Robust Real-world Image Enhancement Based on Multi-Exposure LDR Images

Haoyu Ren Yi Fan Stephen Huang
Oppo Mobile Telecommunications Corp.
3570 Carmel Mountain Road, San Diego, CA, US

Abstract

Robust real-world image enhancement from multi-exposure low dynamic range (LDR) images is a challenging task due to the unexpected inconsistency among the input images, such as the large motion or various exposures. In this paper, we propose a novel end-to-end image enhancement network to solve this problem. After extracting contextual information from the LDR images, we design a novel matching volume to align them by considering the motion and exposure differences among the input images. A stacked hourglass with dilated convolution is further utilized to aggregate the matched feature maps to the final enhanced image. In addition, we design a weakly-supervised pairwise loss function to evaluate the color consistency in the enhanced image, which further boosts the performance. We show the effectiveness of our methods on high dynamic ranging imaging (HDR) and End-to-End image signal processing (E2E-ISP) tasks. Experimental results demonstrate that our model achieves state-of-the-art enhancement performance.

1. Introduction

Image enhancement from low dynamic range (LDR) images with multi-exposure is a practical and challenging problem. Most of the cameras only produce photos with a limited dynamic range. It is necessary to generate the images with high dynamic range (HDR) to satisfy the human vision. With the wide usage of digital overlap (DOL) camera in cellphones, this problem becomes one of the essential functions in the image signal processing (ISP) pipeline. To solve this problem, traditional methods select a reference image and use the rest images to compensate the missing details caused by over or under-exposure. Recently, deep convolutional neural networks (CNNs) have been deployed as they demonstrate significant accuracy improvements.

There are two major challenges of this task. The first challenge is the motion of foreground objects. The prior arts refer to some alignment procedure such as optical flow based image warping. Unfortunately, in the poor lighting scenes such as low-light indoor or night scenes, it is very difficult to calculate an accurate motion. This results in some error in the enhanced image, as shown in the bottom row of Fig. 1. The second challenge is the various exposures among the input images, which makes the algorithm very difficult to compensate the missing content. The prior solutions focus on a limited exposure such as +2/+4. If the network is applied on a totally different exposure (e.g., $> +8$, which is used in real-world night capturing), the results will not be good, as given in the top row of Fig. 1, where the information of the over-exposed area is missed in the HDR output.

Figure 1. Challenges of real-world image enhancement with multi-exposure images. Top - HDR when input images have large exposure difference. Bottom - E2E-ISP when input images have poor lighting condition. Our method handles these scenarios better compared to prior arts.

In this paper, we propose an image enhancement network EMVNet with a novel matching volume (MV). After ex-

tracting features from the input images, a matching volume is applied on the extracted feature maps to check the consistency among the input images. During this procedure, the motion and exposure difference will be measured, with the guidance of an over-exposed area mask from the input images. As a result, the network can compensate motion and understand the exposure difference from the inputs. Then a stacked hourglass with dilated convolution is utilized to aggregate the output feature maps to generate the final enhanced image. We train our EMVNet with a novel pairwise weakly-supervised loss to improve the color consistency in the enhancement output. This further improves the accuracy and the robustness. We utilize the Generative Adversarial Network (GAN) with relativistic discriminator in our end-to-end training. We demonstrate the effectiveness of our method by two tasks, the HDR imaging, and the End-to-End image signal processing (E2E-ISP) which uses single neural network to replace the whole mobile ISP pipeline. The results show that our method outperforms existing approaches with more consistent color and better details in both tasks.

The contributions of this paper are highlighted as follows

- Our proposed matching volume can handle the motion and exposure difference at the same time, which allows us to do 'blind image enhancement' without knowing the exposure difference among the input images.
- Our proposed weakly-supervised loss function is able to boost the enhancement accuracy, which produces images with more reliable color.
- Our proposed EMVNet is flexible and generalized well on various image enhancement applications, such as HDR with two or three input images in both RGB and raw domains, or E2E-ISP where the inputs are raw LDR images, and the output is demosaiced RGB image. To our knowledge, this is the first end-to-end network for mobile ISP with multi-exposure inputs.

2. Related work

2.1. HDR

Traditional HDR algorithms aim to merge several LDR images captured from multiple exposures [4][5]. These methods choose one of the input LDR image as the reference image and align the rest images with the reference one. The missing information of the reference image will be compensated by fusion with hand-crafted features, which limit the accuracy and robustness. Recently, deep learning is widely used. Some researchers focus on reconstructing HDR image from a single LDR image. Eilertsen et al. [3] utilized a U-shape network and gathered a large dataset while simulating sensor saturation for a range of cameras. Lee et al. [12] created HDR images based on the estimated multi-exposure stack using the conditional generative adversarial network structure. Liu et al. [15] modeled the HDR-to-LDR image formation pipeline as the (1) dynamic range clipping, (2) non-linear mapping from a camera response function, and (3) quantization. They proposed to learn three specialized CNNs to reverse these steps.

Single image HDR reconstruction methods do not perform well in the wild due to unexpected illumination and motion. HDR with multiple exposure images is more practical. Recent methods consider HDR reconstruction as an image translation problem from the LDR domain to the HDR domain. Wu et al. [23] estimated the homography transformation and utilized a translation network to hallucinate plausible HDR details in the presence of total occlusion, saturation and under-exposure. Yan et al. [24] proposed to use attention modules to guide the merging according to the reference image. Yan et al. [25] fused all inputs and map the fusion results into a low-dimensional deep feature space and then fed the resultant features into a global non-local module which reconstructs each pixel by weighted averaging all the other pixels. Niu et al. [18] proposed HDR-GAN, with a reference-based residual merging block for aligning large object motions in the feature domain, and a deep HDR supervision scheme for eliminating artifacts of the reconstructed HDR images. Liu et al. [16] presented an attention-guided deformable convolutional network ADNet. They adopted a spatial attention module to adaptively select the most appropriate regions of LDR images for fusion. Huang et al. [7] combined the neuron random field (NERF) with the HDR problem, and utilized the classic volume rendering technique to project the output radiance, colors, and densities into HDR and LDR images.

The above methods propose various solutions to deal with the mis-alignment. But one major problem is that the accuracy strongly relies on the alignment module of the LDR images. If the exposure difference is very large (e.g., in low-light night scenes), existing methods either cannot guarantee to align the LDR images or fail to produce adequately faithful information for the missing image contents. In contrast, our EMVNet considers the motion and exposure difference at the same time, which can generate more reliable information used in the following aggregation procedure. Our method is more robust in real-world scenarios.

2.2. End-to-end ISP

Recently, more and more researchers start working on End-to-End image signal processing (E2E-ISP) [8], where a single neural network is adopted to convert the input raw image into RGB image. Ignatov et al. [9] proposed PyNet for such Raw-to-RGB reconstruction by using a U-shape network with multi-scale encoder-decoder architecture. This method is improved by Kim et al. [11] with the usage of additional channel attention block to improve the performance and re-

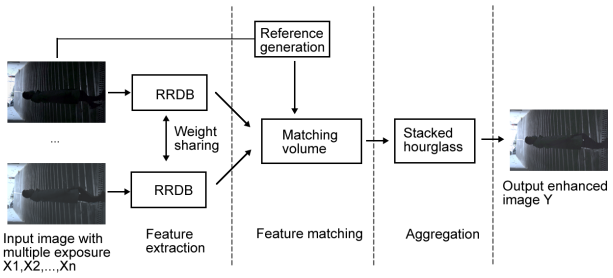

Figure 2. Overall framework of our EMVNet. It consists of three parts, feature extraction, feature matching, and aggregation. A side path 'reference generation' from the input images is utilized to assist the feature matching.

duce the training time. Although these methods work well in a few standard scenarios such as daytime images with good illumination, their performance on low-light scene is limited. In addition, these methods are designed for single input image. In a practical scenario where mobile ISP with dol camera produces images with multiple exposures, these methods don't work well since they cannot handle the motion and exposure mis-alignment. To our best knowledge, our method is the first one who proposes an end-to-end network to handle the ISP with multi-exposure input images.

3. Our methods

Our overall framework is displayed in Fig. 2. Giving n input LDR images $X_1, X_2, ..., X_n$ with different exposures, our EMVNet outputs a single enhanced image Y. The feature extraction module is first applied on each of the input image to extract the key information. Then the matching volume is applied to check the consistency between these features and handle the variations of motion and exposure. The aggregation module will aggregate the output feature maps of the matching volume and convert them into the desired enhancement output.

3.1. Feature extraction

We use the residual-in-residual dense block (RRDB) [22] as the basic unit of the feature extraction. Each dense block consists of 5 convolutional layers. The first 4 convolutional layers are followed by a Relu activation, and there is no Relu activation after the 5th convolutional layer. Dense connections are added between these 5 convolutional layers. The dense block is further inserted into the residual-in-residual architecture to construct the RRDB. We organize several RRDB blocks into a sequential order as the final feature extractor. Same feature extractor will be applied on each of the input image $X_1, X_2, ..., X_n$ to generate the feature maps $F_1, F_2, ..., F_n$.

3.2. Matching volume

In some computer vision tasks with multiple input images such as disparity estimation [2] or optical flow estimation [21], a commonly-used way is designing a cost volume to match the feature maps generated by each of the input image. Inspired by this architecture, we propose a matching volume (MV) for the image enhancement task. Our matching volume takes the deep features extracted by the RRDB from the n inputs images $F_1, F_2, ..., F_n$ as input, and outputs a single feature map F_{MV}. The key idea of MV can be described as 'given a specific motion vector and a specific exposure difference, what is the correlation between the reference image and other input images'. The design of MV is given in Fig. 3. Let $M \in \{M_1, ..., M_{N_m}\}$ be a specific motion vector, and $E \in \{E_1, ..., E_{N_e}\}$ be a specific exposure difference compared to the reference frame I_{ref}, where N_M and N_E are the number of motion and exposure difference we consider in the MV. We first align the feature maps $F_1, ..., F_n$ with motion E and exposure M by

$$F'_{i,M,E} = \begin{cases} C(F_i, M, E) & 1 \leq i \leq n, i \neq ref. \\ F_i & i = ref \end{cases} \quad (1)$$

where C is the alignment function. We compare these aligned feature maps $F'_{i,M,E}, i = 1, ..., n$ with the feature map of the reference frame F_{ref} to generate a correlation $O_{i,M,E}, i = 1, ..., n$. Such 'comparison' is a concatenation of the features generated by the following three operations:

- Feature concatenation: $\{F'_{i,M,E}, F_{ref}\}$
- Feature difference: $|F'_{i,M,E} - F_{ref}|$
- Cross correlation: $< F'_{i,M,E}, F_{ref} >$

The final output of MV F_{MV} is a concatenation of these correlations $O_{i,M,E}, i = 1, ..., n$. Assume the output feature map size of the feature extraction module is $F_i \in R^{C \times H \times W}, i = 1, ..., n$. the size of the corresponding correlation $O_{i,M,E}$ will be $O_{i,M,E} \in R^{4C \times H \times W}$. Since there will be N_M different motion and N_E different exposures evaluated in the matching volume, with n input images, the final size of the output feature map of the matching volume will be $F_{MV} \in R^{4C \times (N_M \times N_E \times n) \times H \times W}$.

The alignment function $C(F_i, M, E)$ consists of two steps. The first step is warping the feature map F_i with the motion M. The second step is multiplying the exposure E onto the warped feature maps. In the ideal scenario, if there is no over-exposed area, $E * W(F_i)$ would be exactly the same as F_{ref}, where W is the warping operator [1]. Unfortunately, in most of the real-world LDR images, there will always be some over-exposed areas. So we add a side path from the input images to extract the following information:

[1] The sensor noise and lens shading are ignored in this assumption

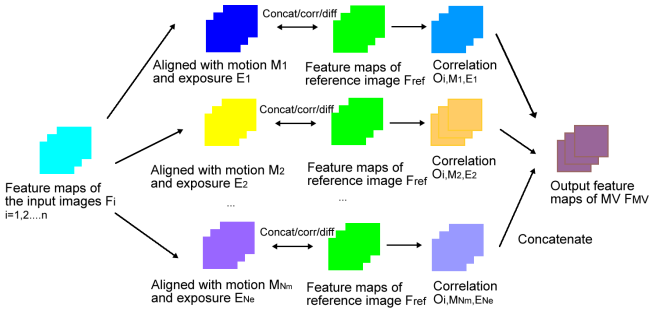

Figure 3. Our proposed matching volume (MV).

- Which of the input image is better to be used as the reference image I_{ref}. This can be achieved either by some prior knowledge or blind image evaluation method [13].

- The over-exposed area mask. It can be generated by thresholding each channel of the input images. This mask will be utilized as a weight mask in the alignment function $C(F_i, M, E)$. When multiplying the exposure E onto the feature maps, if a pixel is saturated in the input image, the corresponding multiplication will be thresholded as well.

There are two major differences of our matching volume compared to existing cost volume used in disparity estimation or optical flow estimation. First, existing cost volume considers the motion only, while our MV considers the exposure differences among the input images as well. Second, our MV allows multiple input images, which is more flexible than the fixed 2 input images in the cost volume.

3.3. Cost aggregation

To generate the enhanced image , we consider aggregating multi-scale contextual information from the output of the matching volume F_{MV}. We adopt a stacked hourglass architecture, where 3 hourglasses are stacked in a sequential order, as given in Fig. 4. Each hourglass consists of 6 layers, the first three layers are 3D convolutional layers with stride 2, and the following three layers are 3D deconvolution layers with scale factor 2. Since image enhancement requires global contextual information, we use dilated convolution instead of standard convolution to further increase the receptive field. The dilation factors increase as the hourglass goes deeper. We extract the intermediate outputs Y'', Y' from the first two hourglasses. These two outputs are used during the training. During the testing, only the final output Y is utilized.

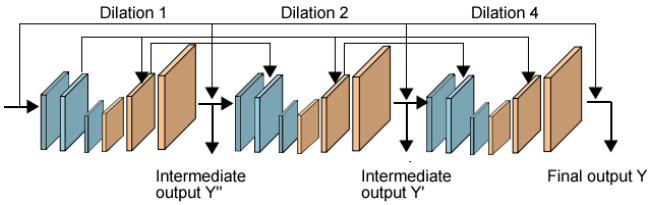

Figure 4. Stacked hourglass for aggregation. All layers have 3D convolution kernels. Blue layers are convolutional layers, and yellow layers are deconvolution layers.

4. Implementation

4.1. Weakly-supervised loss function

Learning with standard image content losses such as L_1 or L_2 has some limitations. The network tries to reach the smallest differences between the output and the ground-truth, but fails to keep the intensity order between different pixels or patches. For instance, if the ground truth intensity of two pixels are [4,7], the network trained by L_1 or L_2 losses might generate enhanced output with corresponding intensities at [6,5]. The intensity order is twisted, which makes the brighter region become a darker region. For 3-channel RGB image, it might result in color distortion in some areas, especially for the low-light images in the wild.

To solve this problem, we propose a pairwise weakly-supervised loss function L_s with two different versions for pixels and patches respectively. Given a pixel pair (i, j), the pixel-wise loss function $L_{s,pix}$ is given as

$$L_{s,pix} = \begin{cases} \log(1 + e^{P_{ij}}) & if P_{ij} \leq S_{pix} \\ \log(1 + e^{\sqrt{P_{ij}}}) + C_{pix} & if P_{ij} > Spix \end{cases} \quad (2)$$

where $P_{ij} = -r_{ij}(\log(I_i) - \log(I_j))$, I_i is the intensity of pixel i in the enhanced image, and r_{ij} is an ordinal intensity indicator, $r_{ij} = 1$ if pixel i is brighter than pixel j in the ground-truth image, otherwise $r_{ij} = -1$. S_{pix} is a threshold setting as 0.25 based on the tuning results, C_{pix} is a constant to make the loss functions continuous. $L_{s,pix}$ encourages the corresponding pixel pair in the output image have same intensity order as the ground-truth image.

In real-world scenario, the LDR images suffer from the unexpected noise. We further propose the patch-wise loss function based on a patch pair $\{x, y, r_{xy}\}$, where (x, y) are rectangle-shape patches, and r_{xy} is an ordinal indicator. x and y have exact same size. $r_{xy} = 1/-1$ depending on whether the average intensity of x is larger than y, which is similar to the pixel-wise version. We define the patch-wise loss function $L_{s,pat}$ as Eq. 3. We set $S_{pat} = 0.5$ based on the tuning results, and C_{pat} is also changed accordingly.

$$L_{s,pat} = \begin{cases} \log(1 + e^{P_{ij}}) & if P_{ij} \leq S_{pat} \\ \log(1 + e^{\sqrt{P_{ij}}}) + C_{pat} & if P_{ij} > S_{pat} \end{cases} \quad (3)$$

The size of the patch pair varies from 5×5 to 15×15. The patch pair and pixel pair are sampled randomly from the non-over-exposed areas:

- Given a pixel pair (i, j), $I_i^*, I_j^* \leq 0.99 * I_{max}$, I_i^* is the intensity of pixel i in the ground-truth image
- Given a patch pair (x, y), for all pixels $i \in x, j \in y$, $I_i^*, I_j^* \leq 0.99 * I_{max}$, I_i^* is the intensity of pixel i in the ground-truth image

4.2. GAN based learning

We follow the ESRGAN [22] framework with the usage of relativistic discriminator. We use the EMVNet described in Section 3 as the generator. During the training, our 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. 4. The hyper-parameters λ, η, α determine the contribution of different components in the final loss function.

$$L_G = L_p + \lambda L_a + \eta L_c + \alpha L_s, \tag{4}$$

The image content loss is based on standard L_1 loss. In Section 3.3, we mentioned that the stacked hourglass will output three enhanced images Y'', Y', Y. The image content loss L_c is formulated as

$$L_c = L_1(Y, Y^*) + 0.5 * L_1(Y'', Y^*) + 0.75 * L_1(Y', Y^*) \tag{5}$$

where Y^* is the ground-truth. The perceptual loss L_p calculates a feature map distance between the final output Y and Y^* with the usage of a pre-trained 19-layer VGG network. Considering the EMVNet output Y as 'fake image', and the ground-truth Y^* as the 'real image', the adversarial loss L_a is based on the relativistic GAN discriminator [22] and is defined as

$$L_a = -\mathbb{E}_{Y^*}[\log(1 - D(Y^*, Y)] - \mathbb{E}_Y[\log(D(Y^*, Y)]. \tag{6}$$

The weakly-supervised loss L_s can either be $L_{s,pix}$ or $L_{s,pat}$ alone, or a combination of both to achieve the best accuracy. In each iteration, we randomly generate a pixel pair or patch pair, and calculate the weakly-supervised loss given in Eq. 3 or Eq. 2.

4.3. Implementation for HDR

For HDR, 12 RRDB blocks are concatenated as the feature extraction module. All convolutional layers in RRDB have 32 3×3 convolutional filters. In the matching volume, we consider 6 different exposures $E \in \{-4, -2, +2, +4, +8, +16\}$[2] with $N_E = 6$, and the motion vector ranges from $\{0, 0\}$ to $\{36, 36\}$. Since it would be time consuming to cover all candidates in such a large motion range, we sample the motion every 3 pixels to make the $N_M = 36 \times 36/3/3 = 144$. Ablation study in section 5.4.1 shows that this will not reduce the accuracy much. The aggregation module consists of three hourglasses given in Fig. 4, where all the convolutional layers have same 32 $3 \times 3 \times 3$ 3D convolutional filters, but with different dilation factors, and different strides for downsampling or upsampling.

We utilize VGG-19 as the discriminator. Since VGG-19 is pre-trained on RGB domain, for raw HDR task, we add a simple demosaicing module before feeding the output HDR raw image into the discriminator. The pixel pairs and patch pairs for the weakly-supervised learning are randomly sampled from each of the channel (e.g.,R/G/B for RGB HDR, or R/G/G/B for raw HDR). We start the training with a learning rate 0.0001 and decrease it by 0.5 every 200K iterations with a batch size 8. The weights of the loss functions are set as $\lambda = 0.001, \alpha = 0.25, \eta = 0.001$. Ablation study of these hyper-parameters are given in Section 3.2 of the supplementary material.

4.4. E2E-ISP

For E2E-ISP EMVNet, we use 16 RRDB blocks for feature extraction. E2E-ISP's aggregation is more complicated since it needs considering the demosaicing. So we set the hourglasses in the aggregation module as 64 filters for each of the convolutional layers, and change the channel of output layer from 4 to 3. Additional sub-pixel convolutional layers are added before each of the output enhanced image Y'', Y', Y respectively. The pixel pairs and patch pairs for the weakly supervised learning are randomly sampled from each of the R/G/B channel. The training hyper-parameters are mostly the same as the HDR training. But the weights of the loss functions are set as $\lambda = 0.005, \alpha = 0.6, \eta = 0.001$.

5. Experiments

5.1. Datasets

5.1.1 HDR

First, we utilize the commonly-used Kalantari's dataset [10] for the task of HDR in RGB domain. This dataset contains 74 image sets for training and 15 image sets for testing. For each training image set, three different LDR images are captured with exposure biases $\{-2, 0, 2\}$ or $\{-3, 0, 3\}$ in TIFF format. We also give the results on another RGB-HDR dataset, the NTIRE 2022 HDR dataset [19] in Section 1 of the supplementary material.

[2] We don't consider exposure -8 and -16 since these two ratios mainly occur in the night scenes. In such scenarios, the short exposure image will be selected as reference image because it has less motion blur.

Next, we check the performance of EMVNet on raw HDR task, which is more practical on mobile devices. For training, we generate synthetic data based on Google HDR+ dataset [6]. This dataset consists of 3,640 bursts with raw burst inputs in DNG format, as well as the merged results for each of the burst. Since all photos in a burst are generally captured with the same exposure time, we generate the synthetic LDR inputs based on their raw burst inputs following [14], and use the merged result as the HDR ground-truth. To simulate the real-world HDR with dol camera, we use two inputs and randomly sample the exposure differences between +2 and +16. We generate the long/short exposure images from different frames of each burst to simulate the motion. 2,000 long/short exposure image pairs are used as our training set, and another 400 image pairs are used as the validation set.

For testing, we collect a real-world mobile raw image test set named RWMR dataset, where the image sequences are captured by OPPO Reno 5 pro+ cellphone in dol mode. The captured raw images cover a variety of illumination levels we would see in our daily life, including indoor, outdoor, daytime and night scenes. 120 sequences are collected, and each sequence consists of 20 to 30 frames.

5.1.2 E2E-ISP

For E2E-ISP, we use HDR+ dataset and same training/testing split as the raw HDR task. For each of the merged burst, HDR+ dataset also provides the high-quality JPEG image processed by Google's ISP. So we train our EMVNet using the synthetic long/short exposure raw image pairs as the input, and use the high-quality JPEG image as ground-truth. Since some of JPEG images are not aligned with the merged burst (e.g., rotated or scaled), we manually remove these images from the 2,000 training images. Our final training set consists of 1,740 long/short raw-JPEG image pairs, and the validation set consists of 317 pairs. Although the HDR+ dataset provides the lens shading map, we don't use them because we are trying to learn an end-to-end model which is expected to handle the lens shading implicitly.

5.1.3 Evaluation metric

Similar to prior arts in image enhancement, we use the conventional fidelity-based Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity index (SSIM) for quantitative feedback. We also perform a user study on the final output images with the Mean Opinion Score (MOS) in the following manner. 16 test candidates were shown a side-by-side comparison of a sample prediction of a certain method and the corresponding reference ground-truth. They were then asked to evaluate the quality of the output image w.r.t. the reference image using the 5-level scale defined as: 0 - 'Perfect', 1 - 'Almost Perfect', 2 - 'Slightly Worse', 3 - 'Worse', 4 - 'Terrible'. The images shown to the participants of the study were composed of zoomed crops. The human study was performed for the top 4 methods of each task according to PSNR ranking.

Table 1. Experimental results on Kalantari's dataset [10]. Bold font indicates the best over the columns. For MOS, the smaller the better. For other metrics, the larger the better.

Method	PSNR/SSIM$_\mu$	PSNR/SSIM$_l$	HV2	MOS
Sen [20]	40.80/0.9808	38.11/ 0.9721	59.38	-
Kalantari [10]	42.67/0.9888	41.23/0.9846	65.05	-
Wu [23]	41.65/0.9860	40.88/0.9858	64.90	-
Yan [24]	43.67/0.9900	41.14/0.9702	64.61	1.80
Niu [18]	43.92/0.9905	41.57/0.9865	65.45	1.81
Liu [16]	44.37/0.9917	41.88/0.9892	66.02	1.72
Our EMVNet	**44.63/0.9932**	**42.12/0.9910**	**66.16**	**1.63**

Figure 5. Example outputs on Kalantari's dataset [10].

5.2. Experimental results of HDR

Table 1 gives the experimental results of our network, compared to prior arts when training/testing on Kalantari's dataset [10]. The subscript of μ indicates that the methods are calculated in the tone mapped domain following the μ-law, and the subscript l denotes that the PSNR/SSIM are calculated in the linear HDR domain. The HDR-VDP-2 (HV2) [17] assesses the visibility and quality of the HDR images in different luminance conditions. During the HDR-VDP-2's calculation, we set the diagonal display size to 24 inches, and the viewing distance to 0.5 meter. We can see that our method outperforms all other methods, on all quantitative and qualitative evaluation metrics. Better MOS score indicates that EMVNet generates more perceptually friendly images in human vision. Our EMVNet is capable of recovering better details from the LDR images, as shown in Fig. 5.

Next, we use the HDR+ dataset to check the performance on the raw HDR task. We train our EMVNet, as well as prior arts [24][18][16] on the same HDR+ images, and re-

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. For MOS, the smaller the better.

	Raw HDR		E2E-ISP	
Method	PSNR/SSIM	MOS	PSNR/SSIM	MOS
Yan [24]	36.06/0.9586	2.28	-	-
Niu [18]	36.29/0.9645	2.08	-	-
Liu [16]	36.55/0.9690	2.13	-	-
PyNet [9]	-	-	35.28/0.9498	2.42
PyNet-CA [11]	-	-	35.35/0.9479	2.48
Ours	**37.38/0.9824**	**1.83**	**36.89/0.9612**	**2.13**

port the PSNR/SSIM/MOS on the validation images in the second and third columns of Table 2. It can be seen that our method achieves 0.8 dB higher PSNR and 0.01 higher SSIM compared to other HDR methods. The MOS score [3] of our method is significantly better than prior arts with a 0.24 gap. This proves that our EMVNet works better than prior arts on raw HDR in human vision.

Moreover, we use the images of RWMR dataset to check the raw HDR on real-world captured images. In Fig. 6, we show a few examples where all input LDR and output HDR images are demosaiced with a simple 4-way interpolation function from OpenCV, and further enhanced with gamma 2.2, otherwise the image will be too dark to visualize. It can be seen that our method is able to provide the HDR outputs with more details and less artifacts, especially on the extreme low-light scenarios where the exposure differences between the inputs are large. In contrast, the prior arts bring some unpleasant artifacts into the output HDR images.

5.3. Experimental results on E2E-ISP

We first give the accuracy comparison of E2E-ISP on the validation images of the HDR+ dataset in the fourth and fifth columns of Table 2. All methods are re-trained on same training set. The input layers of the prior arts [9][11] are modified to accept multi-input images. We find that our method outperforms the state-of-the-art E2E-ISP methods PyNet and PyNet-CA with a large margin in all metrics. The reason is that these two networks are designed for single input image, so that they don't have specific consideration for the motion and exposure differences caused by multi-inputs. Our network benefits from the proposed matching volume, which can generate more reliable color.

We further test these methods on the RWMR raw images captured by dol camera, and give the visualizations of some output RGB images in Fig. 7. It is notable that our results show higher contrast, more neutral color, and better details

[3] In consideration of the workload of human evaluation, we randomly select 50 images from the 400 validation images. Different methods use the same set of 50 images.

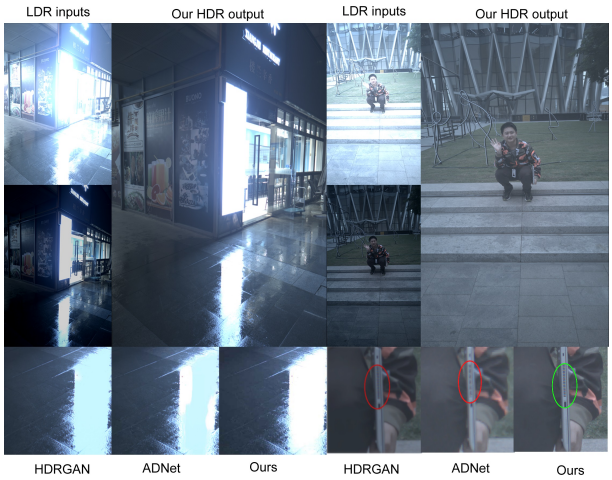

Figure 6. Example outputs of HDR based on raw images of RWMR dataset captured by dol camera. All images are visualized by OpenCV demosaicing with additional gamma 2.2 enhanced.

compared to the prior arts. There is no motion ghost (see the person hand of the right image) as well. This demonstrates the potential of applying our method on real-world scenarios. More example outputs and analysis can be found in Section 2 of our supplementary document.

5.4. Ablation study

5.4.1 Different designs of matching volume

Here we give the ablation study of of using different MVs by the raw HDR task on HDR+ dataset. In Table 3, it can be seen the accuracy of all the DNNs with matching volume outperform the one w/o matching volume (row 2). W/o considering either the motion or the exposure (row 3-4) in the matching volume, the PSNR/SSIM drop significantly. This makes sense since the motion and exposure are the key misalignment among the input images. This result is consistent to [27], which shows that considering color and motion differences at the same time for can improve the enhancement quality for image impainting . We also notice that if the MV is not guided by the over-exposed area mask (row 5), the accuracy also decreases. This reflects that adding penalties on the pixels corresponding to the over-exposed area can give the network more insights. We also provide the results of using a MV where the motion is not downsampled, labels as 'yes(dense)' in the 6th row, compared to our current implementation where the motion vector is x3 sampled (row 7). We observe that with the usage of dense motion, the PSNR/SSIM is slightly improved. But since the computational cost is increased a lot, we still stick to the current version with sub-sampled motion.

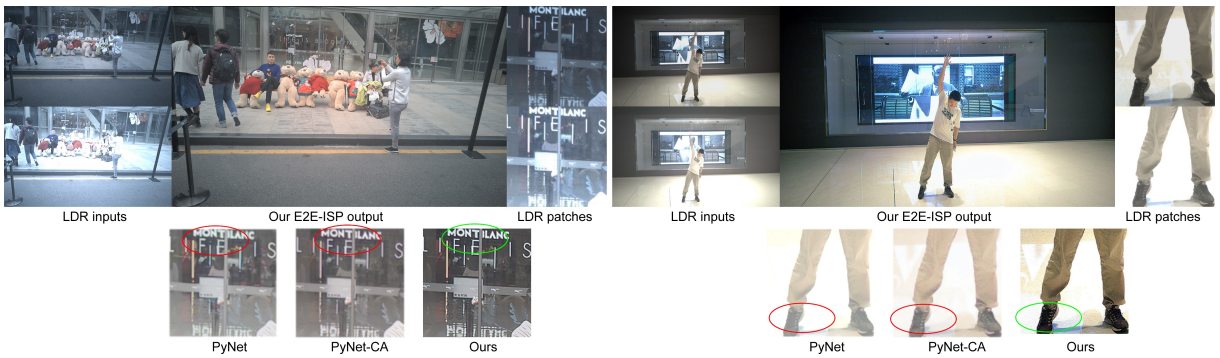

Figure 7. Example outputs of E2E-ISP task on raw images of RWMR dataset captured by dol camera.

Table 3. Raw-HDR accuracy evaluation on HDR+ validation images with different matching volumes. 'mask' means that whether the MV is guided by the over-exposed area mask. Bold font indicates the best over the columns.

	Motion	Exposure	Mask	PSNR	SSIM
EMVNet	no	no	no	36.162	0.9613
EMVNet	no	yes	yes	36.819	0.9708
EMVNet	yes	no	yes	37.064	0.9730
EMVNet	yes	yes	no	37.190	0.9798
EMVNet	yes(dense)	yes	yes	**37.414**	**0.9839**
EMVNet	yes	yes	yes	37.377	0.9824

Table 4. Accuracy evaluation on HDR+ validation images with different loss functions. L_c stands for image content loss, L_p stands for perception loss, L_a is the GAN's adversarial loss, $L_{s,pix}$ stands for pixel-wise weakly-supervised loss, $L_{s,pat}$ stands for the patch-wise weakly-supervised loss. Bold font indicates the best over the columns.

		Raw HDR		E2E-ISP	
	loss	PSNR	SSIM	PSNR	SSIM
EMVNet	L_c	36.966	0.9690	36.395	0.9523
EMVNet	$L_c L_p$	36.999	0.9735	36.484	0.9552
EMVNet	$L_c L_p L_a$	37.021	0.9732	36.491	0.9566
EMVNet	$L_c L_p L_a L_{s,pix}$	37.161	0.9796	36.639	0.9599
EMVNet	$L_c L_p L_a L_{s,pat}$	37.276	0.9789	36.800	0.9604
EMVNet	$L_c L_p L_a L_{s,pix} L_{s,pat}$	**37.377**	**0.9824**	**36.891**	**0.9612**

5.4.2 Different loss functions

Next, we evaluate the accuracy of EMVNet with different loss functions. We use the same EMVNet with both the motion and exposure ratio enabled in the matching volume, but train them with different loss functions. As given in Table 4, comparing row 5 versus row 3 and row 4, we notice that w/o using perceptive loss (L_p) or GAN's learning (L_a), the accuracy drops slightly ($<$ 0.06 dB). If we train with either pixel-wise weakly-supervised loss ($L_{s,pix}$) or patch-wise weakly-supervised loss ($L_{s,pat}$), the accuracy improves up to 0.3 dB, as given in row 6 and row 7. Training with both of these two loss functions (as mentioned in Section 4.2, randomly generate a pixel or patch pair in each iteration) gives us the best accuracy, given in the last row. We give some example outputs in Section 3.1 of the supplementary material to show the effectiveness of training with the proposed loss function.

5.4.3 Computational cost

We evaluate our method using 4K resolution raw images (12M pixels), which is the typical application scenario in mobile devices. On single A100 GPU, our EMVNet takes 1.51 seconds for HDR task, and 2.89 seconds for E2E-ISP. In contrast, the state-of-the-art HDR methods ADNet [16] takes 8.77 seconds due to the huge spatial attention matrix. The state-of-the-art E2E-ISP methods PyNet [9] is slightly slower than our method, which takes 3.22 seconds.

To further improve the efficiency, we reduce the number of RRDBs to 6, and replace the standard convolutional layers by depth-wise convolutional layers in the feature extraction module. The number of the filters in all convolutional layers (including those in the stacked hourglass) are reduced to half. We add a pixel unshuffling layer at the beginning to downsample the feature map x2, and a pixel shuffling layer at the end to retrieve the resolution. The whole network is fine-tuned by knowledge-distillation [1] in a step-to-step way, while using the original EMVNet as the teacher network. This can accelerate the network x30 (0.08 second on 4K resolution image), with a 0.26 dB accuracy lost. Such loss is not very significant in human vision. Details can be found in Section 3.4 of the supplementary material.

6. Conclusion

In this paper, we proposed an effective framework for image enhancement with inputs of different exposures. Our proposed EMVNet utilized the matching volume to measure the variations among different input images. The motion and exposure differences will be evaluated, and further aggregated by the stacked hourglass with dilated convolutions. Along with the usage of weakly-supervised learning, we are able to retrieve the missing information while keeping the confident color information. Our network works well for multiple image enhancement tasks, including HDR, and end-to-end ISP. Experimental results on real-device captured data show the effectiveness of our method.

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] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *arXiv preprint arXiv:2010.13501*, 2020.

[3] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017.

[4] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. Optimal hdr reconstruction with linear digital cameras. In *CVPR*, 2010.

[5] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *CVPR*, 2010.

[6] 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 (ToG)*, 35(6):1–12, 2016.

[7] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *CVPR*, 2022.

[8] Andrey Ignatov, Cheng-Ming Chiang, Hsien-Kai Kuo, Anastasia Sycheva, and Radu Timofte. Learned smartphone isp on mobile npus with deep learning, mobile ai 2021 challenge: Report. In *CVPR*, 2021.

[9] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *CVPR Workshops*, 2020.

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

[11] 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.

[12] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *ECCV*, 2018.

[13] Xin Li. Blind image quality assessment. In *ICIP*, volume 1, 2002.

[14] Yuanzhen Li, Lavanya Sharan, and Edward H Adelson. Compressing and companding high dynamic range images with subband architectures. *ACM transactions on graphics (TOG)*, 24(3):836–844, 2005.

[15] 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.

[16] 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.

[17] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

[18] 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.

[19] 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.

[20] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012.

[21] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*. Springer, 2020.

[22] 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.

[23] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*, 2018.

[24] 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.

[25] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020.

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[27] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2266–2276, 2021.