UPHDR-GAN: Generative Adversarial Network for High Dynamic Range Imaging With Unpaired Data

Ru Li, *Student Member, IEEE*, Chuan Wang, Jue Wang, *Senior Member, IEEE*, Guanghui Liu, *Senior Member, IEEE*, Heng-Yu Zhang, Bing Zeng, *Fellow, IEEE*, and Shuaicheng Liu, *Member, IEEE*

***Abstract*— The paper proposes a method to effectively fuse multi-exposure inputs and generate high-quality high dynamic range (HDR) images with unpaired datasets. Deep learning-based HDR image generation methods rely heavily on paired datasets. The ground truth images play a leading role in generating reasonable HDR images. Datasets without ground truth are hard to be applied to train deep neural networks. Recently, Generative Adversarial Networks (GAN) have demonstrated their potentials of translating images from source domain *X* to target domain *Y* in the absence of paired examples. In this paper, we propose a GAN-based network for solving such problems while generating enjoyable HDR results, named UPHDR-GAN. The proposed method relaxes the constraint of the paired dataset and learns the mapping from the LDR domain to the HDR domain. Although the pair data are missing, UPHDR-GAN can properly handle the ghosting artifacts caused by moving objects or misalignments with the help of the modified GAN loss, the improved discriminator network and the useful initialization phase. The proposed method preserves the details of important regions and improves the total image perceptual quality. Qualitative and quantitative comparisons against the representative methods demonstrate the superiority of the proposed UPHDR-GAN.**

***Index Terms*— Multi-exposure HDR imaging, generative adversarial network, unpaired data.**

I. Introduction

THE dynamic range of commercial imaging products is lower than natural scenes. Most digital photography sensors cannot acquire the irradiance range that is wide enough. High dynamic range (HDR) imaging techniques have been introduced because they can overcome such limitations and generate images with a wider dynamic range. The specialized hardware device [1] has been introduced to directly obtain HDR images, but it is usually too expensive to be widely adopted. An optional strategy is to merge a stack of images with different exposures to produce an informative output [2].

Since its first introduction in 1990s, HDR imaging techniques evolve quickly, whose applications include saliency detection [9] and video compression [10]. Some HDR imaging methods are first proposed to generate the results through two steps: (1) reconstructing an HDR image; (2) applying the tone mapping algorithms for display [11]. These methods are not suitable for handling dynamic scenes because they do not consider the misalignments between different input images. Subsequently, Oh *et al.* proposed a rank minimization algorithm to detect outliers for HDR generation and align input images [12]. Szpak *et al.* introduced the Sampson distance to estimate the homography matrix and applied the homography to align input images [13]. These methods work well when the inputs are aligned properly. However, completely aligning the multi-exposure images is challenging. The aforementioned methods may produce ghosting or blurring artifacts if the alignment process fails to work. To alleviate the problem, some patch-based methods are proposed to generate fully registered image stacks. Sen *et al.* considered the HDR reconstruction as an optimization that includes the alignment and reconstruction [4]. Hu *et al.* built new image stacks using a variant of PatchMatch to handle saturated regions and avoid the ghosting artifacts [3]. However, the patch-based methods lack robustness and cannot produce satisfactory results for complicated scenes.

Inspired by the convolutional neural network (CNN), some learning-based methods are introduced to imitate the fusion process. Kalantari *et al.* [5] and Wu *et al.* [6] adopted similar network architecture but different in the pre-processing. Kalantari *et al.* [5] applied the flow-based pre-processing to align the inputs, while Wu *et al.* [6] embedded the alignment process into the network. Yan *et al.* [14] and Liu *et al.* [15] proposed the attention-guided network to tackle the misalignment and handle the saturation simultaneously. However, due to the unreliability of the image registration, these methods also suffer from unavoidable artifacts. There are also some GAN-based methods that introduce the adversarial loss to improve the unsatisfactory regions by creating realistic

Manuscript received 24 December 2021; revised 16 March 2022, 31 May 2022, and 22 June 2022; accepted 30 June 2022. Date of publication 12 July 2022; date of current version 28 October 2022. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62071097, Grant 61872067, Grant 62031009, and Grant 61720106004. This article was recommended by Associate Editor S. Asif. *(Corresponding authors: Guanghui Liu; Heng-Yu Zhang.)*

Ru Li, Guanghui Liu, Bing Zeng, and Shuaicheng Liu are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: guanghuiliu@uestc.edu.cn).

Chuan Wang is with Megvii Technology, Chengdu 610095, China.

Jue Wang is with Tencent AI Laboratory, Shenzhen 518000, China.

Heng-Yu Zhang is with the Department of Cardiology, West China Hospital, Sichuan University, Chengdu 610041, China (e-mail: zhanghengyu@wchscu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3190057.

Digital Object Identifier 10.1109/TCSVT.2022.3190057

information [8]. Different techniques are introduced to improve the fusion performance. However, the most important problem of deep learning-based fusion methods is that they rely heavily on paired inputs and ground truth.

To relax the constraint of the dataset, we propose a GAN-based fusion method to optimize the network using unpaired dataset, named UPHDR-GAN. First, compared to famous single-image enhancement methods [16]–[19] and some recent GAN-based image fusion methods [8], [20], [21] that are trained on paired datasets, the proposed method trains unpaired datasets and transfers the multi-exposure LDR domain images to HDR domain images. The datasets of common deep learning-based methods require the inputs and the ground truth images. However, obtaining HDR ground truth images is difficult and most existing datasets just include the input images. Some recent datasets [5] generate the ground truth images according to the inputs, but their variety of the scenes is so limited. Training the model on unpaired dataset can relax the constrain of paired training and broaden the application of the dataset. Second, unlike some methods that are designed for unpaired datasets mainly concentrate on processing single-input, our method is a multi-input method with the consideration of moving objects. For example, CycleGAN [22] is designed for training unpaired datasets and processing single-input. The CycleGAN is not suitable for fusing multi-exposure inputs because the forward process (composing multi-exposure images into an HDR output) may be learned properly, while the backward process (decomposing the HDR image into the multi-exposure images) may not converge successfully. The forward process and the backward process in CycleGAN are interactive. Therefore, the forward process will be influenced if the backward process cannot work satisfactorily. Even considering multi-input, simply concatenating multi-exposure inputs will result in severe ghosting.

In contrast, the UPHDR-GAN designs specific modules to solve such problems and produce informative HDR outputs with fewer ghosting artifacts. First, we introduce the initialization phase to maintain the content information between the reference and the output. The initialization phase totally avoids ghosting because it just transfers the reference images to HDR domain. Second, we improve the common adversarial loss to generate images with sharp edges (Fig. 2 (b)). Third, when fusing the information from the under- and over-exposure images, the min-patch training module (Fig. 2 (c)) is adopted to detect and handle the ghosting artifacts. The comparison results with several de-ghosting methods are shown in Fig. 1. The comparison methods have diverse artifacts, while our UPHDR-GAN handles the dynamic objects properly with the balance of the HDR transformation and content preservation.

In summary, the main contributions include:

- We proposed a GAN-based multi-exposure HDR fusion network, which relaxes the constraint of paired training dataset and learns the mapping between input and target domains. To our best knowledge, this work is the first GAN-based approach for unpaired HDR reconstruction.
- The proposed method can not only be trained on unpaired dataset but generate HDR results with fewer ghosting artifacts. We apply the modified GAN loss, the initialization phase and the min-patch training module to avoid ghosting and improve the image quality.
- We provided comprehensive comparisons with several leading methods. The results demonstrate that the proposed UPHDR-GAN outperforms existing methods and works well on challenging cases.

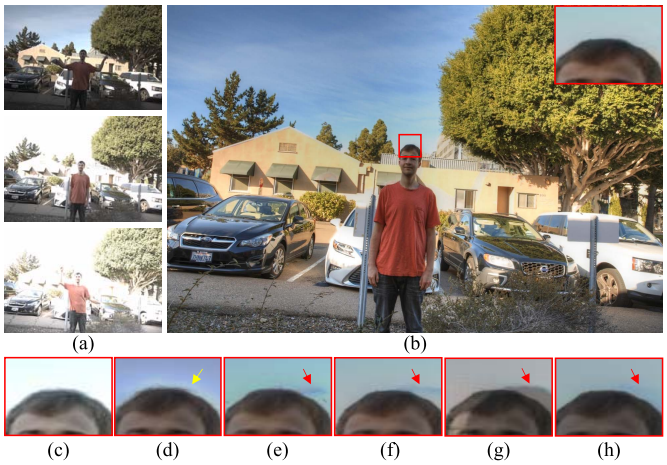


Fig. 1. LDR images with different exposures are shown in (a), and our result is shown in (b). (c) Result of Hu *et al.*'s method [3]. (d) Result of Sen *et al.*'s method [4]. (e) Result of Kalantari *et al.*'s method [5]. (f) Result of Wu *et al.*'s method [6]. (g) Result of Yan *et al.*'s method [7]. (h) Result of Niu *et al.*'s [8]. The proposed UPHDR-GAN handles moving objects better and generates results with fewer ghosting artifacts.

II. Related Works

A. HDR Imaging

HDR imaging has been extensively researched over the past decades. Existing HDR imaging methods can be mainly divided into two groups, static and dynamic scene methods.

a) Static scene methods: Debevec *et al.* first proposed to fuse different exposure images to an HDR image [23]. The original approaches produced spectacular results for static cameras and static scenes. Some variants are then introduced by generating disparity maps or using neural networks [24], [25]. Sun *et al.* computed the disparity map first and applied them to compute the camera response function [25]. Hashimoto *et al.* developed hard-to-view or non-viewable features and content of color images by a new tone reproduction algorithm [24]. There are also numerous static fusion methods that do not generate HDR outputs but directly obtain informative LDR results [26]–[29]. Li *et al.* incorporated the edge-preserving factors into the fusion method to preserve the details [28]. Wang *et al.* [29] presented a unified multi-scale densely connected fusion network to fuse the infrared and visible images. However, due to the lack of an explicit detection for the dynamic objects, the aforementioned methods are unaware of any motion in the scene, so as to be suitable for static scenes only.

b) Dynamic scene methods: Many de-ghosting algorithms are introduced to solve the problem that static methods are not applicable for many scenes [30], [31]. Some methods compute weight maps of input images and eliminate

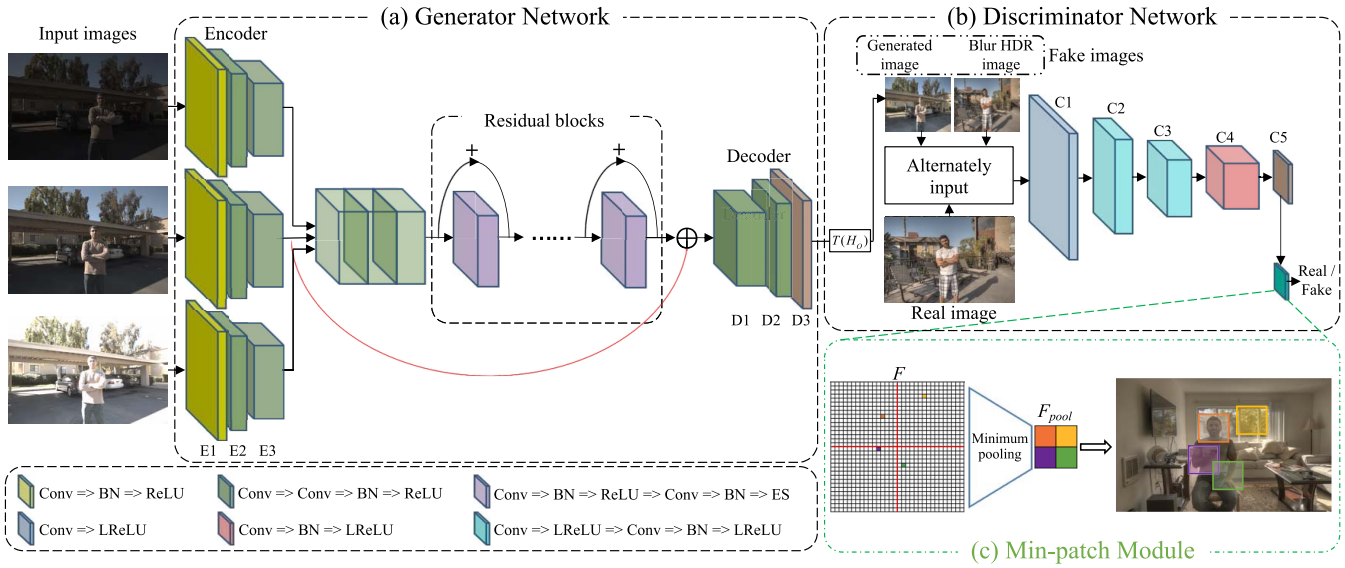


Fig. 2. The proposed method seeks to generate high-quality HDR results with unpaired datasets. **The generator** first extracts features from multi-exposure inputs using identical down-convolution blocks. The encoder features are then concatenated to be sent to the residual blocks. The decoder recovers the features to informative HDR images through up-convolution blocks. **The discriminator** distinguishes the generated and the real HDR images alternately. **The min-patch module** concentrates on the strange part of fake images and helps to avoid ghosting artifacts.

the moving contents together [32], [33]. Complementary, some methods merge images first and resolve ghosting of the results [34]. The misaligned pixels often appear in such methods so that they usually fail to fully utilize available content to generate HDR images. There are also some methods that are applying energy optimization to maintain image consistency or model the noise distribution of color values [35]. Besides, some more complicated methods based on optical flow [2] or patch-based correspondence [3], [4] are proposed to achieve more accurate image registration. Li *et al.* applied the optical flow to roughly align the multi-exposure images which are captured by hand-held cameras and then used the patch-based optimization to obtain full-aligned inputs [2]. Sen *et al.* integrated alignment and reconstruction in a patch-based energy minimization through an HDR image synthesis equation [4]. Hu *et al.* built new image stacks using a variant of PatchMatch to handle saturated regions and avoid the ghosting artifacts [3]. Although flow-based methods are able to align images with complex motions, they usually suffer from deformations in the regions with no correspondences, due to occlusions caused by parallax or dynamic contents. On the other hand, patch-based methods sometimes produce excellent results, while they are less efficient and usually fail in large motions and saturated regions. To overcome above issues, some deep learning approaches have been developed recently [5]–[7], [14]. The deep learning methods can obtain information from the training process to compensate for image regions. However, each of these methods only addresses part of the issues and needs paired data to optimize the network. We propose UPHDR-GAN to comprehensively handle existing issues, including solving ghosting artifacts and relaxing the constrain of paired data.

B. GAN-Based Fusion

GAN was proposed by Goodfellow *et al.* [36], which has achieved impressive results in image blending [37], image generation [38], [39], image style transfer [40], and solving jigsaw puzzles [41]. Generally, the inputs of common GAN-based methods are noise or a single image. Obtaining information from multi-inputs is also an important research topic [42], [43]. Guo *et al.* introduced a GAN-based multi-focus image fusion system, which utilized the generator to produce desired mask maps [44]. Huang *et al.* presented an adaptive weight block to determine whether source pixels are focused or not. [45] Li *et al.* proposed AttentionFGAN that applies the attention mechanism into the GAN framework and uses the attention features to fuse the infrared and visible image [46]. Recently, there are some GAN-based methods are proposed to handle multi-exposure images [8], [20], [21]. Xu *et al.* introduced the self-attention mechanism to solve the luminance variety of multi-exposure images [20]. Yang *et al.* fused the over- and under-exposed image by increasing the number of the discriminators [21]. Niu *et al.* incorporated the adversarial learning and a reference-based residual merging block to solve large motions [8]. However, these GAN-based methods rely heavily on paired training datasets so that their performances are greatly limited. In comparison, we propose UPHDR-GAN to fuse multi-exposure inputs, which is compatible with unpaired datasets, so that the flexibility and robustness of our proposed network are significantly improved.

III. Method

We propose a GAN-based multi-exposure fusion framework, which is the first method designed for handling HDR imaging tasks with unpaired datasets. Like common GAN framework, the generator G transforms inputs of source domain to desired outputs with the characteristics of the target domain, while the discriminator D distinguishes the target domain images from the generated ones to optimize G. Our collected dataset consists of scenes with and without ground truth. By disorganizing the correspondence between the inputs and ground truth, the unpaired training set is

TABLE I
DETAILED PARAMETER SETTINGS OF THE NETWORK, IN WHICH 'ES' INDICATES ELEMENT-WISE SUM

Inputs: 3 × [256, 256, 6]							
	Module		Conv			BN	Activation
			Kernel	Stride	Channel	Channel	
G	Encoder	E1	7	1	64	64	ReLU
		E2	3	2	128	-	-
			3	1	128	128	ReLU
		E3	3	2	256	-	-
			3	1	256	256	ReLU
	Residual blocks		3	2	256	256	ReLU
			3	1	256	256	ES
	Decoder	D1	3	1/2	128	-	-
			3	1	128	128	ReLU
		D2	3	1/2	64	-	-
			3	1	64	64	ReLU
		D3	7	1	3	-	-
D		C1	3	1	32	-	LReLU
		C2	3	2	64	-	LReLU
			3	1	64	64	LReLU
		C3	3	2	128	-	LReLU
			3	1	128	128	LReLU
		C4	3	1	256	256	LReLU
		C5	3	1	1	-	-
Output HDR H_o: [256, 256, 3]							
Tonemapped HDR $T(H_o)$: [256, 256, 3]							

obtained. To better describe the framework, two domain data are collected, including (1) the source LDR domain X, which is constituted by a wide diversity of multi-exposure sequences $x = \{x_1, x_2, x_3\}$, and (2) the target domain Y, which consists of a collection of HDR images. We denote their data distributions as $x \sim p_{\text{data}}(x)$ and $y \sim p_{\text{data}}(y)$, respectively. The proposed UPHDR-GAN can generate HDR images with fewer ghosting artifacts in the absence of paired datasets.

A. Network Architecture

UPHDR-GAN is an images-to-image task with three inputs and one output. The structure of UPHDR-GAN is illustrated in Fig. 2. The detailed layer configurations of the network architecture are displayed in Table I. To improve the efficiency, We crop 256 × 256 overlapped patches from the training images with a stride of 64 rather than optimizing the model with the full-size images. The encoder contains three branches and the input size of each branch is $256^2 \times 6$, which is the concatenation of the inputs $x = \{x_1, x_2, x_3\}$ and their mapped HDR images $H_m = \{H_1, H_2, H_3\}$. H_m is obtained using a simple gamma encoding:

$$H_i = \frac{{x_i}^{\gamma}}{t_i}, \quad \gamma > 1 \tag{1}$$

where x_i is the input image and t_i is the corresponding exposure time. The LDR images and the mapped HDR images are complementary, where the former one detects the saturation and misalignments, and the latter one facilitates the convergence of the network across LDR images.

After getting the HDR output H_o, we add a μ-law [5] post-processing to refine the range of generated HDR images because computing the loss functions on the tone-mapped HDR images is more effective:

$$T(H_o) = \frac{\log(1 + \mu H_o)}{\log(1 + \mu)} \tag{2}$$

where H_o is the output HDR image and real HDR image respectively, μ represents the amount of compression and is set to 5,000 in our implementation.

1) Generator: The generator network is composed of the encoder, the residual blocks and the decoder. Specifically, the encoder consists of three convolutional blocks: E1, E2 and E3, as described in Table. I. Useful signals are extracted in the encoder process and used for following residual blocks to explore high-level features. Two transposed convolutional blocks (D1 and D2) and a convolutional layer (D3) constitute the decoder to recover the features to output images.

2) Discriminator: The discriminator is complementary to the generator. PatchGAN [47] is applied to classify the image patch rather than a full image. We crop 70 × 70 overlapped patches from generated HDR images and real HDR images to train the patch-based discriminator. However, not all regions in the patch contribute to the discriminator optimization during training. If the generator produces images with regions that are strange and different from the real images, the special regions can be considered as undesirable ghosting artifacts. Paying more attention to the strangest parts is essential.

3) Min-Patch Module: We introduce the min-patch training module (Fig. 2 (c)) at the end of the PatchGAN. The implementation of min-patch training is to add an optional minimum pooling layer to the final output of the discriminator [43]. We define F to represent the features after the 'C5' convolutional layer in the discriminator. When training the discriminator, conventional PatchGAN is applied and the network is optimized with F. When training the generator, we add the minimum pooling layer after the 'C5' convolutional layer. The features after the minimum pooling layer (F_{pool}) are used to compute the loss. The generator is optimized with F_{pool}, which plays a vital role in detecting the most important parts of the generated images, such as the error parts or strange parts. The discriminator distinguishes the real image from the fake image using common PatchGAN and is trained with F. In our implementation, the size of features F after 'C5' convolutional layer is 64 × 64. We use 16 × 16 minimum pooling for the min-patch training module and output features F_{pool} with size 4 × 4 to optimize the generator.

B. Loss Function

As GAN is a min-max optimization system, the proposed UPHDR-GAN optimizes the following equation to strike a balance between the generator and the discriminator:

$$G^*, D^* = \arg\min_G \max_D L(G, D) \tag{3}$$

Based on HDR imaging properties, the objective function is designed to have the following two items: (1) the GAN loss $L_{\text{GAN}}(G, D)$ to achieve desired transformation to convert multi-exposure inputs into HDR outputs; (2) the content loss $L_{con}(G)$ to preserve the image semantic information during HDR transformation. The full loss function is:

$$L(G, D) = L_{\text{GAN}}(G, D) + w_{con} L_{con}(G) \tag{4}$$

(a) (b)

Fig. 3. Two examples of the blur dataset. (a) The tone-mapped HDR images. (b) Blur results of (a).

where w_{con} is a hyper-parameter to control the relative importance of the content loss, so as to balance the effects of transformation and content preservation.

1) GAN Loss: The GAN loss helps G to generate results similar to the target domain images in the absence of ground truth, and confuses D using the generated HDR images and real HDR images. However, applying vanilla GAN loss is insufficient, which cannot preserve the edge and boundary information, while such information is important for HDR images. For this reason, Chen *et al.* [48] proposed to confuse D with a blur dataset, which has been proven useful for the style transformation. The blur dataset is considered as fake images to drive the generator to produce images with clear edges. Similarly, we also add a blur HDR dataset to facilitate G to generate high-quality output. Specifically, for the target images $\{y_j\}_{j=1,...,M} \in Y$, we utilize Gaussian filter with kernel size 5×5 to remove their clear edges and generate the blur dataset $\{b_j\}_{j=1,...,M} \in B$. We show two examples of the blur dataset in Fig. 3. The characteristic of blur edges should be avoided in generated images. Selecting the blur dataset as fake images can help the network produce images without blur edges. In other words, there are three categories that need to be classified by the discriminator: $G(x)$, b and y, among which the generated image $G(x)$ and the blurred HDR image b are fake inputs, and the real HDR image y is real input. The modified adversarial loss is designed as:

$$\begin{aligned} L_{\mathrm{GAN}}(G, D) = & \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\big[\log D(y)\big] \\ & + \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\big[\log(1 - D(G(x))\big] \\ & + \mathbb{E}_{b\sim p_{\mathrm{data}}(b)}\big[\log(1 - D(b))\big] \end{aligned} \tag{5}$$

We adopt the negative form of the modified adversarial loss in order to use the min-patch training module properly. Conventional adversarial loss minimizes the generator loss while maximizing the discriminator loss. Now, we train the generator to maximize the loss function and the discriminator to minimize the loss function. The inverse optimization is specifically designed for the min-patch training module, which is only used when training the generator. The modified generator loss tries to maximize the discriminator values after passing the minimum pooling. The lower discriminator outputs imply the fake patches, which may represent the blur or ghosting regions. The modified generator loss can concentrate on these strange parts by maximizing the lower discriminator values.

2) Content Loss: The GAN loss just ensures the generator produces images that are similar to the real HDR domain images. The semantic information preservation cannot be guaranteed by using adversarial loss alone. Adding additional constraints for semantic consistency is necessary. Generally, we select the image with middle-exposure as the reference image, and align images with under- and over-exposure to the reference. The content loss is defined to constrain the paired middle-exposure input x_2 and the generated result $G(x)$ about the semantic similarity. Instead of using common MSE loss function, the perceptual loss [49] is applied to constrain the content differences, which is formulated as:

$$L_{con}(G) = \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\big[||VGG_l\big(G(x)\big) - VGG_l\big(x_2\big)||_1\big] \tag{6}$$

where the selection of layers l is important. Larger l will extract high-level features. We utilize the features of the 'conv4_4' layer from the VGG19 network in our method.

The hyper-parameter w_{con} is added to balance the adversarial loss and content loss. The adversarial loss works on unpair domain translation, while the content loss constrains pair content preservation. A larger w_{con} destroys the domain transformation and generates results that do not like desired HDR images due to the excessive content preservation from inputs, while a small w_{con} concentrates more on unpaired domain translation and the semantic information of the reference image will be destroyed. In order to achieve the balance, w_{con} is empirically set to be 1.5 at the initial stage. After the training process becomes increasingly stable and the content information from the reference is maintained reasonably, w_{con} is gradually decreased to achieve the domain transformation. w_{con} is described as:

$$w_{con} = w_{con} \times 0.96^{\lfloor N_e/10 \rfloor} \tag{7}$$

where N_e is the number of epochs, which is set to 200 in our implementation.

IV. Experiments

The datasets and implementation details are first illustrated in Section IV-A. Comprehensive experiments are then conducted, including quantitative comparisons (Section IV-B), qualitative assessments (Section IV-C) computational complexity (Section IV-D), results on sequences captured by hand-held smartphones (Section IV-E), and ablation studies (Section IV-F). Specifically, we first compare the proposed method with several methods that can only be applied to fuse static inputs [20], [26], [27], [50]–[52], and then compare with several classic de-ghosting methods, including two patch-based methods [3], [4], two deep neural network (DNN) mergers with and without optical flow registration, respectively [5], [6], a non-local network [7], and a GAN-based method [8]. We use the under- and the over-exposed image to produce the results of Xu *et al.*'s method [20] because their method only takes two inputs.

TABLE II
DETAILED SOURCE INFORMATION OF OUR DATASET

Source Name	URL	Number
HDReye	https://mmspg.epfl.ch/downloads/hdr-eye/	46
Fairchild	http://rit-mcsl.org/fairchild/HDR.html	103
EmpaMT	http://empamedia.ethz.ch/hdrdatabase/index.php	30
Kalantari [5]	https://cseweb.ucsd.edu/~viscomp/projects/SIG17HDR	74
Tursun [53]	http://user.ceng.metu.edu.tr/~akyuz/files/eg2016/index.html	17

Fig. 4. An example of the unpaired training dataset, among which three images on the left are the input images, while the rightmost image is the target image in the HDR domain.

A. Datasets and Implementation Details

The datasets of common deep learning-based multi-exposure fusion methods usually include multi-exposure input images and ground truth HDR image. However, obtaining corresponding ground truth HDR images is difficult and most existing datasets just include the input images. Moreover, many existing datasets only include static scenes. Although some of them include moving objects, the dynamic scenes occupy a small proportion. Kalantari *et al.* introduced the first HDR dataset, however, the variety of the scenes is so limited [5]. Our method relaxes the constraints of paired input and learns the transformation from the source LDR domain to the target HDR domain. The network is trained to fuse multi-exposure inputs in the absence of corresponding ground truth. We have collected a total of 270 groups of images from various sources, as seen in Table II for the detailed information. The ground truth images in the test set are required to compute the quantitative scores. The image sequences from Tursun *et al.* [53] and Fairchild do not contain the ground truth images. Therefore, we randomly select dynamic test scenes from other three datasets. Kalatari *et al.*'s dataset [5] only contains dynamic scenes. Twenty static test scenes are randomly selected from remainder two datasets. As for twenty dynamic scenes, 6 sequences originate from the HDReye dataset, 4 sequences originate from the EmpaMT dataset and 10 sequences originate from the Kalantari *et al.*'s dataset [5]. As for twenty static scenes, 11 sequences originate from the HDReye dataset and 9 sequences originate from the EmpaMT dataset. Finally, 40 groups of images are selected as the test set and 230 groups of images are selected as the training set. The test set and the training set are completely distinct. Some of the sequences include approximately 10 multi-exposure inputs, from which we select 3 images with minimum, medium and maximum exposure as training inputs.

By disorganizing the correspondence between the inputs and ground truth, the unpaired training set is obtained. An example of the unpaired training dataset is shown in Fig. 4. The training images are first aligned using a homography before they are sent to the network, which is more effective and helps the network concentrates more on the moving objects. All training images are resized to 1000 $\times$ 1500. Then, we crop 256 $\times$ 256 overlapped patches from the training images with a stride of 64 to improve the training efficiency. The pre-processing will create 54,240 patches. After that, we utilize the data augmentation, including the flipping and rotation to enrich the training data by 8 times. Finally, the training set consists of 433,920 training patches, which is large enough to encompass all the possibilities and train our architecture.

We implement UPHDR-GAN in PyTorch and the model is trained on an NVIDIA RTX 2080Ti GPU for 200 epochs. The entire training process costs 2 days on average. Adam optimizer is selected to iterate the network. The learning rate of the generator and the discriminator is set to 2.0×10^{-4} and 1.0×10^{-4}, respectively. We introduce an initialization phase to help the convergence and guide the network to learn the correct domain transformation. In initialization, the generator is designed to reconstruct the semantic information of middle-exposure input and ignore the domain translation. For this purpose, the generator G is pre-trained using merely the content loss L_{con}. Two examples are presented in Fig. 5 that include the input images and the results after pre-training. Ablation experiments of the initialization phase are also performed in Section IV-F. The initialization phase contributes to controlling the over-exposed regions and enriching the overall colors. Moreover, the network properly reconstructs the content information of middle-exposure input. Since we select the middle-exposure image as the reference, the initialization also helps to avoid ghosting.

B. Quantitative Comparisons

Although the proposed UPHDR-GAN can efficiently fuse multi-exposure images without ground truth, we select the test set for quantitative comparisons from paired datasets that include multi-exposure inputs and HDR images. As the ground truth is available, we can conduct various quantitative evaluations and comparisons. As for the comparisons with static

TABLE III

QUANTITATIVE COMPARISON OF UPHDR-GAN WITH THE COMPARISON METHODS ON TWENTY STATIC SCENES. THE LEFT PART SHOWS THE COMPARISON RESULTS WITH METHODS THAT ARE SUITABLE FOR STATIC SCENES, AND THE RIGHT PART REPRESENTS THE COMPARISON RESULTS WITH METHODS THAT ARE BOTH SUITABLE FOR STATIC AND DYNAMIC SCENES. RED COLOR INDICATES THE BEST PERFORMANCE AND BLUE COLOR INDICATES THE SECOND-BEST RESULTS. THE BEST RESULTS OF STATIC METHODS ARE UNDERLINED

Methods	Mertens [26]	Li2012 [51]	Li2013 [50]	Paul2016 [52]	Ma2019 [27]	Xu2020 [20]	Sen [4]	Hu [3]	Kalantari [5]	Wu [6]	Yan [7]	Niu [8]	Ours
PSNR ↑	29.817	30.289	31.020	31.876	<u>33.469</u>	32.582	39.105	32.192	40.008	39.505	39.994	40.637	40.601
SSIM ↑	0.9511	0.9527	0.9575	0.9584	<u>0.9675</u>	0.9651	0.9664	0.9655	0.9701	0.9681	0.9692	0.9715	0.9717
HDR-VDP-2.2 ↑	54.904	53.432	56.058	56.935	<u>57.910</u>	54.151	56.345	55.973	59.782	60.155	58.362	61.348	61.916
TMQI ↑	0.854	0.859	0.871	0.874	<u>0.887</u>	0.872	0.882	0.878	0.889	0.890	0.887	0.893	0.895

(a) (b)

Fig. 5. Results of the initialization phase. (a) The middle-exposure inputs. (b) The generated results after the pre-training with 10 epochs.

scenes, we compute four metrics, including the PSNR values [54], the SSIM values [55], the HDR-VDP-2.2 scores [56] and the tone mapped image quality index (TMQI) scores [57]. The PSNR value approaches infinity as the MSE approaches zero and a higher PSNR value provides a higher image quality. The SSIM is considered to be correlated with the quality perception of the human visual system [55]. HDR-VDP-2.2 is a calibrated objective method that can tackle both HDR and LDR signals [56]. The TMQI score combines the multi-scale signal fidelity measure and a naturalness measure to evaluate the tome mapped images [57]. As for the comparisons with dynamic scenes, we further compute the PU-PSNR and PU-SSIM values [58] with 1,000 cd/m^2 display, which represents current commercial HDR display technology. The two perceptually uniform (PU)-encoding metrics convert absolute HDR linear color values into approximately perceptually uniform values and expect that the values in images correspond to the luminance emitted from the HDR display. The higher PSNR, SSIM, HDR-VDP-2.2, TMQI, PU-PSNR and PU-SSIM scores indicate better image quality. The quantitative comparison results are presented in Table III and IV.

Twenty static scenes and twenty dynamic scenes, which include multi-inputs and corresponding ground truth, are collected as the test set for quantitative comparisons. The test set is completely distinct from the training set to ensure the evaluation is fair. The proposed method is first compared with several classic methods that can only be applied to fuse static inputs [20], [26], [27], [50]–[52]. The left part in Table III displays the quantitative comparison results with the static methods. Some of static methods fuse multi-exposure inputs with the absence of ground truth, and therefore resulting in lower scores when computing the evaluation metrics between the generated image and the ground truth. The comparison results with several de-ghosting methods [3]–[8] on these static scenes are then reported in the right part of Table III. These methods are designed for handling sequences with moving objects, which can solve the slight movements (such as the moving leaves caused by the wind and the flowing water) and obtain higher scores than the aforementioned static methods. The proposed UPHDR-GAN abandons the constraint of ground truth, but can extract information from the target HDR dataset, hence providing results with better PSNR, SSIM, HDR-VDP-2.2 and TMQI values on average.

Table IV exhibits the comparison results of UPHDR-GAN with several de-ghosting methods [3]–[8] on twenty dynamic scenes. Two patch-based methods [3], [4] generate the registered image stacks according to the patch match-oriented optimization. Kalantari *et al.* [5] and Wu *et al.* [6] obtain HDR results through deep neural networks. Yan *et al.* use the non-local correlation to tackle the ghosting artifacts [7]. Niu *et al.* introduce the adversarial loss to improve the unsatisfactory regions by creating realistic information [8]. These deep learning-based algorithms have demonstrated significant performance advantages over patch-based methods. However, the deep learning-based methods are not sensitive to large motions and lack robustness. These comparison methods focus on fusing the multi-exposure images but cannot handle the dynamic objects well, which affects their performance. On the contrary, the proposed initialization phase totally avoids ghosting because it just transfers the reference images to the HDR domain. Then, when fusing the information from the under- and over-exposure images, the min-patch training module helps to detect and avoid ghosting artifacts. Overall, by incorporating the initialization phase and the min-patch training module, our method owns superior performance.

C. Qualitative Comparisons

In this section, our method is first compared with [20], [26], [27], [50]–[52] on static scenes (Fig. 6). The comparison methods are mature enough to handle images that are static, but ignore the tiny motions, such as the moving leaves caused by wind. The comparison methods produce the ghosting artifacts in the left case in Fig. 6, which are caused by the slight movements of the leaves. Some of them

TABLE IV

QUANTITATIVE COMPARISON OF UPHDR-GAN WITH THE DYNAMIC METHODS ON TWENTY DYNAMIC SCENES. RED COLOR INDICATES THE BEST PERFORMANCE AND BLUE COLOR INDICATES THE SECOND BEST RESULTS

Methods	PSNR ↑	SSIM ↑	PU-PSNR ↑	PU-SSIM ↑	HDR-VDP-2.2 ↑
Sen [4]	40.924	0.9806	41.856	0.9832	57.249
Hu [3]	34.785	0.9725	38.604	0.9760	56.427
Kalantari [5]	42.532	0.9871	40.710	0.9821	61.988
Wu [6]	41.660	0.9844	41.054	0.9854	62.345
Yan [7]	42.321	0.9869	40.942	0.9855	59.417
Niu [8]	43.113	0.9877	41.969	0.9860	63.050
Ours	43.005	0.9880	42.115	0.9860	63.542

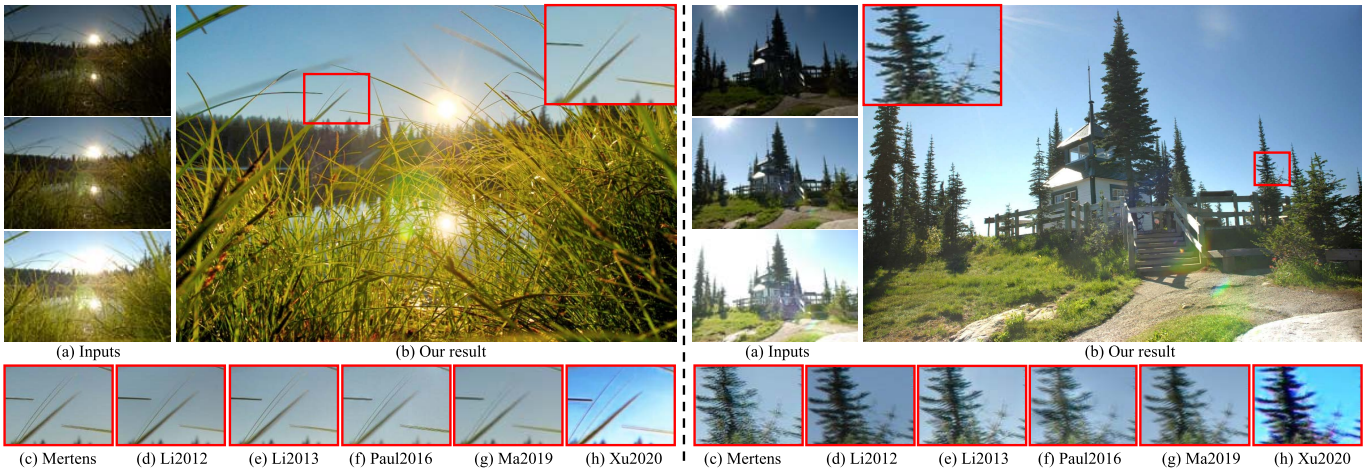


Fig. 6. Visual comparisons with several representative static methods. (a) Input images. (b) Our result. (c) Result of Mertens *et al.*'s method [26]. (d) Result of Li *et al.*'s method [51]. (e) Result of Li *et al.*'s method [50]. (f) Result of Paul *et al.*'s method [52]. (g) Result of Ma *et al.*'s method [27]. (h) Result of Xu *et al.*'s method [20].

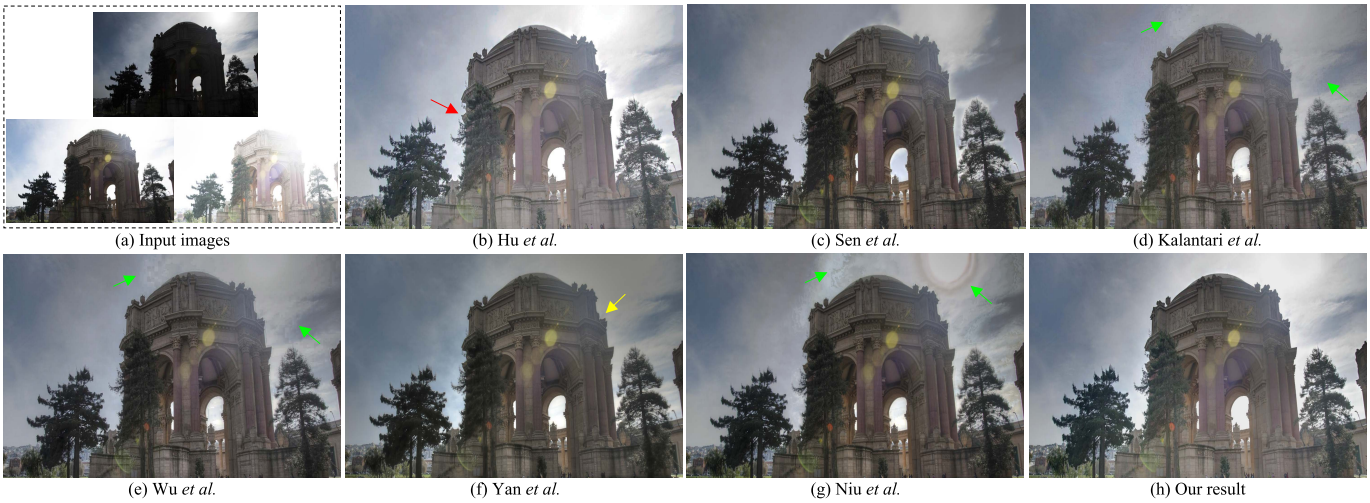


Fig. 7. Visual comparisons with de-ghosting methods. (a) Input images. (b) Result of Hu *et al.*'s method [3]. (c) Result of Sen *et al.*'s method [4]. (d) Result of Kalantari *et al.*'s method [5]. (e) Result of Wu *et al.*'s method [6]. (f) Result of Yan *et al.*'s method [7]. (g) Result of Niu *et al.*'s [8]. (h) Our result. Note that, Hu *et al.*'s method [3] produces noise around the building in (b). Please zoom in for details.

design specific strategies to detect and solve the dynamic contents, such as guided filtering [50]. However, the results are still unsatisfactory. In the right case, the static methods suffer from the blurring artifacts around the tree. Xu *et al.* solely obtained information from the under- and over-exposed images [20], which leads to the mediocre result with color deviation.

Fig. 7 and Fig. 8 show the qualitative comparisons against several state-of-the-art de-ghosting methods [3]–[8]. Two patch-based methods [3], [4] tend to generate fully registered input image stacks, but cannot reconstruct the regions with rich textures or large motions. Hu *et al.*'s method [3] generates results with noise around the building (red arrow) in Fig. 7 and unclear edges in Fig. 8. Sen *et al.*'s method [4] produces

Fig. 8. Visual comparisons with de-ghosting methods. (a) Input images. (b) Our result. (c) Result of Hu *et al.*'s method [3]. (d) Result of Sen *et al.*'s method [4]. (e) Result of Kalantari *et al.*'s method [5]. (f) Result of Wu *et al.*'s method [6]. (g) Result of Yan *et al.*'s method [7]. (h) Result of Niu *et al.*'s [8]. (i) Zoomed-in areas of our result. The scene is challenging because there are large foreground motions between input LDR images. The proposed UPHDR-GAN can properly deal with the motions caused by moving people.

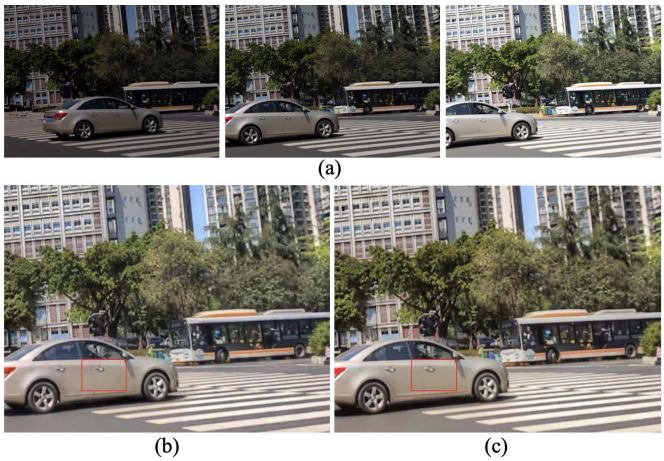


Fig. 9. Comparisons with Niu *et al.*'s work [8] on scene with large motions. (a) Input images with a moving car, which causes large motions. (b) Result of Niu *et al.*'s work [8]. (c) Our result.

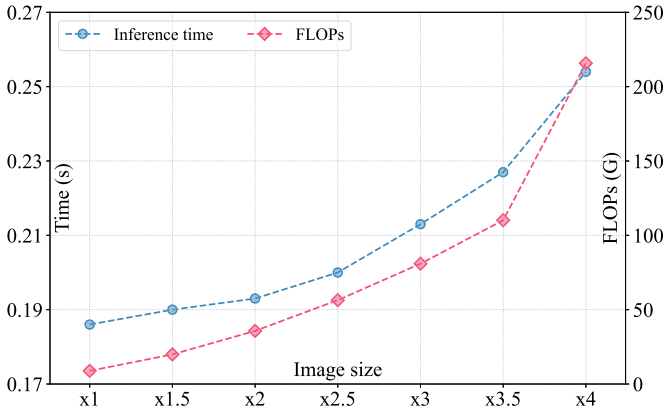


Fig. 10. The variation trend of the FLOPs and the inference times when selecting test images with different resolutions. The smallest image size in the figure is 256 × 384, which is labeled as ×1. The image size of ×1.5 in the figure is 384 × 576. Therefore, the largest image size ×4 is 1024 × 1536.

results with serious halo artifacts in Fig. 7 and ghosting artifacts in Fig. 8. The deep learning-based methods can obtain information from the training process to compensate for image regions. However, they only perform well in one way or another. Kalantari *et al.* [5] and Wu *et al.* [6] adopt similar network architecture but different in pre-processing. Kalantari *et al.* [5] apply flow-based pre-processing to align the inputs, while Wu *et al.* [6] process the alignment and the fusion together. The two methods suffer from similar artifacts, including the problematic transformation in the junction regions of the sky and the cloud (green arrows) in Fig. 7, and the ghosting artifacts in Fig. 8. Yan *et al.* decrease the ghosting artifacts by using the non-local module, which is designed based on the pixel correspondence [7]. However, their method cannot generate sharp edges (yellow arrow) in Fig. 7 and cannot avoid the ghosting artifacts in Fig. 8. Niu *et al.* incorporated the adversarial learning to produce faithful information in the regions with missing content [8]. Their method also suffers from the problematic transformation in the junction regions (green arrows) in Fig. 7 and the unreasonable color reconstruction in Fig. 8. Our method is more sensitive to ghosting artifacts and handles them properly. We further show the comparisons with Niu *et al.*'s method on scene with large motions. Fig. 9 show the input images (Fig. 9 (a)) and results of Niu *et al.*'s method (Fig. 9 (b)) and our method (Fig. 9 (c)). Overall, our method achieves comparable result with Niu *et al.*'s method. Specifically, our method preserves more details than Niu *et al.*'s method, such as the crevice between two car doors (red box).

D. Computational Complexity

Computing efficiency is also an important factor for evaluating the fusion performance. The comparisons of inference time and parameters are then conducted. The results for fusion images with size 1000 × 1500 on the test set are reported in Table V. There is a large difference between different methods. Two patch match-based methods [3], [4] take approximately 60*s* and 80*s*, respectively. The deep learning-based methods are faster than patch patch-based methods due to the training environment. Kalantari *et al.*'s method [5] costs about 30*s*, which is mainly spent on the optical flow pre-processing. Wu *et al.*'s method [6] and Yan *et al.*'s method [7] take less inference time but their networks include a large number of parameters. Niu *et al.*'s method [8] and the proposed UPHDR-GAN have similar performance on the computational complexity. However, the proposed method needs fewer para-

TABLE V

THE INFERENCE TIME AND PARAMETERS OF DIFFERENT METHODS ON THE TESTING SET WITH SIZE 1000 × 1500. THE '-' DENOTES THAT THE PATCH MATCH-BASED METHODS DO NOT HAVE PARAMETERS

Methods	Sen [4]	Hu [3]	Kalantari [5]	Wu [6]	Yan [7]	Niu [8]	PBR-GAN
Environment	CPU	CPU	CPU+GPU	GPU	GPU	GPU	GPU
Time (s)	61.81	79.77	29.14	0.24	0.31	0.29	0.25
Parameters (M)	-	-	0.3	20.4	38.1	2.56	2.21

Fig. 11. The results on real-life scenes captured by HUAWEI Mate 10 smartphones. Each case includes three inputs with different exposures and corresponding results generated by UPHDR-GAN.

TABLE VI

ABLATION EXPERIMENTS OF DIFFERENT COMPONENTS. RED COLOR INDICATES THE BEST PERFORMANCE AND BLUE COLOR INDICATES THE SECOND BEST RESULTS

	$w_{con}=0.25$	$w_{con}=0.5$	$w_{con}=1$	MSE	w/o. initialization	w/o. min-patch	w/o. blur dataset	only Kalantari's dataset [5]	Ours
PSNR ↑	36.587	39.745	41.623	36.209	41.231	40.724	42.441	42.995	43.005
SSIM ↑	0.9789	0.9803	0.9840	0.9751	0.9837	0.9811	0.9861	0.9881	0.9880
PU-PSNR ↑	35.512	37.559	40.023	34.725	39.574	38.827	41.721	42.107	42.115
PU-SSIM ↑	0.9766	0.9779	0.9829	0.9758	0.9820	0.9782	0.9852	0.9859	0.9860
HDR-VDP-2.2 ↑	56.842	58.196	60.018	56.131	59.877	58.624	61.596	63.491	63.542

meters and costs less inference time by taking the advantage of the well-designed architecture.

To better illustrate the computing efficiency of the proposed method, Fig. 10 shows the variation trend of the FLOPs and the inference time when selecting test images with different resolutions. The smallest image size in Fig. 10 is 256 × 384, which is labeled as ×1. The largest image size ×4 is 1024 × 1536. Obviously, the FLOPs and the inference time increase with the increase of image resolution. When the resolution changes from ×3.5 (896 × 1344) to ×4 (1024 × 1536), the FLOPs increases dramatically. If we continue to enlarge the image size, the curve of the FLOPs will have a larger slope. In order to be consistent with other methods and obtain the balance between network performance and computational complexity, we set the size of test images as 1000 × 1500.

E. Results on Sequences Captured by Hand-Held Smartphones

We also conduct experiments on multi-exposure images captured by hand-held smartphones. We apply the HUAWEI Mate 10 to capture the input sequences, whose exposure time is adjusted manually. The captured scenes may have two problems: large-scale shaking and dynamic objects. To solve the first problem, we adopt the homograph registration from [59] to achieve the background alignment. Then, the proposed architecture can handle the artifacts caused by dynamic objects. The fusion results on real-life images are shown in Fig. 11. The proposed method also performs well because the training dataset contains diverse scenes, including many real-life sequences captured by different devices.

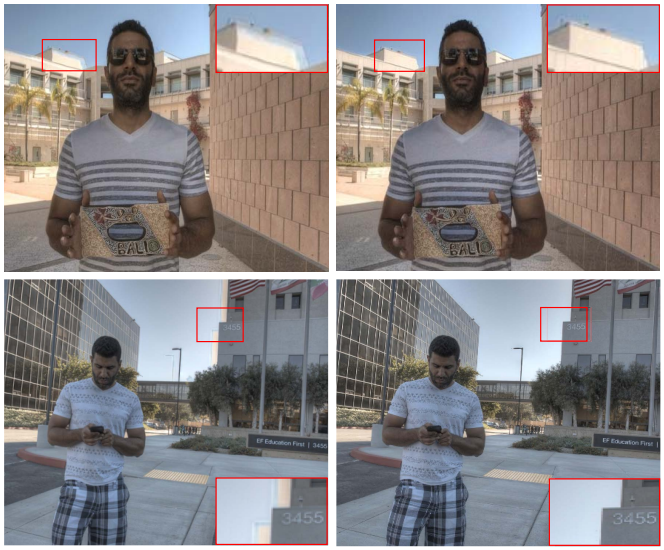
F. Ablation Studies

We conduct the ablation studies of different items in the architecture to understand the effectiveness of our designed modules. Table VI displays the ablation results of different components. First, the results from the second column to the fourth column show the importance of selecting suitable weights of the content loss. Second, the fifth column shows the evaluation scores when applying the MSE loss as the content loss. Third, the results when we remove the initialization phase are listed in the sixth column in Table VI. Fourth, the seventh column shows the results when removing the min-patch training module. Fifth, the results without blur dataset are exhibited in the eighth column. Last, the ninth column displays the results when we merely train our network on Kalantari *et al.*'s dataset. The results demonstrate that each component contributes to the final results.

1) Ablation Study of w_{con}: We first conduct the experiments of selecting different w_{con} to illustrate why we set the weight to 1.5. The results when we select different w_{con} are shown in

(a) w_{con}=0.25 (b) w_{con}=0.5 (c) w_{con}=1 (d) w_{con}=1.5

Fig. 12. The effect of different w_{con}. We set w_{con} to be 1.5 at the initial stage to keep a balance between HDR transformation and content preservation.



(a) Without min-patch training (b) With min-patch training

Fig. 13. The influence of min-patch training. (a) The generated results without min-patch training. (b) The generated results with min-patch training.

(a) Without blur dataset (b) With blur dataset

Fig. 14. The influence of blur dataset. (a) Results without blur dataset. (b) Results with blur dataset.

Table VI and Fig. 12. From the second column to the fourth column in Table VI, we can conclude that unsuitable weights of the content loss apparently degrade the results, which has a consistent performance with the qualitative results in Fig. 12. Smaller w_{con} cannot generate desired details or suffer from ghosting artifacts because they tend to learn the translation but ignore preserving the semantic content information (Fig. 12 (a)-(c)). We set w_{con} to be 1.5 when the network in the initialization to strike a balance between unpaired domain transformation and paired semantic information preservation. If we continue to increase the value of w_{con}, the results will be similar to the middle-exposure LDR image because they bring more content information from the input so that the dynamic range of the result is limited.

2) Ablation Study of Min-Patch Training Module: Conventional discriminator can distinguish the real HDR images and generate HDR images. However, not all regions contribute to the discriminator optimization during training. If a small part of the generated image is so strange as to be different from the real image, it can be considered as ghosting artifact. We add the min-patch training module to detect such regions and avoid ghosting artifacts. The quantitative results when removing the min-patch training module in Table VI (the seventh column) are worse than the complete UPHDR-GAN. Fig. 13 shows the effectiveness of the min-patch training module. After using the min-patch training module, UPHDR-GAN generates results with fewer artifacts (Fig. 13 (b)) compared to results without the min-patch training module (Fig. 13 (a)).

Fig. 15. Fusion results when selecting different input images as the reference. (a) Input images. (b) Results when selecting the middle-exposure image as the reference. (c) Results when selecting the under-exposure image as the reference. (d) Results when selecting the over-exposure image as the reference.

3) Ablation Study of Blur Dataset: Simply applying GAN loss is not sufficient for generating sharp HDR images. Having clear edges is an important characteristic of HDR images, but common GAN loss may produce results with unclear edges. To solve the problem, we add a blur dataset B as fake images to confuse the discriminator to produce images with sharp edges. The eighth column in Table VI presents the quantitative results when we remove the blur dataset. Corresponding evaluation scores are lower than the final results. Fig. 14 shows the qualitative results of without and with the blur dataset, among which the results with blur dataset (Fig. 14 (b)) have more sharp edges, such as the line shadow in the window region of the top case and the boundaries of the ceramic tiles in the bottom case.

4) Ablation Study of Different Reference: We also conduct the experiments when selecting different input images as the reference. Fig. 15 (a) are the input images with different exposures. Fig. 15 (b) are the results when selecting the middle-exposure image as the reference, while Fig. 15 (c) and Fig. 15 (d) are the results when selecting the under-exposure image and the over-exposure image as the reference, respectively. The two scenes in Fig. 15 have background misalignments between the input images. Furthermore, there is a moving person in the right case. The proposed method can handle the misalignments and solve the moving objects well no matter which input image is chosen as the reference. For example, in the right case, when we select the under-exposure image as the reference, the semantic information of the fusion result (Fig. 15 (c)) is the same to the under-exposure image. The proposed method can properly handle the moving objects when fusing information from other exposure images. However, the image quality between (b), (c) and (d) are different. The under-exposure image has large black regions due to the insufficient exposure time. If we select the under-exposure image as the reference, the result may suffer from color-drift (green boxes in Fig. 15 (c)). On the contrary, if we choose the over-exposure image as the reference, the content of over-exposed regions cannot recover well because noise can be easily introduced (Fig. 15 (d)). Obtaining information from near exposure is easy. It is challenging to acquire information from over-exposure image when the under-exposure image is selected as the reference, and vice versa. It is reasonable that the image quality of Fig. 15 (c) and (d) is slightly inferior to Fig. 15 (b) because the target HDR domain is the collection of HDR images that correspond to the distribution of 2-nd input images. Suitable techniques to adjust the exposure are necessary to generate high-quality results when selecting the under- and over- exposure images as the reference.

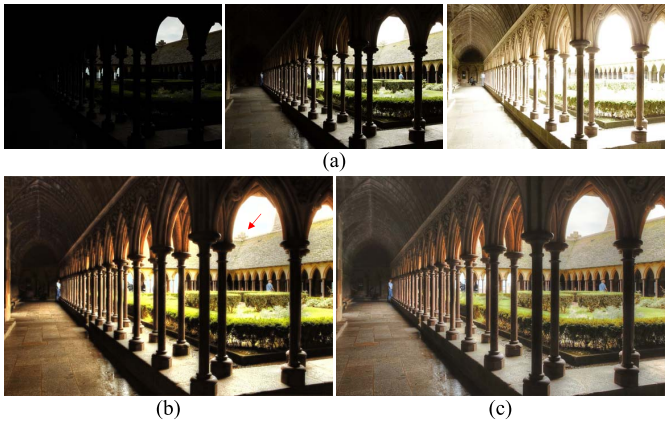


Fig. 16. Ablation experiment of different content loss. (a) Input images. (b) Result when selecting the MSE loss as content loss. (c) Result when selecting the perceptual loss as content loss.

5) Ablation Study of Different Content Loss: We show the results when applying different forms of the content loss in the fifth column of Table VI and Fig. 16. Our method adopts the perceptual loss (Fig. 16 (c)) as the content loss to achieve high-level feature abstraction, which keeps the content information between the middle-exposure image and the generated image although the middle-exposure image and the result have different styles. Fig. 16 (b) shows the result

when selecting the MSE loss as the content loss, which means $L_{con}(G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[(G(x) - x_2)^2\right]$. The MSE loss is more strict than the perceptual loss because it directly minimizes the difference between two images. As for our task, the MSE loss tends to constrain the generated image to be similar to the reference, and cannot learn the domain transformation satisfactorily. The result in Fig. 16 (b) has large black regions and cannot acquire the details of under- and over-exposed regions from other exposure images (red arrow). In Table VI, the quantitative scores when selecting the MSE loss as content loss are also lower than the perceptual loss.

G. Discussion

Multi-exposure image fusion is a challenging topic, especially considering the image quality of generated images (related to the under- or over-exposed regions) and the ghosting artifacts (caused by the moving objects). Although we have collected a dataset that includes a variety of scenes and can satisfy recent requirements, creating a larger comprehensive dataset with more diverse scenes is helpful for the development of image fusion. Besides, as for deep learning-based methods, the number of input images is commonly fixed to three due to the network architecture. We also consider increasing the flexibility of input exposure numbers as our future work. This may be implemented by using a fully convolutional network, which is shared by different exposed images, enabling the network to process arbitrary spatial resolution and arbitrary number of exposures.

V. Conclusion

We have proposed a novel method to generate HDR images from multi-exposure inputs with unpaired datasets. The proposed method relaxes the constraints that deep learning-based methods need paired inputs and ground truth by introducing generative adversarial networks. The proposed method learns the translation between the input domain and the target domain and transforms the multi-inputs into an informative HDR output. However, generative adversarial networks obtain unclear results sometimes. We designed specific techniques to generate images with sharp edges and clear content information, including the initialization phage, the improved adversarial loss and the designed min-patch training module. Comprehensive experiments have been conducted to demonstrate the effectiveness of the proposed UPHDR-GAN.

References

[1] J. Tumblin, A. Agrawal, and R. Raskar, "Why I want a gradient camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 103–110.

[2] R. Li, S. Liu, G. Liu, and B. Zeng, "Hybrid synthesis for exposure fusion from hand-held camera inputs," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4639–4643.

[3] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1163–1170.

[4] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, B. G. Dan, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–11, Nov. 2012.

[5] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017.

[6] S. Wu, J. Xu, Y. Tai, and C. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. ECCV*, 2018, pp. 120–135.

[7] Q. Yan *et al.*, "Deep HDR imaging via a non-local network," *IEEE Trans. Image Process.*, vol. 29, pp. 4308–4322, 2020.

[8] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. H. Lau, "HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions," *IEEE Trans. Image Process.*, vol. 30, pp. 3885–3896, 2021.

[9] X. Wang *et al.*, "Multi-exposure decomposition-fusion model for high dynamic range image saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4409–4420, Dec. 2020.

[10] Y. Zhang, M. Naccari, D. Agrafiotis, M. Mrak, and D. R. Bull, "High dynamic range video compression exploiting luminance masking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 950–964, May 2016.

[11] G. Qiu, J. Duan, and G. D. Finlayson, "Learning to display high dynamic range images," *Pattern Recognit.*, vol. 40, no. 10, pp. 2641–2655, Oct. 2007.

[12] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. NIPS*, 2015, pp. 1486–1494.

[13] Z. L. Szpak, W. Chojnacki, A. Eriksson, and A. van den Hengel, "Sampson distance based joint estimation of multiple homographies with uncalibrated cameras," *Comput. Vis. Image Understand.*, vol. 125, pp. 200–213, Aug. 2014.

[14] Q. Yan *et al.*, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1751–1760.

[15] Z. Liu *et al.*, "ADNet: Attention-guided deformable convolutional network for high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 463–470.

[16] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[17] Y. Ren, Z. Ying, T. H. Li, and G. Li, "LECARM: Low-light image enhancement using the camera response model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 968–981, Apr. 2018.

[18] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107038.

[19] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4227–4240, Nov. 2021.

[20] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.

[21] Z. Yang, Y. Chen, Z. Le, and Y. Ma, "GANFuse: A novel multi-exposure image fusion method based on generative adversarial networks," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 6133–6145, Jun. 2021.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[23] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. ACM SIGGRAPH Classes (SIGGRAPH)*, 2008, pp. 369–378.

[24] A. R. Várkonyi-Kóczy, A. Rövid, and T. Hashimoto, "Gradient-based synthesized multiple exposure time color HDR image," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 8, pp. 1779–1785, Aug. 2008.

[25] N. Sun, H. Mansour, and R. Ward, "HDR image construction from multi-exposed stereo LDR images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2973–2976.

[26] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," *Comput. Graph. Forum*, vol. 28, no. 1, pp. 382–390, 2007.

[27] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 2808–2819, 2020.

[28] H. Li, T. Chan, X. Qi, and W. Xie, "Detail-preserving multi-exposure fusion with edge-preserving structural patch decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4293–4304, Nov. 2021.

[29] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3360–3374, Jun. 2022.

[30] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2017.
[31] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "The state of the art in HDR deghosting: A survey and evaluation," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 683–707, May 2015.
[32] K. Jacobs, C. Loscos, and G. Ward, "Automatic high-dynamic range image generation for dynamic scenes," *IEEE Comput. Graph. Appl.*, vol. 28, no. 2, pp. 84–93, Mar./Apr. 2008.
[33] R. Li, S. Liu, G. Liu, T. Sun, and J. Guo, "Multi-exposure photomontage with hand-held cameras," *Comput. Vis. Image Understand.*, vol. 193, Apr. 2020, Art. no. 102929.
[34] S. Raman and S. Chaudhuri, "Reconstruction of high contrast images for dynamic scenes," *Vis. Comput.*, vol. 27, no. 12, pp. 1099–1114, Dec. 2011.
[35] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free HDR reconstruction," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–10, Nov. 2013.
[36] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[37] H. Wu, S. Zhang, J. Zhang, and K. Huang, "GP-GAN: Towards realistic high-resolution image blending," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2487–2495.
[38] Y. Lu, S. Wu, Y. Tai, and C. Tang, "Image generation from sketch constraint using contextual GAN," in *Proc. ECCV*, 2018, pp. 205–220.
[39] M. Yuan and Y. Peng, "Bridge-GAN: Interpretable representation learning for text-to-image synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4258–4268, Nov. 2020.
[40] R. Li *et al.*, "SDP-GAN: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 374–385, 2021.
[41] R. Li, S. Liu, G. Wang, G. Liu, and B. Zeng, "JigsawGAN: Auxiliary learning for solving jigsaw puzzles with generative adversarial networks," *IEEE Trans. Image Process.*, vol. 31, pp. 513–524, 2022.
[42] P. Perera, M. Abavisani, and V. M. Patel, "In2I: Unsupervised multi-image-to-image translation using generative adversarial networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 140–146.
[43] D. Joo, D. Kim, and J. Kim, "Generating a fusion image: One's identity and another's shape," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1635–1643.
[44] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1982–1996, Aug. 2019.
[45] J. Huang, Z. Le, Y. Ma, X. Mei, and F. Fan, "A generative adversarial network with adaptive constraints for multi-focus image fusion," *Neural Comput. Appl.*, vol. 32, no. 18, pp. 15119–15129, Sep. 2020.
[46] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.
[47] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[48] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9465–9474.
[49] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.
[50] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.
[51] S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 626–632, May 2012.
[52] S. Paul, I. Sevcenco, and P. Agathoklis, "Multi-exposure and multi-focus image fusion in gradient domain," *J. Circuits, Syst., Comput.*, vol. 25, no. 10, pp. 1650123-1–1650123-18, 2016.
[53] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "An objective deghosting quality metric for HDR images," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 139–152, May 2016.
[54] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[56] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images," *J. Electron. Imag.*, vol. 24, no. 1, 2015, Art. no. 010501.
[57] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.
[58] R. K. Mantiuk and M. Azimi, "PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.
[59] H. Guo, S. Liu, T. He, S. Zhu, B. Zeng, and M. Gabbouj, "Joint video stitching and stabilization from moving cameras," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5491–5503, Nov. 2016.

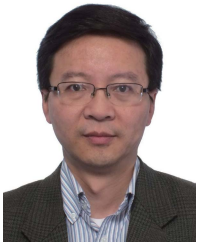
Ru Li (Student Member, IEEE) received the B.E. degree in electronic information engineering from the China University of Petroleum, Qingdao, China, in 2016. She is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. She visited the University of Oxford from July 2019 to October 2019 and worked on image enhancement. Her research interests include image processing and computer vision.

Chuan Wang received the B.Eng. degree from the University of Science and Technology of China in 2010 and the Ph.D. degree from The University of Hong Kong in 2015. He was a Staff Researcher of computer vision at Lenovo Group Ltd., Hong Kong. He was a Visiting Scholar at the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2009, and the State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou, China, in 2010. He started his training program with Megvii in 2018. His research interests include video analysis and computer vision.

Jue Wang (Senior Member, IEEE) received the B.E. and M.Sc. degrees from the Department of Automation, Tsinghua University, Beijing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2007. He was a Senior Director of Megvii from 2017 to 2020, and a Principle Research Scientist at Adobe Research. He is the Research Manager with the Visual Computing Center, Tencent AI Laboratory. His research interests include image and video processing and computational photography. He is a Senior Member of ACM. He received the Microsoft Research Fellowship and the Yang Research Award from the University of Washington in 2006.

Guanghui Liu (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2002 and 2005, respectively. He joined as a Senior Engineer at Samsung Electronics, Suwon, South Korea, in 2005. In 2009, he became an Associate Professor with the School of Electronics Engineering, UESTC, where he has been a Full Professor, since 2014, and is currently with the School of Information and Communication Engineering. He has published more than ten papers in journals or conferences, and received over 60 patents (six U.S. granted patents) in his research areas. His research interests include digital signal processing and telecommunications, with emphasis on digital video processing and transmission. He was a recipient of the Natural Science Award and the Science and Technology Progress Award, both from the Ministry of Education of China, in 2015. He was the Publication Chair of IEEE Intelligent Signal Processing and Communication Systems (ISPACS) 2010 and IEEE Visual Communications and Image Processing (VCIP) 2016.

Heng-Yu Zhang received the master's and Ph.D. degrees from the West China Medicine School, Sichuan University, in 1996 and 2010, respectively. From February 2008 to February 2009, he was a Visiting Scientist at the School of Medicine, National University of Singapore. Since 2021, he has been the Director of the West China Syncope Center, where medical images are heavily involved. He is currently a Professor and the Deputy Chief Physician of the West China Hospital with the Sichuan University, engaging cardiac pacing, electrophysiology, and pacemaker implantation with international leading level.



Bing Zeng (Fellow, IEEE) received the B.E. and M.Sc. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from the Tampere University of Technology, Tampere, Finland, in 1991. He was a Post-Doctoral Fellow at the University of Toronto from September 1991 to July 1992 and a Researcher at Concordia University from August 1992 to January 1993. He joined the Hong Kong University of Science and Technology (HKUST). After 20 years of service, he returned to the UESTC in the summer of 2013, through Chinas 1000-Talent-Scheme. He is the Leader of the Institute of Image Processing to work on image and video processing, 3D and multiview video technology, and visual big data, at UESTC. During his tenure with the HKUST and UESTC, he graduated more than 30 master's and Ph.D. students, received about 20 research grants, filed eight international patents, and published more than 250 papers. Three representing works are as follows: one paper on fast block motion estimation, published in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) in 1994, has so far been SCI-cited more than 1000 times (Google-cited more than 2300 times), and currently stands at the 7th position among all papers published in this Transactions; one paper on smart padding for arbitrarily-shaped image blocks, published in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) in 2001, leads to a patent that has been successfully licensed to companies; and one paper on directional discrete cosine transform, published in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) in 2008. He was the General Co-Chair of the IEEE Visual Communications and Image Processing (VCIP) 2016, Chengdu, in November 2016. He is currently on the Editorial Board of *Journal of Visual Communication and Image Representation* and is the General Co-Chair of PCM-2017. He was a recipient of the Best Paper Award at China-Com for three times (2009 Xi'an, 2010 Beijing, and 2012 Kunming) and the 2011 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) Transactions Best Paper Award, and the Second Class Natural Science Award (the first recipient) from the Chinese Ministry of Education in 2014. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) eight years and received the Best Associate Editor Award in 2011.

Shuaicheng Liu (Member, IEEE) received the B.E. degree from Sichuan University, Chengdu, China, in 2008, and the M.Sc. and Ph.D. degrees from the National University of Singapore, Singapore, in 2010 and 2014, respectively. In 2014, he joined the University of Electronic Science and Technology of China, where he is currently an Associate Professor with the School of Information and Communication Engineering, Institute of Image Processing, Chengdu. His research interests include computer vision and computer graphics.