Supplementry 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.

Table 1. Experimental results on the NTIRE 2022 HDR dataset [8]. Bold font indicates the best over the columns. All networks are trained on the same dataset. For MOS, the lower the better.

Method	$PSNR_\mu$	$SSIM_\mu$	MOS
Kalantari et al. [3]	35.24	0.9544	-
Yan et al. [10]	35.79	0.9601	1.99
Niu et al. [7]	36.13	0.9622	2.01
Liu et al. [6]	36.26	0.9627	1.96
Our EMVNet	**36.44**	**0.9691**	**1.85**

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].

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 IMM766). 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

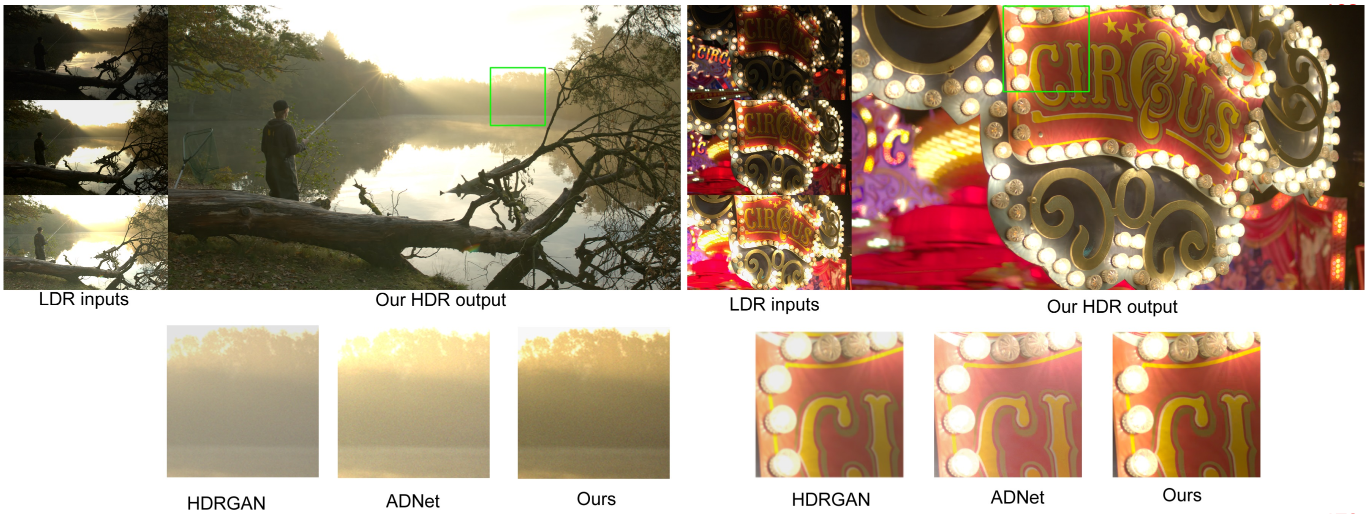


Figure 1. Comparisons of our EMVNet and state-of-the-art HDR methods on NTIRE 2022 images.

Table 2. Comparison to the state-of-the-art methods on the validation images of HDR+ dataset. For raw HDR, we calculate the PSNR/SSIM in the linear raw domain using the merged bursts as the ground-truth. For E2E-ISP, we calculate the PSNR/SSIM in the RGB domain using the ISP processed JPEG images as the ground-truth. Bold font indicates the best over the columns. All the approaches are trained on the same training set.

	Raw HDR		E2E-ISP	
Method	PSNR	SSIM	PSNR	SSIM
ADNet	36.351	0.9670	-	-
ADNet + our loss	36.513	0.9730	-	-
PyNet-CA	-	-	35.349	0.9479
PyNet-CA + our loss	-	-	35.669	0.9533
Ours EMVNet	**37.377**	**0.9824**	**36.891**	**0.9612**

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. λ, η, α determine the contribution of adversarial loss, content loss, and the weakly-supervised loss.

$$L_G = L_p + \lambda L_a + \eta L_c + \alpha L_s, \qquad (1)$$

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 β_1, β_2 in Eq. 2.

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

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 β_1, β_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, β_1, β_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 the accuracy doesn't change much along with different combinations of β_1, β_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,

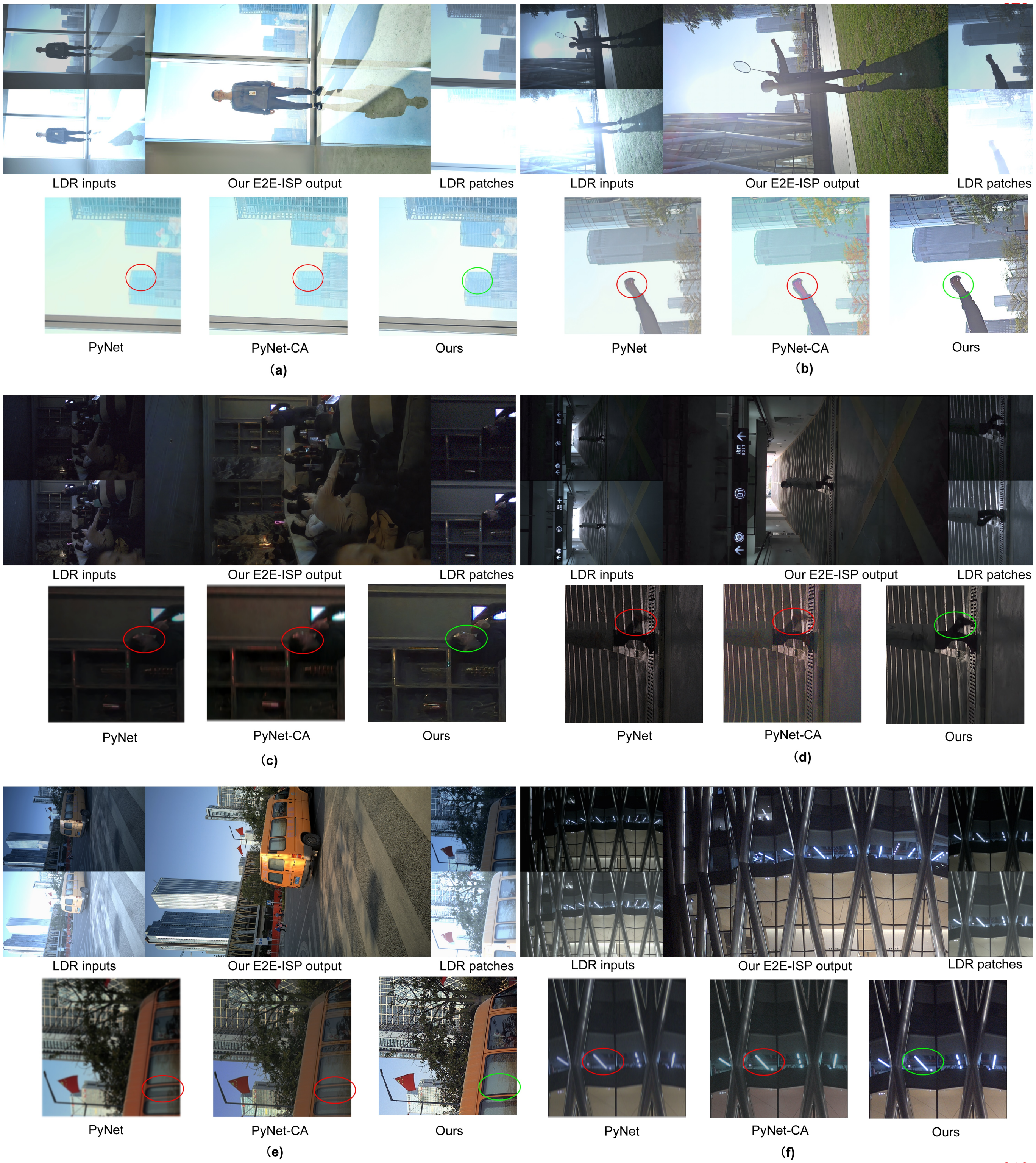
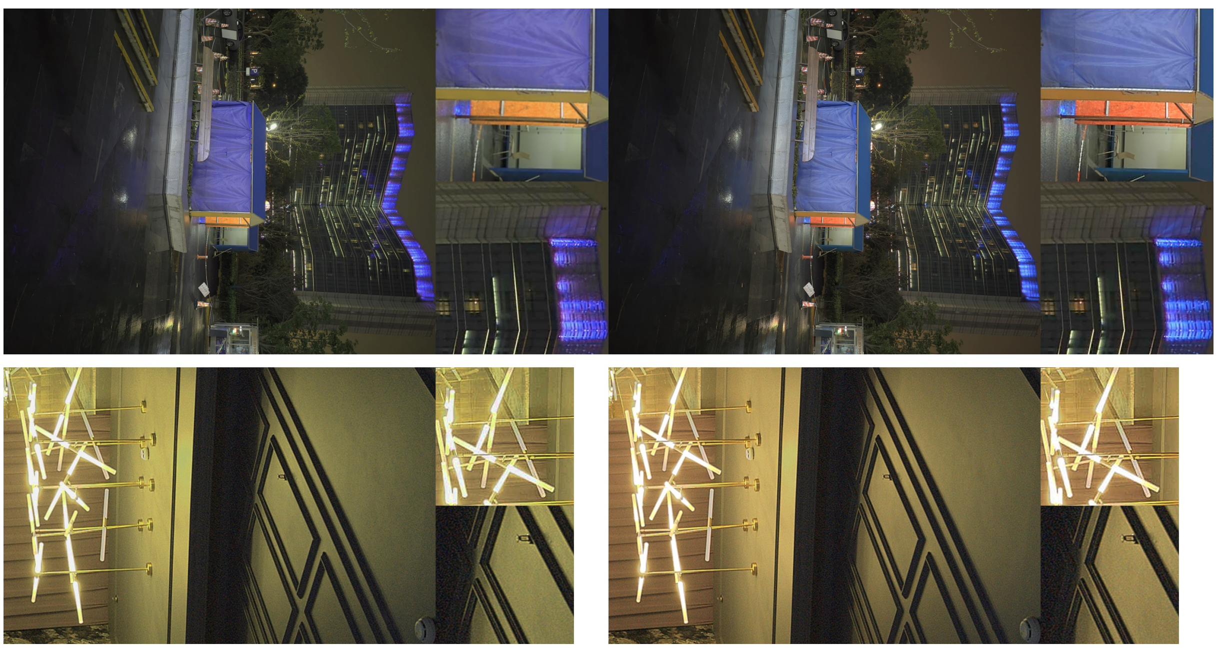


Figure 2. More comparisons of our EMVNet and state-of-the-art E2E-ISP methods on real-world raw images. (a)(b) have large exposure differences. (c)(d)(e) have significant motion between long/short-exposure images, see the person head at bottom left of (c), the person leg in (d), and the school bus in (e). (c)(d)(f) are captured in extreme low-light scenarios. Our EVMNet produces better results consistently.

EMVNet output for E2E-ISP task trained on HDR+ only

EMVNet output for E2E-ISP task trained on HDR+ and fine-tuned on manually labeled data from IMM766 sensor

Figure 3. Fine-tuning with manually labeled ground-truth with same sensor as the testing images can further improve the output image quality.



EMVNet output w/o weakly supervised loss for E2E-ISP

EMVNet output with weakly supervised loss for E2E-ISP

Figure 4. Expample output of EMVNets with and w/o weakly-supervised loss functions.

if we increase β_1, β_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 λ, α, η 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 $\beta_1 = 0.75, \beta_2 = 0.5, S_{pix} = 0.25, S_{pat} = 0.5$.

First, we notice that using different values of λ (row 2-4) will not lead to significant difference on the accuracy

Table 3. Raw HDR accuracy evaluation on HDR+ validation images with different hyper-parameters of loss functions. $\lambda = 0.001, \alpha = 0.25, \eta = 0.001, S_{pix} = 0.25, S_{pat} = 0.5$ are fixed for all rows.

	β_1	β_2	PSNR	SSIM
EMVNet	0	0	37.299	0.9802
EMVNet	0	0.25	37.331	0.9816
EMVNet	0.25	0.5	37.343	0.9819
EMVNet	0.5	0.5	37.346	0.9827
EMVNet	0.5	0.75	37.377	0.9824
EMVNet	1	1	37.326	0.9816

Table 4. Raw HDR accuracy evaluation on HDR+ validation images with different hyper-parameters of loss functions. $\beta_1 = 0.75, \beta_2 = 0.5, S_{pix} = 0.25, S_{pat} = 0.5$ are fixed for all rows.

	λ	α	η	PSNR	SSIM
EMVNet	0.001	0.25	0.001	37.377	0.9824
EMVNet	0.01	0.25	0.001	37.333	0.9811
EMVNet	0.1	0.25	0.001	37.360	0.9817
EMVNet	0.001	0.1	0.001	37.322	0.9799
EMVNet	0.001	0.5	0.001	37.346	0.9821
EMVNet	0.001	1	0.001	37.272	0.9780
EMVNet	0.001	0.25	0.005	37.379	0.9820
EMVNet	0.001	0.25	0.01	37.391	0.9798
EMVNet	0.001	0.25	0.1	37.427	0.9762

(< 0.05 dB). This observation is consistent to [9], where adversarial loss has less impact on the image enhancement accuracy as well. Second, by using different α 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 dB). Third, we observe that increasing η will lead to better PSNR (row 8-10), but with a cost of SSIM reduction. Since η 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.

3.2.2 Different thresholds in the weakly-supervised losses

Next, we evaluate different thresholds S_{pix} and S_{pat} in the weakly supervised loss function. We train EMVNet with the usage of $L_{s,pix}$ and $L_{s,pat}$ respectively to find the best S_{pix} and S_{pat}. The weights of the loss functions are fixed as $\lambda = 0.001, \alpha = 0.25, \eta = 0.001, \beta_1 = 0.75, \beta_2 = 0.5$. From Table 5, we find that the by setting different thresholds for the weakly-supervised loss functions, the PSNR/SSIM don't change much, within a range about 0.08 dB/0.003. We also notice that the EMVNets trained with the patch-version loss have better accuracy than the pixel-version. This makes sense because the patch-version loss is more robust to noise.

Table 5. Raw HDR accuracy evaluation on HDR+ validation images with different thresholds of weakly-supervised losses.

	S_{pix}	S_{pat}	PSNR	SSIM
EMVNet	0.1	-	37.126	0.9800
EMVNet	0.25	-	37.161	0.9798
EMVNet	0.5	-	37.111	0.9781
EMVNet	1	-	37.130	0.9774
EMVNet	-	0.1	37.192	0.9766
EMVNet	-	0.25	37.173	0.9782
EMVNet	-	0.5	37.276	0.9789
EMVNet	-	1	37.211	0.9793

Table 6. Raw HDR accuracy comparison on the validation images of HDR+ dataset. All the approaches are trained on the same training set.

Method	Number of inputs	PSNR/SSIM
Liu et al. [5]	1	35.32/0.9538
Chen et al. [2]	1	35.79/0.9512
EMVNet	2	37.38/0.9824
EMVNet	3	37.55/0.9835

In our final implementation, we select the ones with the best PSNR/SSIM, as $S_{pix} = 0.25, S_{pat} = 0.5$ (row 3, row 8).

3.3. Different number of input images

Since the EMVNet is not limited to the number of the 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 the two-input EMVNet shown in our paper, we train another EMVNet with 3 input images, while the exposure biases are $\{-a, 0, a\}, a \in \{2, 4, 8, 16\}$. We also compare the state-of-the-art single-input HDR methods [5][2], which is re-trained by their official code on the same HDR+ datasets with single LDR image as input. In Table 6. it can be seen that the multi-inputs HDR methods demonstrate a large margin compared to single-input HDR methods [5][2] (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 input images, the accuracy can be further enhanced around 0.17 dB (row 5 vs. row 4). This accuracy improvement in 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 cellphone in contrast to current two-image version of Dol 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) .

	Raw HDR			E2E-ISP		
Method	PSNR	SSIM	runtime	PSNR	SSIM	runtime
EMVNet-lite	37.111	0.9781	0.052	36.651	0.9557	0.081
EMVNet-lite-os	36.899	0.9712	0.052	36.365	0.9506	0.081
EMVNet	37.377	0.9824	1.513	36.891	0.9612	2.892

(including those in the stacked hourglass) are trimmed to half. We further add a pixel unshuffling layer at the beginning to downsample the feature map x2, and a pixel shuffling layer just before the output layer to upsample the feature map x2. The number of RRDB in the feature extraction is reduced to 6 for both HDR and E2E-ISP.

We fine-tune the network with the usage of knowledge distillation of GAN learning [1]. We follow a step-to-step way during the fine-tuning. Starting from the original EMVNet:

- Step 1: Reduce the number of RRDB and the number of filters in the feature extraction, but keep other parts of EMVNet unchanged and inherit the weights from the original MVNet, fine-tune the network w/o any weights frozen.

- Step 2: Replace the standard convolutional layers by depthwise convolutional layers in the feature extraction of the output model of Step 1, but keep other parts unchanged and inherit the weights, fine-tune the network w/o any weights frozen.

- Step 3: Trim the convolutional layers in the aggregation part of the output model of Step 2, but keep other parts unchanged and inherit the weights, fine-tune the network w/o any weights frozen.

- Step 4: Add the pixel-unshuffling layer and pixel-shuffling layer at the output model of Step 3, fine-tune the whole network to get the final EMVNet-lite model.

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

[1] Angeline Aguinaldo, Ping-Yeh Chiang, Alex Gain, Ameya Patil, Kolten Pearson, and Soheil Feizi. Compressing gans using knowledge distillation. *arXiv preprint arXiv:1902.00159*, 2019.

[2] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 354–363, 2021.

[3] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

[4] Byung-Hoon Kim, Joonyoung Song, Jong Chul Ye, and Jae-Hyun Baek. Pynet-ca: enhanced pynet with channel attention for end-to-end mobile image signal processing. In *ECCV*. Springer, 2020.

[5] 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 *CVPR*, 2020.

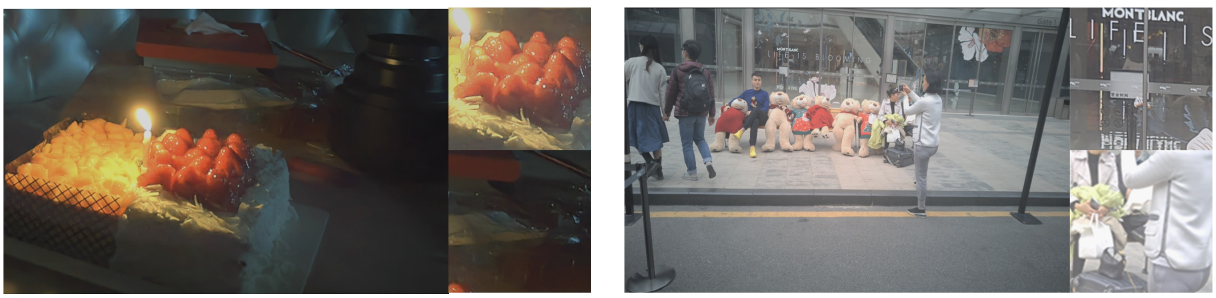
[6] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *CVPR*, 2021.

[7] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

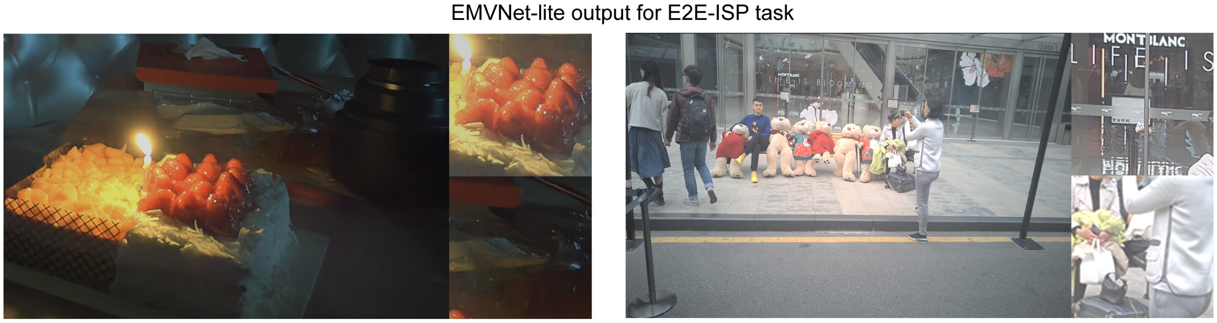
[8] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Aleš Leonardis, Radu Timofte, Zexin Zhang, Cen Liu, Yunbo Peng, Yue Lin, Gaocheng Yu, et al. Ntire 2022 challenge on high dynamic range imaging: Methods and results. In *CVPR*, 2022.

[9] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV workshops*, 2018.

[10] 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. In *CVPR*, 2019.



EMVNet-lite output for E2E-ISP task



EMVNet output for E2E-ISP task

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