Supplementary Materials of Real-world Image Enhancement Based on Multi-Exposure LDR Images

Anonymous WACV 2023 Algorithms Track submission
Paper ID 0259

1. HDR evaluation on NTIRE 2022 dataset

The NTIRE 2022 HDR dataset [8] consists of approximately 1,500 training, 60 validation and 201 testing examples in RGB domain. Each example in the training set is composed of three input LDR images, i.e. short, medium and long exposures, and a related ground-truth HDR image aligned with the central medium frame. Input images are obtained using a pixel-measurement model, which includes several sources of noise. Since this dataset doesn’t release the ground-truth of its original validation and testing set, we divide its training set into two subsets, and use them for training and testing respectively. 1,200 images are randomly selected as the training set, and the remaining 300 images are used as the testing set. In consideration of the workload, we randomly select 50 images from the testing set for MOS score evaluation.

As given in Table 1, it can be seen that our method outperforms the prior arts with PSNR/SSIM/MOS at 0.18 dB/0.007/0.11. Fig. 1 indicates that our EMVNet is able to produce HDR output with better perceptual quality.

2. Real-world E2E-ISP

In this section, we give more details of the evaluation on real-world E2E-ISP task. E2E-ISP is more complicated than HDR since it is a hybrid problem which composes of various tasks such as multi-image fusion (HDR), domain transfer (demosaicing), and color tuning (auto white balancing). Similar to raw HDR, the inputs to the E2E-ISP networks are 10-bit raw images, while the output is 3-channel RGB image. As given in Fig. 2, in scenarios with different lighting conditions and exposure variations, our EMVNet is able to produce RGB images with more neural color, and w/o any motion ghost (in contrast to prior arts’ results as given in Fig. 2(d)(e)). Since the real-world cellphone adopts dol sensor to capture the raw LDR images, the time gap between the two captures will be less than 100ms. As a result, there will not be much scenario with motion differences as large as Kalantari’s dataset [3].

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>SSIM</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalantari et al. [3]</td>
<td>35.24</td>
<td>0.9544</td>
<td>-</td>
</tr>
<tr>
<td>Yan et al. [10]</td>
<td>35.79</td>
<td>0.9601</td>
<td>1.99</td>
</tr>
<tr>
<td>Niu et al. [7]</td>
<td>36.13</td>
<td>0.9622</td>
<td>2.01</td>
</tr>
<tr>
<td>Liu et al. [6]</td>
<td>36.26</td>
<td>0.9627</td>
<td>1.96</td>
</tr>
<tr>
<td>Our EMVNet</td>
<td>36.44</td>
<td>0.9691</td>
<td>1.85</td>
</tr>
</tbody>
</table>

Our training RGB ground-truth images are HDR+ images processed by the ISP of Google phones, and the testing raw images are captured by OPPO phones with a different sensor (Sony IM766). Such mismatch might influence the accuracy. To solve this problem, we create a small tuning dataset with 16 RWMR raw images as input, process them by our EMVNet, and manually tune the color tone by Photoshop to make the output image have better perceptual quality. We further fine-tune our EMVNet on this 16 manually tuned images for 100 epochs. From Fig. 3, we notice that the color is further enhanced after fine-tuning. That means with human label, we are able to handle the mismatch between different sensors.

3. Ablation studies

3.1. Usage of weakly-supervised loss function

In Fig. 4, we give the illustrations of applying our weakly-supervised loss functions for E2E-ISP task (in raw HDR task the visualization difference is not that significant due to lacking appropriate AWB and demosaicing modules). We observe that with the usage of weakly-supervised loss function, we receive more reliable color in the output RGB images. w/o the usage of weakly-supervised loss, the blue lights and orange bars turn into purple and yellow in the top image, and the yellow light in the bottom image becomes yellow-green. Since the E2E-ISP is more complicated than HDR, keeping the color consistency will be more important.

Our proposed weakly-supervised loss function is useful to deal with this issue.

Our proposed weakly-supervised loss function can be
applied as an additional component to any of image enhancement networks. In Table 2, we use the state-of-the-art HDR method ADNet [6] and E2E-ISP method PyNet-CA [4] as baseline, retrain their official code with additional weakly-supervised loss proposed in this paper. It can be seen that the accuracy of the baseline methods is significantly improved. This demonstrates the effectiveness of the proposed weakly-supervised loss function. We also notice that the accuracy of prior arts with weakly-supervised learning are still lower than our EMVNet, because the matching volume is another key component which contributes to our considerable performance.

3.2. hyper-parameter tuning

We evaluate the accuracy of EMVNet trained with different hyper-parameters in the loss functions. As mentioned in Section 4 of our paper, the generator loss $L_G$ consists of the image content loss $L_c$, the perceptual loss $L_p$, the adversarial loss $L_a$, and the weakly-supervised loss $L_s$, as described in Eq. 1. $\lambda$, $\eta$, $\alpha$ determine the contribution of adversarial loss, content loss, and the weakly-supervised loss.

$$L_G = L_p + \lambda L_a + \eta L_c + \alpha L_s. \quad (1)$$

Since the content loss consists of three terms due to the two intermediate outputs $Y'$, $Y''$ from stacked hourglass, there will be two additional hyper-parameters here, given as the $\beta_1, \beta_2$ in Eq. 2.

$$L_c = L_1(Y, Y^*) + \beta_2 * L_1(Y'', Y^*) + \beta_1 * L_1(Y', Y^*) \quad (2)$$

The weakly-supervised loss functions $L_s$ also have two thresholds $S_{pix}, S_{pat}$ for the pixel version $L_{s,pix}$ and the patch version $L_{s,pat}$ respectively, as given in Eq. 2 and Eq. 3 in Section 4.1 of our paper. As a result, there are $3 + 2 + 2 = 7$ hyper-parameters in total. In our paper, we preset $\beta_1 = 0.75, \beta_2 = 0.5, S_{pix} = 0.25, S_{pat} = 0.5$. In this section, we train EMVNet with different combinations of all these 7 hyper-parameters on HDR+ images for raw HDR to check the network robustness.

3.2.1 Different weights of the loss functions

First we fix $\lambda = 0.001, \alpha = 0.25, \eta = 0.001, S_{pix} = 0.25, S_{pat} = 0.5$, and evaluate different values of $\beta_1, \beta_2$ to see whether adding constraint on the intermediate outputs of the stacked hourglass will be helpful. Since the intermediate outputs $Y', Y''$ are only used during the training, $\beta_1, \beta_2$ should be set to some values smaller than 1. $Y''$ is the output of the first hourglass, and $Y'$ is the output of the second hourglass, so empirically we set $\beta_2 \leq \beta_1$. From the results given in Table 3, it can be seen that accuracy doesn't change much along with different combi-nations of $\beta_1, \beta_2$. But if set both of them to 0 (row 2), which means that the intermediate outputs are not utilized during the training, the PSNR will decrease 0.08 dB. In contrast,
Figure 2. More comparisons of our EMVNet and state-of-the-art E2E-ISP methods on real-world raw images. (a)(b) have large exposure differences. (c)(d)(e) have significant motion between long/short-exposure images, see the person head at bottom left of (c), the person leg in (d), and the school bus in (e). (c)(d)(f) are captured in extreme low-light scenarios. Our EVMNet produces better results consistently.
if we increase $\beta_1, \beta_2$ to 1 (row 7), which means that the intermediate outputs have equal weights as the final output, the accuracy will also decrease 0.05 dB. In our final implementation, we select the one with the highest PSNR/SSIM combination, which is $\beta_1 = 0.75, \beta_2 = 0.5$ as given in row 6 of Table 3.

In Table 4, we give EMVNet trained with different combinations of $\lambda, \alpha, \eta$ to evaluate the robustness when giving different weights for content loss, GAN learning, and weakly supervised loss. During the evaluation, other hyper-parameters are fixed as $\beta_1 = 0.75, \beta_2 = 0.5, S_{\text{pix}} = 0.25, S_{\text{pat}} = 0.5$.

First, we notice that using different values of $\lambda$ (row 2-4) will not lead to significant difference on the accuracy.
(\(< 0.05 \text{ dB})\). This observation is consistent to [9], where adversarial loss has less impact on the image enhancement accuracy as well. Second, by using different \(\alpha\) for weakly-supervised loss \(L_s\), the accuracy varies. The PSNR drops 0.1 dB when setting \(\alpha = 1\) (row 7). This tells us that over-emphasizing the pair-wise constraint will also decrease the performance. In Table 4 of our paper, we already show that ignoring the weakly-supervised loss function (\(\alpha = 0\)) will decrease the PSNR 0.2-0.3 dB. But if we have this loss, using different weights in an appropriate range (row 5-7) will not lead to significant accuracy change (\(< 0.1 \text{ dB}\)). Third, we observe that increasing \(\eta\) will lead to better PSNR (row 8-10), but with a cost of SSIM reduction. Since \(\eta\) is the weight of \(L_1\) loss, if we emphasize it too much, the network might overfit on the training images and receive lower perceptual quality in the unseen scenarios. So finally we choose the group of \(\lambda = 0.001, \alpha = 0.25, \eta = 0.001\) given in row 2 of Table 4.

### 3.3. Different number of input images

Since the EMVNet is not limited to the number of the 607 input images, we did an ablation study on the accuracy versus the number of input images. We use the raw HDR task and HDR+ dataset for this purpose. Besides using the two-input EMVNet shown in our paper, we train an additional EMVNet with 3 input images, while the exposure biases are \(\{-\alpha, 0, \alpha\}, \alpha \in \{2, 4, 8, 16\}\). We also compare the state-of-the-art single-input HDR methods [5][2][4] which is re-trained by their official code on the same HDR+ dataset with single LDR image as input. In Table 6. it can be seen that the multi-inputs HDR methods demonstrate a larger margin compared to single-input HDR methods [5][2][4] (row 4-5 vs. row 2-3). The major reason is that most of the current single image HDR methods are evaluated on images captured by DSLR cameras, which has less sensor noise and limited scenarios. For cellphone images which are captured in extreme lighting conditions like night scenes, the single-image HDR methods don’t work well. If we use three inputs, the accuracy can be further enhanced around 0.17 dB (row 5 vs. row 4). This accuracy improvement is not significant because the more input images we use, the more difficulty we will have during the motion and exposure alignment. In addition, in consideration of the power consumption, capturing three images is not very practical in a cellphone in contrast to current two-image version of Dolby sensor.

### 3.4. Efficient version EMVNet-lite

As mentioned in section 5.4.3 of our paper, we create an simplified version of our EMVNet in consideration of the efficiency, called EMVNet-lite. We replace the standard convolutional layers by depthwise convolutional layers in the feature extraction module. All the convolutional layers...
Table 7. PSNR/SSIM/runtime of the simplified EMVNets compared to original version on HDR+ dataset. The runtime (second) is calculated on single A100 GPU with 4K resolution images (12M pixels).

<table>
<thead>
<tr>
<th>Method</th>
<th>Raw HDR</th>
<th>E2E-ISP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>EMVNet-lite</td>
<td>37.111</td>
<td>0.9781</td>
</tr>
<tr>
<td>EMVNet-lite-os</td>
<td>36.899</td>
<td>0.9712</td>
</tr>
<tr>
<td>EMVNet</td>
<td>37.377</td>
<td>0.9824</td>
</tr>
</tbody>
</table>

This accelerates the network x30 compared to original EMVNet, with a 0.26 dB/0.005 PSNR/SSIM drop in total (row 3 vs. row 5), as given in Table 7. If we don’t follow the above step-to-step fine-tuning, but directly do an one-shot training from scratch, the accuracy of the resulting EMVNet-lite model will decrease 0.47 dB/0.011 PSNR/SSIM compared to original EMVNet, given as EMVNet-lite-os in Table 7 (row 4). The quality reduction of the output images of the EMVNet-lite is not very significant in human vision, as shown in Fig. 5. With further network quantization and revision based on NPU’s requirement, the EMVNet-lite has potential to fit on current cellphone.

References


Figure 5. The difference between the EMVNet-lite and standard EMVNet is not very significant in human vision.