finally, we carefully construct training losses, which assess good and bad generated images, in two folds. Brightness enhancement is attained with unsupervised loss of well-exposure pixel values and supervised loss of similarity between generated and label images in low frequency. Noise reduction and edge enhancement are achieved with a loss of the similarity between high-frequency components of generated and label images and a generative adversarial loss.

In summary, our contributions are as follows:

- We proposed an E2EISP that transforms the input of not only a single Bayer array but also camera settings to a visually pleasing image of an HDR scene regarding labels of high-quality multi-frame merged images.

- We designed a neural network architecture capable of discriminating shadows and highlights pixels for brightness enhancement and color processing employing camera set-

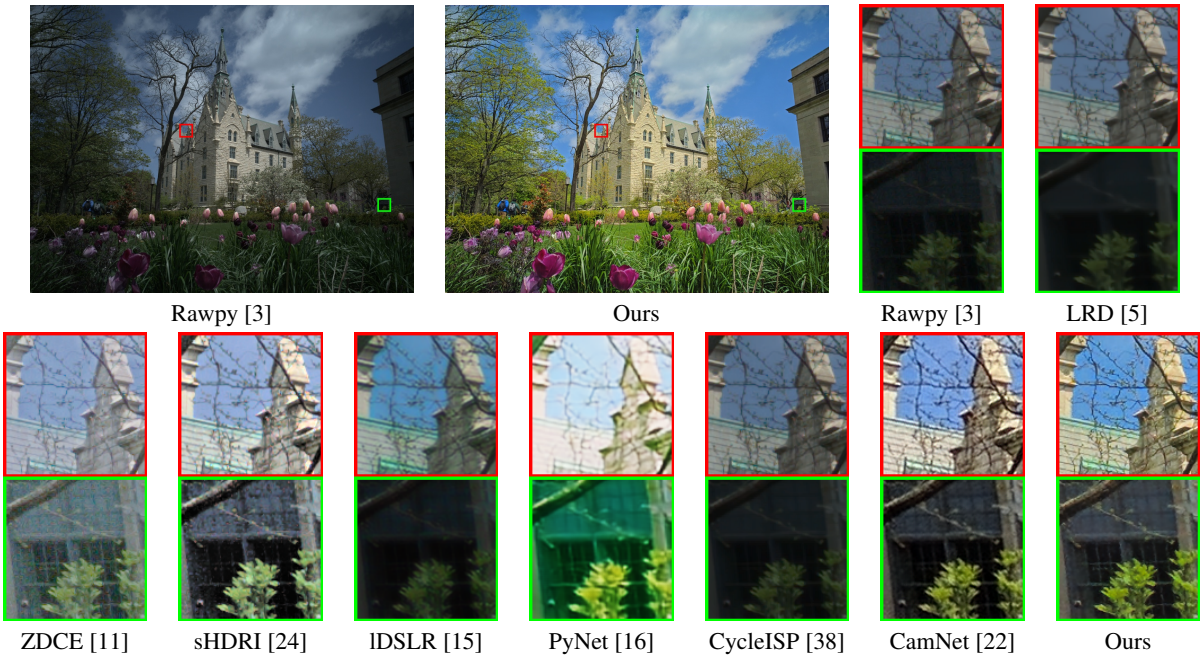

Figure 1. Our enhanced image of a high dynamic range scene in HDR+ dataset [12] generated from a single Bayer array (processed with Rawpy [3] for displaying purpose) and its camera settings. Our method 1) avoids suffering from capturing noise and image signal processing artifacts compared to methods working in sRGB domain of low-light image enhancement (ZDCE) [11] and inverse tone mapping (sHDRI) [24]; 2) effectively brightens shadows (bottom crop) with proper details in highlights (top crop) compared to Bayer array denoising (LRD) [5]; 3) refrains from incorrect color happened in converting Bayer array without camera settings PyNet [16] (failed to generate black color of the window) or CamNet [22] (failed to reproduce cyan color in the roof); and 4) produces sharper edges with more detail in overall.

tings to flexibly mimic the desired transformation.

- We combined supervised, unsupervised, and generative adversarial losses concerning low- and high-frequency components separately for noise reduction, brightness, edge, and detail enhancements.

2. Related work

HDR imaging [29] is a natural solution to deal with HDR scenes, which can be roughly categorized into two groups. One is to capture HDR scenes directly using customized imaging hardware being capable of acquiring light energy in a wider dynamic range [28, 32, 14], nonetheless, this approach is not affordable to mobile phones. Another multi-frame approach [19, 35, 37, 36] assumes that each frame contains some amount of information for various lighting conditions of the scene, subsequently, the merged frame can have the details of both shadows and highlights. However, this approach has potential merging artifacts [36] and the high cost of capturing multiple frames.

Inverse tone mapping [8, 27, 21, 31, 24] is to convert a single low dynamic range (LDR) image into an HDR image first and the result can be tone mapped for displaying later if necessary. The works [8, 21] generate multiple exposure images from an LDR image and then merge them into an HDR image; while the works [24, 31, 27] generate an HDR image from an LDR image directly. These methods share with us the concept of using a single frame, however, as inputs are in the sRGB domain, their results are possibly affected by ISP artifacts which are severe in the case of mobile phone images, e.g., we can observe the bright but noisy result of sHDRI [24] in Fig. 1.

Low-light image enhancement [7, 39, 11] is another approach that mainly focuses on low-light scenes. The works [7, 39] are based on pairs of low-light images taken with a short exposure and clean images captured with sufficiently long exposure. Though this setting gives good pairs for training and generates good results, we cannot apply this setting for HDR scenes because labels taken with long exposure will contain over-saturated highlights. The work [24] proposed to enhance the sRGB image with unsupervised losses, which gets rid of the expensive process of constructing a training dataset. But similar to the inverse tone mapping approach, this reveals more details of shadows but affects by image processing artifacts being in input sRGB images, e.g., in the result of ZDCE [11] in Fig. 1.

Neural network-based ISP can replace the whole ISP or its part(s) with neural network(s) for better performance. For example, Bayer array denoising [5, 38] proposed an unprocessing procedure that synthesizes a Bayer array from a clean sRGB image; and the pairs of clean Bayer arrays and noise-injected Bayer arrays are used to train a Bayer array

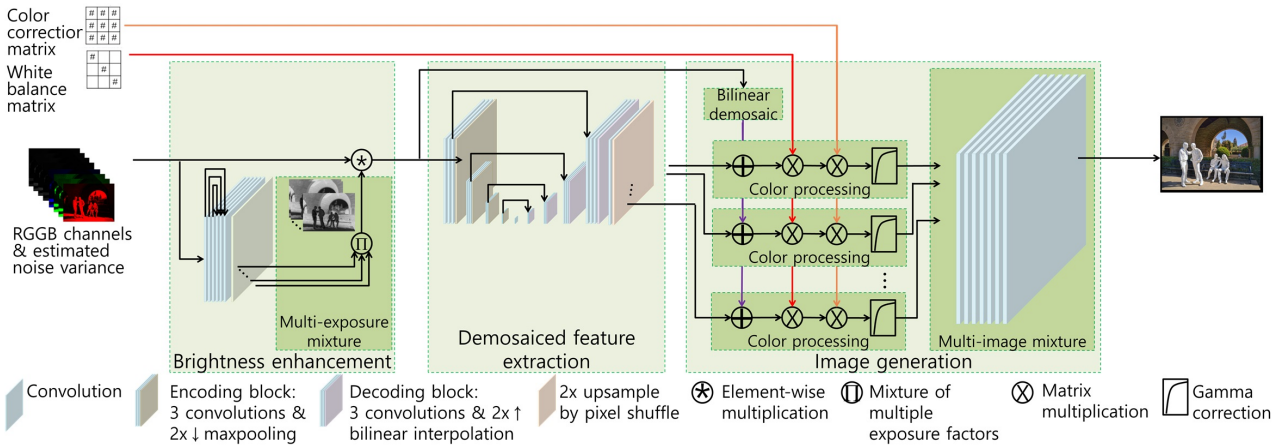

Figure 2. Neural network architecture of proposed E2EISP.

denoising neural network. Though the result in LRD [5] in Fig. 1 is clean compared to conventional ISP of Rawpy [3], it cannot brighten shadows which is one problem of capturing HDR scenes. The work [5] motivates us to synthesize Bayer arrays to pair with multi-frame merged images. Another work [9] tackles both denoising and demosaicking at the same time. As post-processing, the work [15] enhances mobile images using the labels of DSLR images of the same scenes. The whole ISP can also be replaced, as in [16] using the same labels of DSLR images or in [22] using labels of multi-frame merged images. Though these works share the concept of working on Bayer array, the purpose of [16] is not for HDR images which are more challenging and require different dataset, network architecture, and training losses; while the work [22] needs to have two stages training with two different datasets (which is costly) and ignore the camera setting leading to the suffering of failed color as shown in results of CamNet [22] in Fig. 1.

3. Proposed method

3.1. Network architecture

We design our E2EISP as in Fig. 2 to learn the transform of interest with the input of $[x, \sigma_n^2]$, where x and σ_n^2 are $RGGB$ Bayer array and its noise variance of size $h \times w \times 4$.

Brightness enhancement. This subnetwork enhances brightness to reveal details of shadows without oversaturating highlights, hence, discriminates shadow pixels from highlight pixels. Accordingly, we proposed to adjust the intensity of the Bayer array, pixel-by-pixel, so that each pixel in the Bayer array has its own adjusting factor. We learn the adjusting factors directly from the Bayer array since it contains all available information about the scene.

The brightness of the image depends on three factors of ISO speed (i.e., sensor sensitivity), shutter speed, and aperture diameter. The effect of those three factors on brightness can be combined into exposure measured in "stop", where a stop of exposure increased/decreased by 1 means either double or half the amount of light captured while taking the image. Thus, we use the stop of exposure as an interchangeable measure for all three factors and let the brightness enhancement subnetwork learn to adjust the stop of exposure. To improve the reliability of learning, we learn a mixture of S different stops of exposure, x_s^{stop}, with their corresponding S different mixture weights, w_s, as:

$$x^{adjust} = 2^{x^{stop}}, \quad \text{where } x^{stop} = \sum_{s=1,\dots,S} w_s x_s^{stop} \tag{1}$$

Let $x_s^{adjust} = 2^{w_s x_s^{stop}}$, brightness enhancement returns:

$$x^{bright} = [x; \sigma_n^2]\, x^{adjust} = [x; \sigma_n^2] \prod_{s=1,\dots,S} x_s^{adjust} \tag{2}$$

Though the brightness enhancement subnetwork can output both w_s and x_s^{stop}, we simplify the subnetwork by producing a x_s^{adjust} directly as one output channel; and the subnetwork generates output with S channels. We use a neural network in the shape of UNet [30] without downsampling consisting of convolutional layer and $ReLU$ activation [10].

Demosaiced-feature extraction. Though pixels may have a better intensity range after brightness enhancement, they are still in the mosaiced domain. Accordingly, our next processing is to extract feature in demosaiced domain, denoted $x^{feature}$ of size $H \times W \times N$, where $H = 2 \times h, W = 2 \times w$ and N is the number of feature channels. We employ a neural network of UNet [30] with downsampling, which is well known for high capacity due to its multiscale nature, followed by a convolutional layer. Pixel shuffling is added to convert feature to demosaiced resolution $x^{feature}$.

Image generation. Though technically the desired transform can be learned without camera setting (i.e., white balance and color correction matrices) as in [16], and a demosaiced feature with only 3 channels can be considered as

sRGB output. However, it is better to embed the camera settings into the E2EISP because each image was taken with a different camera setting for which camera color may vary. We add basic color processing functions (i.e., skip connection with bilinear demosaic, white balance, color correction, and gamma correction) to our E2EISP to convert data from the camera color space to the sRGB color space. Similar to the brightness enhancement, we employ a mixture model to generate a more reliable output image. We split N channels of $x^{feature}$ into P groups, each of 3 channels denoted $x_p^{feature}$, to generate P color-processed candidates x_p^{sRGB}.

First, because the residual learning [13] with a skip connection is well proven to improve the learning ability of the neural network in not only computer vision but also image processing, we employ a skip connection to our color processing, where we learn a residual image (in RGB linear domain) from the brightened Bayer array. Note that the output of the demosaiced-feature extraction is already high resolution, we demosaic the brighten Bayer array, x^{bright}, by simple bilinear interpolation, because bilinear interpolation can be integrated into our network easily.

$$x_p^{skp} = x_p^{feature} + LinearDemosaic\left(x^{bright}\right) \tag{3}$$

Then, given the white balance matrix and color correction matrix from the camera setting, and the default gamma factor of 2.2, the sRGB candidate is derived as follows:

$$x_p^{sRGB} = Gamma(ColorCorrec.(WhiteBalan.(x_p^{skp}))) \tag{4}$$

The concatenation of those multi-head color-processed candidates will be mixed into a final image $\hat{x}$ by a subnetwork composed of multiple convolutional layers followed by $Leaky\_ReLU$ activation [25].

3.2. Training losses

We construct our training losses using supervised, unsupervised, and generative adversarial losses regarding the generated and label images, denoted $\hat{x}$ and y, respectively.

Exposure loss. The main concern of HDR scenes is to properly brighten shadows without oversaturating the highlights. One of the effective ways is to employ an unsupervised loss to impose the pixel intensity to a range that is easily perceived by human eyes. Hence, we employ the exposure loss [26, 11] as:

$$\ell_{exposure} = \frac{b^2}{3HW}\left\|\hat{x}^{avg} - L\right\|_2^2, \tag{5}$$

where $\hat{x}^{avg}$ is the average of each non-overlapped $b \times b, b = 16$, block of the output $\hat{x}$ and L is the expected brightness of pixels intensity in the enhanced images. We set the expected brightness L to 0.6, following the work [11].

Multi-frequency losses. Since the label images have good brightness enhancement after carefully merging multiple shots, learning brightness from label images helps to complement the unsupervised loss of exposure. Because the low-frequency component contains sufficient information on brightness, we set this brightness loss as:

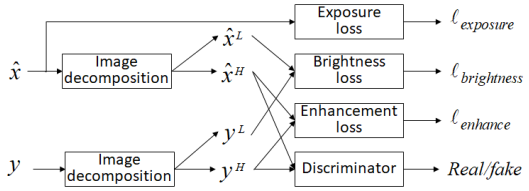

Figure 3. Multi-frequency losses for proposed E2EISP.

$$\ell_{brightness} = \frac{1}{3HW}\left\|\hat{x}^L - y^L\right\|_1, \tag{6}$$

in which $\hat{x}^L$ and y^L are the low-frequency component of the generated and label images, respectively. Each either R, G, or B pixel of $\hat{x}^L$ and y^L is the average of its $b' \times b', b' = 9$, neighbor pixels in the corresponding color channel.

Besides, because label images also have good edge enhancement and noise reduction due to the plentiful amount of information coming from multiple shots, it would be better to learn these effects via an enhancement loss. Note that both edge and noise can be distinguished well in the high-frequency components, we set the enhancement loss as:

$$\ell_{enhance} = \frac{1}{3HW}\left\|\hat{x}^H - y^H\right\|_1 \tag{7}$$

The high-frequency components are calculated as $\hat{x}^H = \hat{x} - \hat{x}^L$ and $y^H = y - y^L$. This loss gives more freedom in balancing between edge enhancement/noise reduction in eq. (7) and brightness enhancement in eqs. (5) and (6), so that combined loss is not dominated by the brightness difference between generated and label images.

Generative adversarial loss. Enforcing edge enhancement and noise reduction in the loss of $\ell_{enhance}$ might lead to loss in delicate details, as ℓ_1 minimization in the loss $\ell_{enhance}$ promotes sparse output [6], i.e., either smooth or abrupt changes in pixel intensity. To deal with this unwanted effect, we employ the concept of the generative adversarial network (GAN) so that our E2EISP can learn the detail and sharpness more effectively from label images.

We implement relativistic average GAN (RaGAN) [17, 18] in our losses, inspired by their discussion that relatively estimating probability of real or fake data (that are the probability of real data given how on-average realistic fake data is or vice versa) is significantly more stable and be able to construct more plausible high-resolution images. Besides, because we focus on using RaGAN [17, 18] to generate sharper edges and more details, the RaGAN is applied to high-frequency components of generated and label images (i.e., fake and real images). Given a discriminator network

Table 1. Ablation studies on network architecture (trained with ℓ_1 loss, average of the last 10 validation steps). Bests are in red.

Configuration	A	B	C	D	E	F
Brightness enhancement	Off	Single	Mixture	Mixture	Mixture	Mixture
Color processing	Mixture	Mixture	Off	OffPlus	Single	Mixture
Validation loss (ℓ_1)	0.056	0.055	0.056	0.056	0.053	0.052
Validation PSNR (dB)	24.07	24.16	24.06	24.11	24.57	24.84

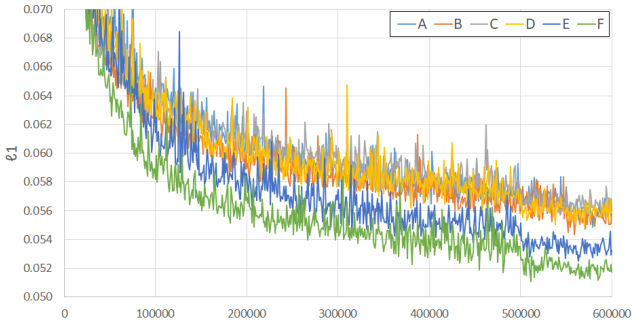

Figure 4. Validation loss for ablation architectures in Table 1.

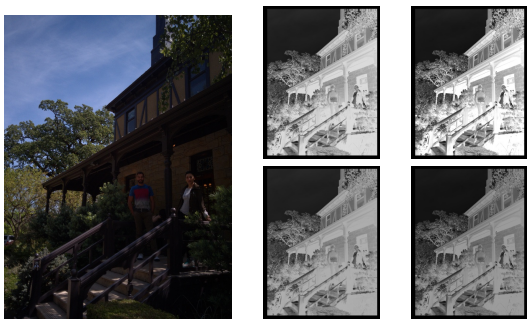

Figure 5. Brightness enhancement subnetwork with a Bayer array input in the left (processed with Rawpy for displaying) constructs detail adjusting factors for $RGGB$ channels, in which large (bright) adjusting factors are for shadows and vice versa.

D, the adversarial and generative losses are:

$$\ell_D^{GAN} = -\mathrm{E}_{y^H}\left[\log\left(\sigma\left(D\left(y^H\right) - \mathrm{E}_{\hat{x}^H}\left[D\left(\hat{x}^H\right)\right]\right)\right)\right] - \mathrm{E}_{\hat{x}^H}\left[\log\left(1 - \sigma\left(D\left(\hat{x}^H\right) - \mathrm{E}_{y^H}\left[D\left(y^H\right)\right]\right)\right)\right]$$

$$\ell_G^{GAN} = -\mathrm{E}_{\hat{x}^H}\left[\log\left(\sigma\left(D\left(\hat{x}^H\right) - \mathrm{E}_{y^H}\left[D\left(y^H\right)\right]\right)\right)\right]$$

4. Experiments

4.1. Experimental setting

We evaluated our method with the HDR+ dataset [12, 2] because this dataset is an extensive dataset with 3640 HDR scenes taken with various mobile phones (Nexus 5/6/5X/6P, Pixel, and Pixel XL). We used the curated subset of 153 images [2] as test images; the rest was divided into training and valid sets. Because of merging, enhancement, and post-processing, label images are not necessarily aligned with their corresponding Bayer arrays. Therefore, we need to register pairs of Bayer arrays and multi-frame merged images by 1) constructing sRGB with conventional ISP (black level subtraction, demosaicking, white balancing, color correction, and gamma correction), 2) registering label images to generated sRGB images (similar to [18], finding matching key points and descriptors with by ORB technique [35], calculating perspective transformation between two images using RANSAC [11], and warping label images into registered ones using the perspective transformation). Furthermore, label images might be enhanced by some image processing techniques like edge enhancement and dehazing [2, 16], they may contain some unwanted artifacts, e.g., sharpening halo artifacts. Accordingly, similar to [5], both sRGB and registered label images are downsampled by the factor of two to reduce potential misalignment after registration and unwanted artifacts. We then synthesize Bayer array from downsampled sRGB images on-the-fly by inversions of gamma correction, color correction, and white balance using metadata; and then mosaicking as in [5]. We also visually inspect and remove those pairs which are obviously not aligned. Our E2EISP, having about 12M parameters, is trained using Adam optimizer [20] with learning rate of (2e-5, 0.5e-5, 0.125e-5) for (500k, 50k, 50k) iterations, respectively. Our training takes roughly 2 days for each training phase of without and with RaGAN loss on a Tesla V100 GPU. Parameters for losses are $\lambda_{exposure} = 1.0$, $\lambda_{brightness} = 1.0$, $\lambda_{enhance} = 3.0$, and $\lambda_{GAN} = 0.003$.

4.2. Ablation study

Network architecture. Our E2EISP composed of two main blocks of brightness enhancement and color processing subnetworks, to transparently verify effectiveness of each subnetwork and their corresponding techniques, we studied multiple variations of our network as shown in Table 1. We trained these variants with ℓ_1 to verify their expressive power, where final validation loss (or PSNR) is shown also in Table 1 and training progress is shown in Fig. 4.

Brightness enhancement subnetwork, though trained in an end-to-end optimization, effectively generates meaningful adjusting factors for each pixel in the Bayer array, as depicted in Fig. 5. The subnetwork produces large adjusting factors (bright) for shadows, e.g., the house. Oppositely, small adjusting factors (dark) are given to highlights of clouds in the sky. Due to pixel-by-pixel brightness adjustment, the adjusting factors are detailed enough to adapt to fine intensity changes, e.g., thin branches of the trees. The mixture model, eqs. 1 and 2, in this subnetwork for

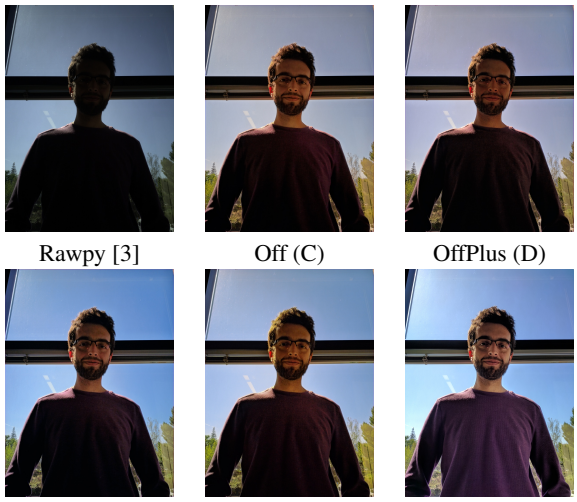

Figure 6. Visual effect of color processing in our E2EISP.

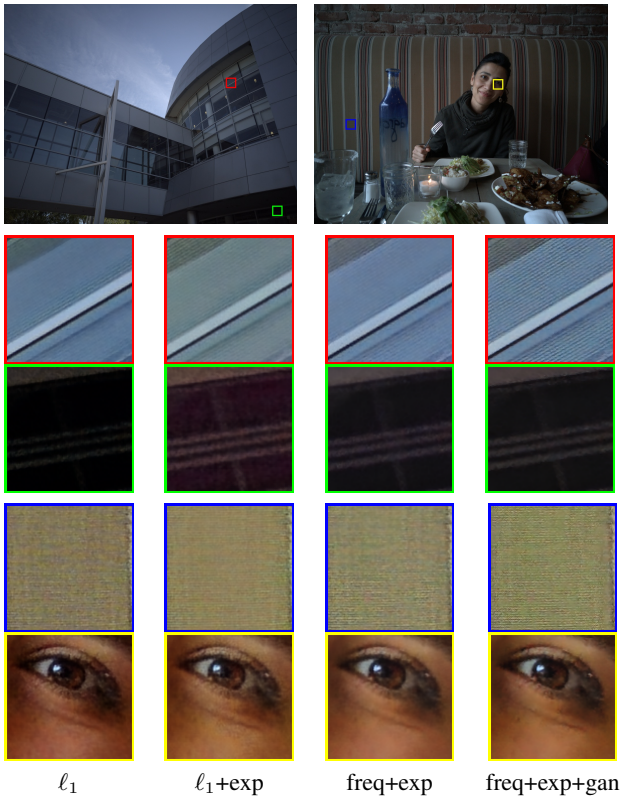

Figure 7. Ablation study on training losses (Bayer array inputs are in top row, processed with Rawpy for displaying). Exposure loss helps to brighten shadows, frequency-based losses reduces noise much, and generative adversarial loss enhances edges and details.

reliable adjusting factor is verified in Table 1 by comparing three cases: (A) brightness enhancement network removed, (B) single model, and (F) mixture model. Best validation PSNR of mixture model (F) among three confirms the effectiveness of brightness enhancement and its mixture model.

Table 2. Comparison to state-of-the-arts in terms of common image quality assessment metrics. Bests (2^{nd} bests) are in red (blue).

Quality metric	PSNR	SSIM	MSSSIM	FSIM
Rawpy [3]	13.545	0.446	0.624	0.730
LRD [5]	13.622	0.505	0.650	0.772
ZDCE [11]	13.545	0.446	0.624	0.730
sHDRI [24]	15.835	0.390	0.576	0.740
lDSLR [15]	15.425	0.548	0.657	0.765
PyNet [16]	15.977	0.573	0.677	0.765
CycleISP [38]	13.621	0.501	0.649	0.773
CamNet [22] ℓ_1	19.178	0.606	0.701	0.794
CamNet [22]	19.107	0.617	0.709	0.795
Ours ℓ_1	19.464	0.640	0.721	0.796
Ours	19.147	0.617	0.716	0.791

Explicit color processing with camera setting helps, as shown in Fig. 6. The networks (C) without the camera setting (i.e., color processing off) and (D) with color processing off but equipped with a post-processing network with a similar design as the multi-image mixture network, produces a false color of the sky compared to conventional ISP of Rawpy [3] and HDR+ [12], supposed to be the true color. With camera settings, either single (E) or mixture (F) model, the sky color matches that of Rawpy and the mixture model (F) has a more vivid color. Objectively, color processing single (E) and mixture (F) models have better training with smaller losses and better PSNRs, see Table 1.

Losses. We verified the effectiveness of losses in Fig. 7, starting with the common ℓ_1 loss, which measures the mean absolute difference between the generated $\hat{x}$ and the label y. The exposure loss in eq. (5) helps to enhance the dark regions more compared to ℓ_1 loss as in the second crop row, but noise is still irritating as in the first crop row. Employing multi-frequency losses keeps brightening and effectively suppresses noise, however, at a cost of some details lost. Our proposed losses together, including generative adversarial loss, can generate proper brightness and detail enhancement without irritating noise.

4.3. Comparison to state-of-the-arts

Singe-frame approaches. In Figs. 1, 8, 9, and 10, we compared our results with sRGB-domain image enhancement methods of low-light image enhancement (ZDCE) [11], inverse tone mapping (sHDRI) [24] (tone-mapped with [23] for displaying), and learning from high-quality DSLR images (lDSLR) [15]. sRGB inputs for these methods are produced by conventional ISP of Rawpy [3] with the default configuration. As the output of Rawpy contains much noise due to capturing on mobile phones, the results of ZDCE [11] and sHDRI [24] contain much noise and artifacts, though they significantly enhance the brightness. Besides, with clean labels, lDSLR [15] outputs clean images, however, cannot enhance the brightness of shadows

Figure 8. Visual comparison to state-of-the-art methods on an HDR scene.

Figure 9. Visual comparison to state-of-the-art methods on an HDR scene.

as it does not focus on HDR scenes. Our work is then compared to approaches working with Bayer array: Bayer array denoising of LRD [5] and CycleISP [38]; and transforming Bayer array to high-quality DLSR images (PyNet) [16]. Note that these work trained their neural networks with clean labels, they effectively denoise and generate clean images. However, LRD [5] and CycleISP [38] cannot improve the shadows, which are still in dark, while PyNet [16] and CamNet [22] suffers from not only saturated highlights, e.g., rocks of Fig. 8 or incorrect color as camera settings are not involved in the generation process, e.g., man's face in Fig. 8. Compared to these single-frame methods, our work properly enhances shadows, e.g., the face in Fig. 8 and the foot of the tree in Fig. 9, while keeping sufficient details in the highlight of rocks in Fig. 8 and hill ground in Fig. 9. Note our results do not suffer from either noise or demo-

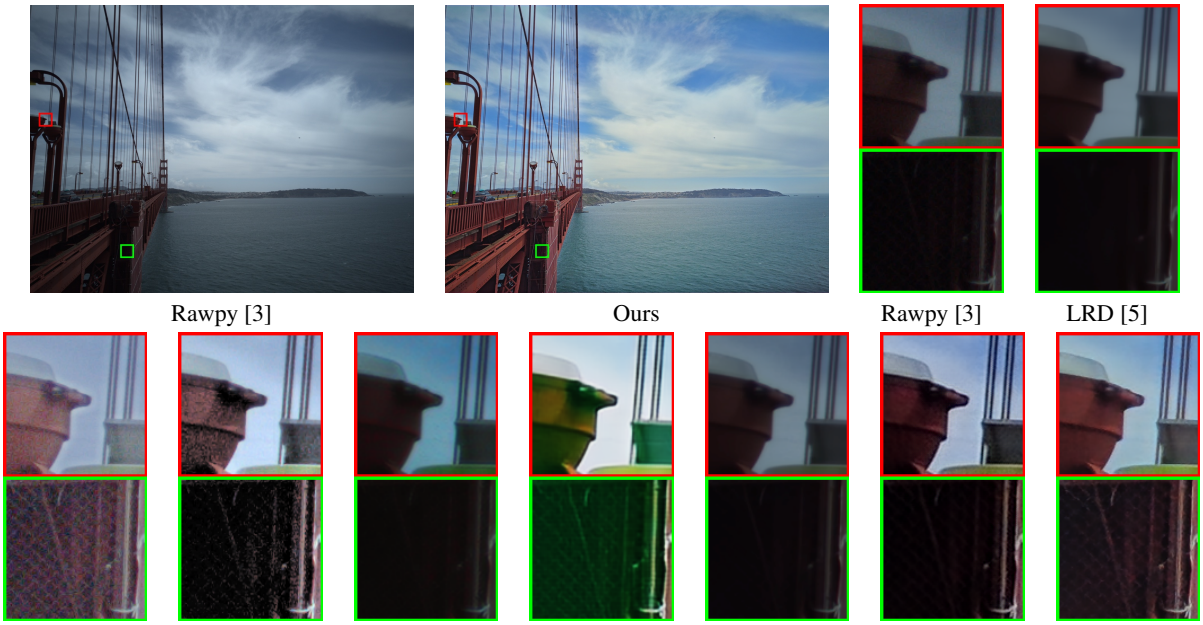

Figure 10. Visual comparison to state-of-the-art methods on an HDR scene.

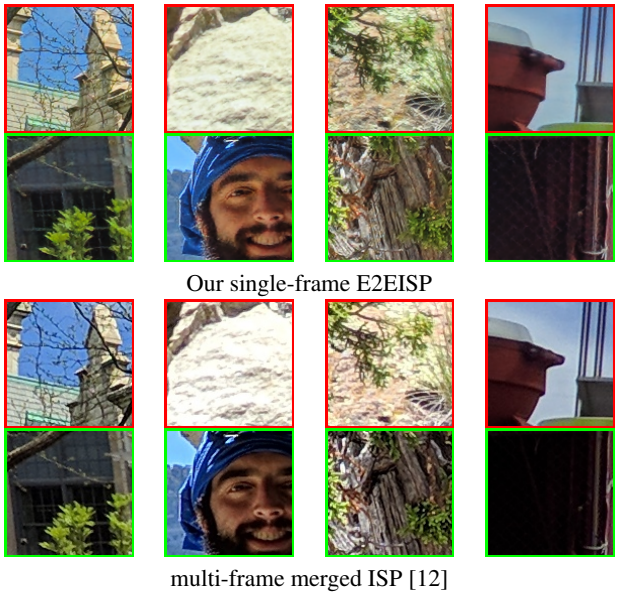

Figure 11. Comparison to multi-frame merged label images [12]. Locations of crops can be found in Figs. 1, 8, 9, and 10.

saicking artifacts, while edges are enhanced without much halo artifacts and details are well generated.

In addition to visual inspection, we verify those methods in terms of objective quality measured in well-known metrics of PSNR, SSIM [33], MSSSIM [34], and FSIM [40] regarding the labels of multi-frame merge images [12] in Table 2. Note that except for ours and CamNet [22], other methods were not trained with multi-frame merge labels, however, we include them in the table for completeness and simple assessment of how close their methods are to the multi-frame output. From the Table 2, our network design trained with ℓ_1 generates outputs closest to the multi-frame merge images. Our design loss terms, including GAN, lower the quality metrics due to the generation of novel textures/details apart from labels but gains in subjective quality as discussed in ablation studies.

Multi-frame approach. As depicted in Fig. 11, multi-frame labels of HDR+ [12] with much more information from multiple frames generally constructs cleaner images with more details. However, with our unsupervised exposure loss in eq. (5) we can enhance the brightness of shadows better, such as the face and foot of the tree without losing details in rocks or hill ground. Further, with proper training of RGAN [17, 18] and multi-frequency losses, our E2EISP generates thin and sharp edges with fewer halo artifacts, e.g., of tree branches or grass in hill ground.

5. Conclusion

This paper proposed an end-to-end optimized image signal processing that transforms Bayer arrays and camera settings to enhanced sRGB images for high dynamic range scenes taken on mobile phones. We designed a suitable neural network architecture to express the transformation and end-to-end optimized it with proper training losses considering desired properties of the enhanced images and labels of multi-frame merged images. The proposed method constructs clean images with enhanced shadows, detail highlights, sharp edges, and noise reduction.

References

[1] Apple iphone 12 pro, accessed October 28, 2020. https://www.apple.com/iphone-12-pro/.

[2] Hdr+ burst photography dataset, accessed October 28, 2020. https://hdrplusdata.org/dataset.html.

[3] Rawpy, accessed October 28, 2020. https://pypi.org/project/rawpy/.

[4] Samsung galaxy s20 5g, accessed October 28, 2020. https://www.samsung.com/us/mobile/galaxy-s20-5g/camera/.

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2017)*, 36(6), Nov. 2017.

[9] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. 35(6), Nov. 2016.

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 315–323, 2011.

[11] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] 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 (Proc. SIGGRAPH Asia)*, 35(6), 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid PajÄ…k, Dikpal Reddy, Orazio Gallo, Jing Liu abd Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, and Kari Pulli. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2014)*, 33(6), December 2014.

[15] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] A. Ignatov, L. Van Gool, and R. Timofte. Replacing mobile camera isp with a single deep learning model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2275–2285, 2020.

[17] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[18] Alexia Jolicoeur-Martineau. On relativistic f-divergences. *arXiv preprint arXiv:1901.02474*, 2019.

[19] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[22] Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang. Cameranet: A two-stage framework for effective camera isp learning. *IEEE Transactions on Image Processing*, 30:2248–2262, 2021.

[23] Z. Liang, J. Xu, D. Zhang, Z. Cao, and L. Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4758–4766, 2018.

[24] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[25] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML) (2013)*, 2013.

[26] T. Mertens, J. Kautz, and F. Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. *Computer Graphics Forum*, 28(1):161–171, 2009.

[27] Kenta Moriwaki, Ryota Yoshihashi, Rei Kawakami, Shaodi You, and Takeshi Naemura. Hybrid loss for learning single-image-based hdr reconstruction, 2018. https://arxiv.org/abs/1812.07134.

[28] S. K. Nayar and T. Mitsunaga. High dynamic range imaging: spatially varying pixel exposures. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 1, pages 472–479 vol.1, 2000.

[29] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. Morgan Kaufmann, 2nd edition, 2010.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation.

In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[31] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Trans. Graph.*, 39(4), July 2020.

[32] J. Tumblin, A. Agrawal, and R. Raskar. Why i want a gradient camera. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 103–110 vol. 1, 2005.

[33] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[34] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[35] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[36] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. pages 1751–1760, 2019.

[37] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. pages 41–50, 2019.

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[39] C. Zhang, Q. Yan, Y. Zhu, X. Li, J. Sun, and Y. Zhang. Attention-based network for low-light image enhancement. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Los Alamitos, CA, USA, jul 2020. IEEE Computer Society.

[40] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.