Joint HDR Denoising and Fusion: A Real-World Mobile HDR Image Dataset

Shuaizheng Liu[1,2], Xindong Zhang[1,2], Lingchen Sun[1,2], Zhetong Liang[2], Hui Zeng[2], Lei Zhang[1,2*]
[1]The HongKong Polytechnic University, [2]OPPO Research
{*shuaizhengliu21, cshzeng*}*@gmail.com*; *ling-chen.sun@connect.polyu.hk*;
liangzhetong@oppo.com; {*csxdzhang, cslzhang*}*@comp.polyu.edu.hk*;

Abstract

Mobile phones have become a ubiquitous and indispensable photographing device in our daily life, while the small aperture and sensor size make mobile phones more susceptible to noise and over-saturation, resulting in low dynamic range (LDR) and low image quality. It is thus crucial to develop high dynamic range (HDR) imaging techniques for mobile phones. Unfortunately, the existing HDR image datasets are mostly constructed by DSLR cameras in daytime, limiting their applicability to the study of HDR imaging for mobile phones. In this work, we develop, for the first time to our best knowledge, an HDR image dataset by using mobile phone cameras, namely ***Mobile-HDR*** *dataset. Specifically, we utilize three mobile phone cameras to collect paired LDR-HDR images in the raw image domain, covering both daytime and nighttime scenes with different noise levels. We then propose a transformer based model with a pyramid cross-attention alignment module to aggregate highly correlated features from different exposure frames to perform joint HDR denoising and fusion. Experiments validate the advantages of our dataset and our method on mobile HDR imaging. Dataset and codes are available at* https://github.com/shuaizhengliu/Joint-HDRDN.

1. Introduction

With the rapid development of mobile communication techniques and digital imaging sensors, mobile phones have surpassed DSLR cameras and become the most prevalent device for photography in our daily life. Nonetheless, due to the low dynamic range (LDR) of mobile phone sensors [7], the captured images may lose details in dark and bright regions under challenging lighting conditions. Therefore, high dynamic range (HDR) imaging [3] is critical for improving the quality of mobile phone photography.

*Corresponding author. This work is supported by the Hong Kong RGC RIF grant (R5001-18) and the PolyU-OPPO Joint Innovation Lab.

Actually, HDR imaging has been a long standing research topic in computational photography, even for DSLR cameras. An effective and commonly used way to construct an HDR image is to fuse a stack of LDR frames with different exposure levels. If the multiple LDR frames can be well aligned (*e.g.*, in static scenes), they can be easily fused to generate the HDR image [3, 23]. Unfortunately, in dynamic scenes where there exist camera shaking and/or object motion, the fused HDR image may introduce ghost artifacts caused by inaccurate alignment [45]. Some deghosting methods have been proposed to reject pixels which can be hardly registered [12, 16]. However, precisely detecting moving pixels is challenging and rejecting too many pixels will sacrifice useful information for HDR fusion.

In the past decade, deep learning [15] has demonstrated its powerful capability to learn image priors from a rich amount of data [4]. Unfortunately, the development of deep models for HDR imaging is relatively slow, mainly due to the lack of suitable training datasets. Kalantari *et al*. [10] built the first dataset with LDR-HDR image pairs by DSLR cameras in daytime. Benefiting from this dataset, many deep learning algorithms have been proposed for HDR imaging. Some works [10] employ the convolutional neural network (CNN) for fusion after aligning multiple frames with optical flow [18], which is however unreliable under occlusion and large motions. Subsequent works resort to employing various networks to directly reconstruct the HDR image from LDR frames. Liu *et al*. [19] developed a deformable convolution based module to align the features of input frames. Yan *et al*. [40] proposed a spatial attention mechanism to suppress undesired features and employed a dilated convolution network [42] for frame fusion. Following this spatial attention mechanism, some fusion networks have been developed with larger receptive fields, such as non-local networks [41] and Transformer networks [21].

Though the dataset developed in [10] has largely facilitated the research of deep learning on HDR imaging, it is not well suited for the investigation of HDR imaging techniques for mobile phone cameras. Firstly, due to the small aperture and sensor size, images captured by mobile phones

are more susceptible to noise than DSLR cameras, especially in nighttime. However, the images in dataset [10] are generally very clean since they are collected by DSLR cameras in daytime. Compared with DSLR cameras, the normal exposed frame of mobile phone cameras contains stronger noise, which should be reduced by fusing with other frames. Secondly, mobile phone cameras often has fewer recording bits (12 bit) than DSLR (14 bit), resulting in larger overexposure areas in the reference frame. Therefore, mobile HDR imaging is a more challenging problem, and new dataset and new solutions are demanded.

To address the above limitations of existing HDR datasets and facilitate the research on real-world mobile HDR imaging, we establish a new HDR dataset, namely Mobile-HDR, by using mobile phone cameras. Specifically, we utilize three mobile phones to collect LDR-HDR image pairs in raw image domain, covering both daytime and nighttime scenes with different noise levels. In order to obtain high-quality ground truth of HDR images, we first collect noise-free LDR images under each exposure by multi-frame averaging, and then synthesize the ground truth HDR image by fusing the generated clean LDR frames. For dynamic scenes with object motion, we follow [10] to first capture multiple exposed frames from static scenes to synthesize the ground truth HDR images, and then replace the non-reference frames with the images captured in dynamic scenes as input. To our best knowledge, this is the first mobile HDR dataset with paired training data.

With the established dataset, we propose a new transformer based model for joint HDR denoising and fusion. To enhance denoising and achieve alignment, we design a pyramid cross-attention module to implicitly align and fuse input features. The cross-attention operation enables searching and aggregating highly correlated features from different frames, while the pyramid structure facilitates the feature alignment under severe noise, large overexposure and large motion. A transformer module is then applied to fuse the aligned features for HDR image recovery.

The contributions of our work can be summarized as follows. First, we build the first mobile HDR dataset with LDR-HDR image pairs under various scenes. Second, we propose a cross-attention based alignment module to perform effective joint HDR denoising and fusion. Third, we perform extensive experiment to validate the advantages of our dataset and model. Our work provides a new platform for researchers to investigate and evaluate real-world mobile HDR imaging techniques.

2. Related work

HDR Image Datasets. Datasets are the cornerstone of algorithm development and evaluation. Before the era of deep learning, Sen *et al*. [33] and Tursun *et al*. [36] provided 8 and 16 scenes of real-world HDR data without ground-truth HDR images, respectively, for qualitative evaluation and comparison of different algorithms. In [10], Kalantari *et al*. proposed the first paired LDR-HDR dataset, including 74 training and 15 test pairs, making the learning of deep HDR models possible. Prabhakar *et al*. [28] later built a dataset with 582 LDR-HDR pairs. These two datasets regard the medium-exposed image as the reference frame. In order to explore the cases when other exposures should be used as the reference, Li *et al*. [17] collected a dataset where different LDR frames can be taken as the reference frame, but it is not publicly available.

All the above datasets are collected by DSLR cameras, and they are unsuitable for investigating mobile HDR imaging methods due to the different characteristics of camera sensors and lens, especially for nighttime scenes with strong noise. In order to facilitate the development of mobile HDR imaging techniques, we construct an HDR dataset using mobile phones, covering different scenes and noise levels.

HDR Image Reconstruction. In case the LDR frames can be strictly aligned, the HDR image can be easily obtained by fusing them with different weight functions [3, 23]. In practice, however, ghost artifacts can be generated by camera motion and subject moving. A number of methods have been proposed for HDR deghosting in dynamic scenes. Early works can be divided into two classes. The first class aligns LDR frames and fuses them to an HDR image. The global rigid alignment by translation or homography [35, 38] is simple to use but can fail to handle the foreground motion. Bogoni *et al*. [2] and Kang *et al*. [11] utilized optical flow to deal with moving objects, which are not robust to occlusion, large motion and saturated areas. Some works [8, 33, 43, 44] perform patch-based registration, which are more robust to motion but suffer from heavy computation. The other class of methods detect inconsistent pixels and discard them after global alignment, such as local entropy [9], color consistency [5, 6, 31], median threshold bitmaps [26], and rank minimization [16, 25]. However, the accurate detection of such pixels is difficult and the rejection strategy may lose much useful information for fusion.

Recently, with the availability of paired LDR-HDR image dataset [10], deep learning based HDR reconstruction methods have been developed. Kalantari *et al*. [10] and Prabhakar *et al*. [27, 28] employed optical flow [18] or flow net [34] to align input frames and utilized CNN to merge them. Wu *et al*. [39] used an encoder-decoder network to synthesize HDR image directly from input frames without alignment. Liu *et al*. and Pu *et al*. [30] adopted deformable convolution to align features implicitly. Yan *et al*. [40] designed a spatial attention module to detect unaligned area to suppress the ghosting artifacts. Following this spatial attention module, various CNNs with large receptive fields [40, 41] have been developed. Since transformers could better model long-range dependencies than CNN, Liu *et*

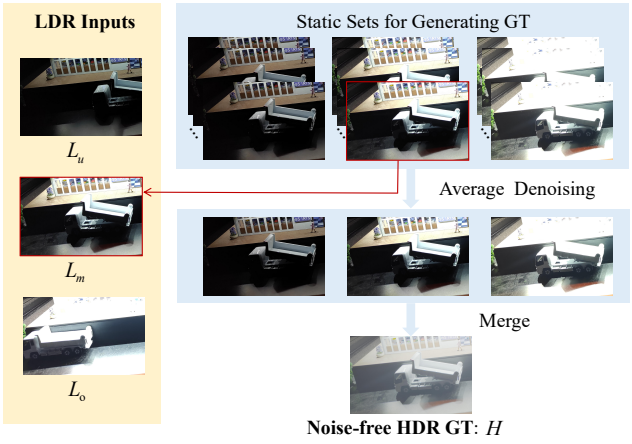


Figure 1. The synthesis process of LDR-HDR data pairs for dynamic scenes. We first keep the foreground object still to capture three static sets of successive raw images with under-, middle- and over-exposures, and use them to generate three noise-free LDR images by average denoising, which are merged to synthesize the noise-free HDR image as ground-truth (GT). One middle-exposure frame, denoted by L_m, is extracted from the corresponding static set. We then move the foreground object and tripods to capture another two LDR frames, denoted by L_u and L_o, with under- and over-exposures, respectively. The final LDR inputs are composed of L_u, L_m and L_o. The LDR frames and synthesized HDR image are visualized through a simple ISP pipeline.

al. [21] proposed a transformer network for HDR fusion and achieved state-of-the-art results. Meanwhile, there were some attempts to utilize GAN networks [24] and few-shot learning [29] to hallucinate HDR details or relieve the dependency on abundant training data.

Most of the existing methods are developed on images collected by DSLR cameras, which may not fit well to mobile phone images. Recently, Lecouat *et al*. [14] proposed to perform joint HDR and super-resolution with mobile raw burst images. However, they utilized synthetic raw images for training, resulting in color halos in the saturated area and failing to deal with large motion. In this work, we construct a real-world mobile HDR image dataset and propose a pyramid cross-attention module to achieve alignment against noise, saturated area and large motion.

3. The Established Dataset

To facilitate the research on mobile HDR imaging, we establish a paired LDR-HDR image dataset in raw image domain by using mobile phone cameras. Specifically, we utilized four mobile phones equipped with three types of mobile sensors (IMX586, IMX766 and IMX800) to capture images under different exposures and lighting conditions, including indoor, outdoor, daytime and nighttime scenes. The ISO settings in our dataset range from 100 to 6400, covering a variety of noise levels. Our dataset is composed of three subsets: a subset of static scenes with ground-truth (GT) HDR images, a subset of dynamic scenes with GT HDR images, and a subset of dynamic scenes without GT HDR images (used only for visual comparison).

For the subset of static scenes, we capture LDR sequences with three exposures (*i.e*., under-, middle- and over-exposures) by a mobile phone fixed on tripod with a customized app, which detects and removes the defective pixels in each shot. Under each exposure, we take 120 to 400 successive images to facilitate denoising. Generally, the number of shots increases with the increase of ISO and/or the decrease of exposure. We then average the shots to obtain the noise-free LDR image for each exposure. After the noise-free LDR frames of the three exposures are acquired, we merge them using the weighting function proposed in [3] to generate the high-quality HDR image as GT, denoted by H. Then three LDR frames, denoted by L_u, L_m, L_o, are extracted from the captured successive LDR images with under-, middle- and over-exposures, respectively, as the LDR inputs, building the LDR-HDR data pairs. We check the quality of each sequence and discard the outliers. Finally, 136 static scenes are collected, including 49 daytime and 87 nighttime scenes.

For the subset of dynamic scenes with foreground object motion, we utilize controllable objects to simulate the motion between LDR frames, following the strategy in [10]. The process is illustrated in Fig. 1. We first keep the object still and capture three static sets of images with three exposures, and synthesize the noise-free HDR GT image H of this scene by the method applied in static scenes. Meanwhile, we extract one middle-exposure LDR frame L_m from the static set as one of the LDR inputs. Then we move the object and tripod to capture an under-exposure LDR frame L_u and an over-exposure LDR frame L_o. Finally, L_u, L_m, L_o and H are taken as the LDR inputs and HDR GT, respectively. In total, we collected 115 dynamic scenes, including 15 daytime and 100 nighttime ones.

The above subsets of static and dynamic scenes with HDR GT can be used to train HDR reconstruction models and evaluate them quantitatively. In addition, we capture 30 scenes without HDR GT, which contain uncontrolled moving or static objects captured by hand-held mobile phones, for qualitative evaluation of different models.

We compare the statistics of our Mobile-HDR dataset and the Sig17 dataset [10] in Table 1. One can clearly see that our dataset covers more diverse real-world scenarios than the Sig17 dataset. Unlike Sig17, which only contains daytime dynamic scenes, our dataset covers both daytime and nighttime, dynamic and static scenes, representing the diverse lighting conditions and various noise levels in practical scenarios. In addition, the image resolution (4K in general) of our dataset is much higher than that in Sig17 (1500×1000). Fig. 2 show some typical scenes from our

Table 1. The statistics comparison between Sig17 dataset [10] and our Mobile-HDR dataset.

Data	Camera Sensor	Image Resolution	Dynamic Scenes w/ HDR GT		Static Scenes w/ HDR GT		Dynamic Scenes w/o HDR GT		Total
			daytime	nighttime	daytime	nighttime	daytime	nighttime	
Sig17 [10]	Canon EOS-5D Mark II	1500×1000	89	None	None	None	None	None	89
Mobile-HDR (ours)	OPPO A95 OPPO FindX3 Pro HONOR 70 Pro	4000×3000 4608×3456 4096×3072	15	100	49	87	10	20	281

Figure 2. Sample LDR frames in our Mobile-HDR dateset. Various noise levels can be observed in the zoomed-in regions. The images are visualized through a simple ISP pipeline.

dataset. One can see the strong noise in dark areas and the severely over-exposed regions in the LDR images.

4. Method

Overview. Different from HDR reconstruction on DSLR data, where the normally exposed area of the reference frame can be directly used for HDR recovery, the normally exposed areas in the reference frame of mobile camera data still have strong noise, which needs to be suppressed by fusing with other frames. In addition, the larger overexposed areas in mobile camera data rely on information from underexposed frames to recover detail. Without an effective alignment module, ghosting artifacts will become severe in mobile HDR images. We therefore propose a transformer based network with a novel pyramid cross-attention alignment module to aggregate and align the correlated features from LDR frames more effectively, achieving denoising and HDR reconstruction jointly.

Given the set of noisy LDR frames of three exposures $\{L_1, L_2, L_3\}$ (sorted by their exposure times), we aim to reconstruct the noise-free HDR frame H. The middle exposure frame L_2 is regraded as the reference frame, so the estimated noise-free HDR $\widehat{H}$ should be consistent with L_2 in structure but contain the dynamic range information from all frames. Since LDR images in our dataset are in the RAW format, they have linear response curve with ambient lighting. So we do not need to linearize the LDR images by using the camera response function (CRF) or gamma correction. In order to facilitate the alignment, we map the input LDR images $\{L_i\}$ to the domain of brightness constancy based on the exposure time to get the corresponding set of $\{H_i\}$:

$$H_i = \frac{L_i}{t_i}, \forall i = 1, 2, 3, \tag{1}$$

where t_i denotes the exposure time of the image L_i. Following [40], images I_i and H_i are concatenated along the channel dimension to obtain the tensors $X_i = [L_i, H_i], i = 1, 2, 3$ as the network input. The network outputs the estimated noise-free HDR image $\widehat{H}$.

Fig. 3(a) illustrates the overall architecture of our proposed model. It mainly consists of three components, *i.e.*, the pyramid cross-attention alignment module for aligning neighborhood frames to the reference frame, the attention fusion module for fusing aligned features and the merging subnet for final HDR reconstruction. For each input tensor $X_i, i = 1, 2, 3$, we first extract the shallow features F_i by convolution layers. Then, a pyramid cross-attention alignment module is used to align non-reference features $F_i, i = 1, 3$ to reference features F_2. A local skip connection is used for better training. The aligned features $\widetilde{F_1}$, $\widetilde{F_3}$ and F_2 are fed into the attention fusion module to get the fused features. Finally, a merging subnet, which consists of context-aware transformer blocks, takes the fused features as input to generate the HDR image. A global residual connection is used to accelerate the training process.

Pyramid Cross Attention Alignment Module. We propose a pyramid cross attention alignment module to align features from neighborhood frames to the reference frame. Since the computation of cross attention will also aggregate correlated features, the cross attention facilitates the alignment and denoising at the same time.

Given the neighbor-frame features $F_i, i = 1, 3$ and reference-frame features F_2 of the same size $H \times W \times C$, we first partition them into non-overlapping $M \times M$ local windows to get two reshaped inputs of size $\frac{HW}{M^2} \times M^2 \times C$, where $\frac{HW}{M^2}$ is the total number of windows. Then we compute the cross-attention separately for each window. For the local window feature from neighbor-frame $F_i \in \mathbb{R}^{M^2 \times C}, i = 1, 3$ and the ones from reference-frame $F_2 \in \mathbb{R}^{M^2 \times C}$, the query, key, and value matrices Q, K and

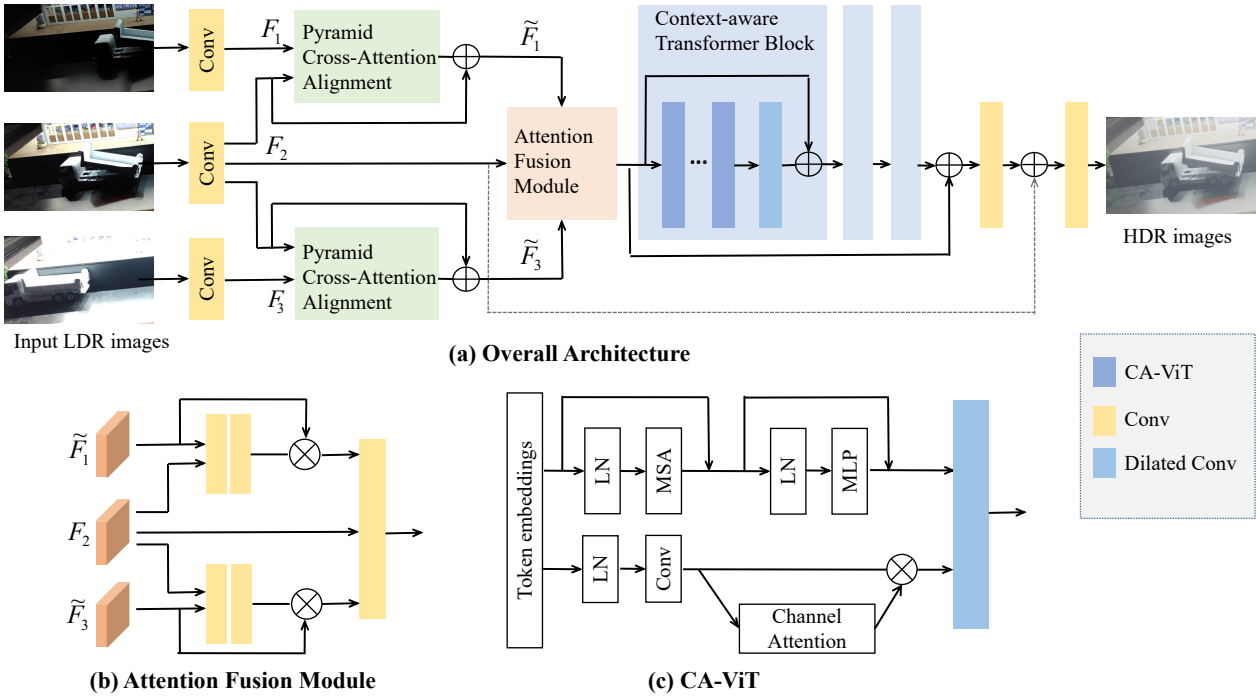


Figure 3. (a) The overall architecture of the proposed joint HDR denoising and fusion model. (b) The structure of attention fusion module. (c) The structure of Context-Aware Vision Transformer (CA-ViT) module used in Context-Aware Transformer Block.

V are computed as:

$$Q = F_2 P_Q, K = F_i P_K, V = F_i P_V, \forall i = 1, 3, \quad (2)$$

where P_Q, P_K, P_V are projection matrices shared across different windows. The attention matrix is thus computed in a local window as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V, \quad (3)$$

where B is the learnable relative positional encoding. Following [20], the attention function are performed for h times in parallel and the result are concatenated for multi-head cross attention.

In order to handle complex motion, we adopt the pyramidal processing and cascading operation like the the PCD alignment module [37] used in video super-resolution. Considering that the saturated regions or severely noisy regions in the reference image are generally difficult to perform reliable feature matching, we propose an attention transfer mechanism. Intuitively, if the size of query patch is enlarged, the patch may include some region with details, which will enable more reliable matching. Since features at coarse scales are extracted from larger receptive fields, we perform feature matching in a coarse scale and transfer the attention coefficients to finer scales.

The proposed pyramid cross-attention module is illustrated in Fig. 4(a). We generate an L-level pyramid of feature representation for each LDR frame. Given the features F_i^l at l-level, we use strided convolution filters to get the downsampled features with a factor of 2 at the $(l+1)$-th pyramid level. At the l-th level, cross attention is performed on reference feature F_2^l and neighborhood features $F_i^l, i = 1, 3$ to get features F_i^{lc}. The attention coefficient A_i^{l+1} computed by F_2^{l+1} and F_i^{l+1} from the upper $(l+1)$-th level is multiplied with neighborhood feature F_i^l to get F_i^{lt}. The specific computation process is illustrated in Fig. 4(b). Finally, the aligned features at the l-th level $\widetilde{F}_i^l$ are predicted by using F_i^{lc}, F_i^{lt}, and $\times 2$ upsampled aligned features from the the upper $(l+1)$-th level $\widetilde{F}_i^{l+1}$ as follows:

$$\widetilde{F}_i^l = \text{Conv}(F_i^{lc}, F_i^{lt}, (\widetilde{F}_i^{l+1})^{\uparrow 2}). \quad (4)$$

Attention Fusion Module. After obtaining the aligned features, we adopt the attention module proposed in [40] to suppress harmful features from misaligned, over-exposed and under-exposed areas, as illustrated in Fig. 3(b). For each aligned feature from non-reference LDR image (*i.e.*, $\widetilde{F_1}$ and $\widetilde{F_3}$), we concatenate it with the reference feature F_2 as the input of two convolutional layers, generating a spatial attention map $m_i, i = 1, 3$ ranging between 0 and 1. We then perform the element-wise multiplication of m_i and $\widetilde{F}_i$ to get the attentioned features F_i':

$$F_i' = m_i \odot \widetilde{F}_i, \; i = 1, 3. \quad (5)$$

The features F_1', F_2, F_3' are concatenated and passed through a convolution layer to obtain the fused features.

Merging Network. The merging network is composed of several context-aware Transformer blocks (CTB) [21], which consists of a dual-branch context aware vision Transformer (CA-ViT) and a dilated convolution layer, as illustrated in Fig. 3(a). The structure of CA-ViT is shown in Fig. 3(c), which employs a window-based multi-head Transformer encoder [20] to extract globally long-range features, and a convolution block with channel attention as another parallel branch to capture the local information.

Loss Function. The proposed model outputs the estimated HDR image $\widehat{H}$ within range [0, 1]. If the loss is applied to $\widehat{H}$ and ground truth H directly, the training will be dominated by the brighter areas, hindering the restoration of dark areas. So we apply the μ-law tone-mapping function to the HDR image in HDR domain:

$$\mathcal{T}(H) = \frac{\log(1+\mu H)}{\log(1+\mu)}, \tag{6}$$

where $\mathcal{T}(H)$ is the tone-mapped HDR image and μ is set to 5000 in our work. Given the estimated HDR image $\widehat{H}$ and the ground truth HDR image H, we utilize the ℓ_1 loss in tone-mapped domain to optimize the network:

$$\mathcal{L} = \|\mathcal{T}(\widehat{H}) - \mathcal{T}(H)\|_1. \tag{7}$$

5. Experiments

Training Data Preparation. Since there is a lack of paired HDR datasets captured by mobile phones, we constructed such a paired dataset, *i.e.*, Mobile-HDR, in this work for HDR model training and evaluation. We apply black level correction and range normalization to the raw data to obtain each L_i, and conduct joint HDR denoising and fusion in a raw-in-raw-out manner. For static scenes, we add random global motions to the non-reference frames, *i.e.*, random translation in the range of [0,20] pixels. We divide our Mobile-HDR dataset with ground-truth into 223 training samples and 28 test samples. For the training samples, 102 samples are taken from dynamic scenes and 121 from static scenes. While for the test samples, 13 samples are taken from dynamic scenes and 15 samples are taken from static scenes. Meanwhile, there are 30 test samples without ground-truth for visual comparison. Each sample is composed of three LDR frames with exposure values {-2,0,2} or {-3,0,3} and the corresponding HDR frame. Before training, we crop the images into 512×512 patches with stride 200. During training, we randomly crop 128×128 regions from the 512×512 patch as training samples.

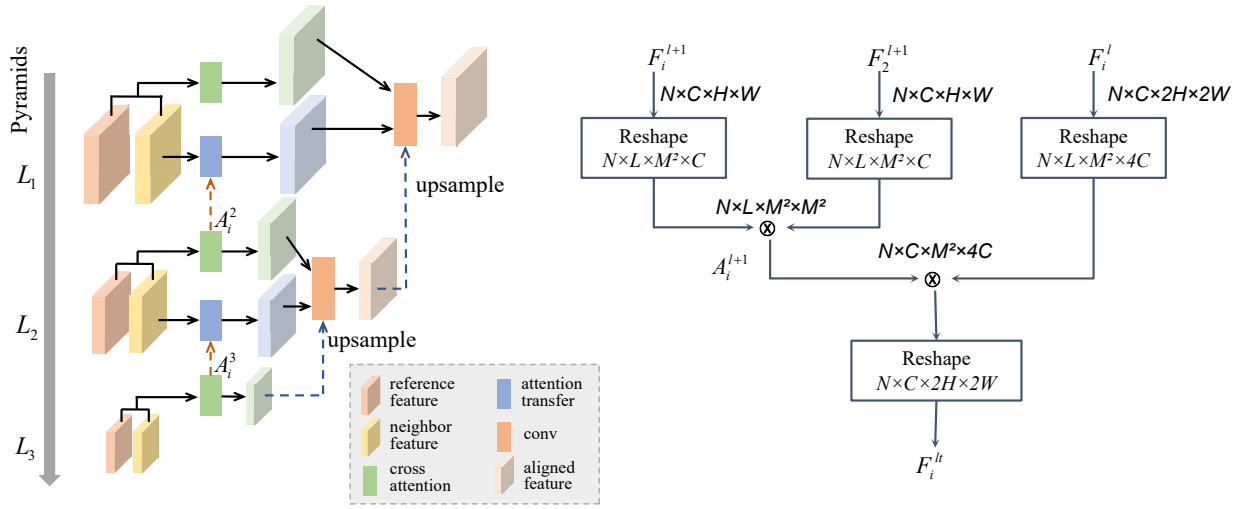
Implement Details. We call our method as Joint-HDRDN for that it performs HDR denoising and fusion jointly. The overall model is optimized by the Adam optimizer [13] with default parameters. The batch size is set as 16, and the initial learning rate is 2e-4 and halved after 500 epochs. Our pyramid cross attention alignment module adopts 3-layer pyramid with partition window size M of 8. The number of channels is set as 60, and there are 3 context-aware transformer blocks in our merging subnet. Different from the HDR-Tranformer that employs 6 CA-ViT in each context-aware transformer block, our Joint-HDRDN has only 4 CA-ViT in each block. Benefiting from our proposed pyramid cross-attention alignment module, our merging network does not need to stack many transformer blocks to enlarge the receptive field. The whole training is conducted on four NVIDIA V100 GPUs and costs about three days to converge.

Evaluation Metrics. We use PSNR and SSIM in both raw domain and sRGB domain, as well as HDR-VDP-2 in sRGB domain [22], as evaluation metrics. The HDR results in sRGB domain is obtained by passing the HDR results in raw domain through a simple ISP pipeline as in SIDD [1], which involves white balance, demosaicking, color correction and sRGB space transfer. The parameters are from the metadata of the reference frame. For both raw domain and sRGB domain, PSNR and SSIM are evaluated in both linear domain (*i.e.*, PSNR-l and SSIM-μ) and the tone-mapped domain with μ-law (*i.e.* SSIM-l and SSIM-μ). Moreover, since HDR-VDP-2 is developed specifically for qualitative evaluation of HDR images, we compute it in sRGB domain.

Comparison with State-of-the-Arts. We compare our proposed Joint-HDRDN method with state-of-the-art HDR reconstruction methods, including DeepHDR [39], AHDRNet [40], NHDRRNet [41] and HDR-Transformer [21]. For fair comparison, we retrain these deep HDR models on our training dataset and then evaluate them on our test dataset.

Table 2 compares the quantitative results of competing methods. It can be observed that transformer based algorithms outperform CNN based methods, while our proposed Joint-HDRDN surpasses HDR-Transformer, which is the previous state-of-the-art, by up to 0.38dB and 0.95dB in terms of PSNR-μ and PSNR-l in raw domain. Furthermore, after rendering images from raw domain to sRGB domain, our model still outperforms other competitors by a large margin, which demonstrates the effectiveness of our strategy of performing HDR fusion and denoising jointly by using pyramid cross-attention.

Fig. 5 compares the visual results of our method and its competitors on some challenging scenarios in our dataset. All HDR results are first passed through a simple ISP and then tone-mapped by the Reinhard [32] operator. It can be seen that our method achieves significantly better visual results. On the reference frames which have high-exposure or severe noise, our method can recover fine details without introducing much artifacts. In comparison, the other methods



(a) Pyramid Cross-Attention Alignment Module **(b) The Computation of Attention Transfer**

Figure 4. (a) The structure of pyramid cross-attention alignment module. (b) The attention transfer mechanism between two scales.

(a) Scene 1 from our test dataset. (b) Scene 2 from our test dataset.

Figure 5. Visual comparison between our Joint-HDRDN and other state-of-the-arts HDR reconstruction methods on two scenes from our test dataset. All images go through a simple ISP pipeline for visualization. The HDR results are tone-mapped for better comparison.

suffer from the ghosting artifacts or residual noise. Previous methods don't consider the impact of noise on the final HDR image quality. Meanwhile, they usually just resort to modeling long-range dependency to hallucinate reasonable content for over-exposed areas and attention module to suppress unaligned areas to alleviate ghost artifacts. So they do not make efficient use of other frames to recover details. Additionally, they will produce ghosting artifacts inevitably due to the lack of specific alignment design, especially for the large over-exposed areas. In contrast, our proposed pyramid cross-attention alignment module searches and aggregates beneficial features from other frames more effectively, which can better reproduce the details and alleviate the artifacts. More visual comparison examples can be found in the **supplementary file**.

Ablation Study on Training Dataset. To further demonstrate the necessity of constructing our Mobile-HDR dataset to develop mobile HDR techniques, we train the HDR-Transformer [21] and our Joint-HDRDN models on the DSLR camera captured Sig17 dataset [10] and our Mobile-HDR dataset, respectively, and evaluate them on our test dataset, which consists of data captured by mobile phone cameras. Since Sig17 provides the demosaiced data, we compare the results in the sRGB domain.

The quantitative results are shown in Table 3, which verifies that the DSLR dataset cannot well support the research

Table 2. Quantitative comparison of proposed Joint-HDRDN method with state-of-the-art HDR reconstruction methods.

Methods	Raw				sRGB				
	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP-2
DeepHDR [39]	37.94	35.29	0.9559	0.9869	34.91	36.43	0.9537	0.9799	44.23
AHDRNet [40]	38.16	35.49	0.9571	0.9872	34.85	36.64	0.9524	0.9803	44.98
NHDRRNet [41]	37.57	34.54	0.9517	0.9850	34.27	36.04	0.946	0.9774	44.81
HDR-Transformer [21]	38.54	35.42	0.9599	0.9873	35.50	36.67	0.9572	0.9809	45.60
Joint-HDRDN (ours)	38.92	36.37	0.9616	0.9888	35.70	37.46	0.9588	0.9825	45.96

Table 3. Quantitative comparison between models trained on different training sets.

	DSLR training set		Our training set	
	HDR-Transformer [21]	Ours	HDR-Transformer [21]	Ours
PSNR-l(sRGB)	23.59	23.45	36.67	37.46
SSIM-l(sRGB)	0.4276	0.4134	0.9809	0.9825

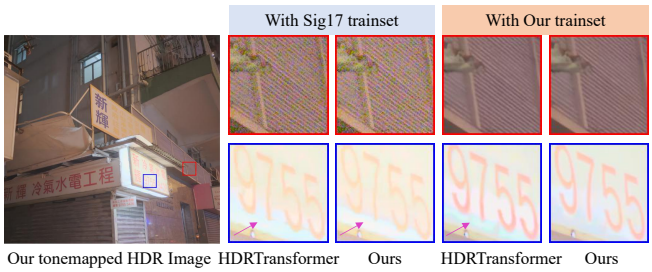


Figure 6. Visual comparison between models trained on DSLR dataset [10] and our Mobile-HDR dataset.

of HDR imaging on mobile phones. This is mainly because the mobile phone images have stronger noise and larger overexposed areas caused by the smaller aperture and sensor size. As shown in Fig. 6, models trained on the DSLR dataset are hard to remove the heavy noise in mobile phone images, and the over-exposed areas have obvious ghost artifacts. In addition, existing methods such as HDRTransformer are not designed for mobile HDR imaging. Even retrained on our Mobile-HDR datasset, they still show many ghosting artifacts in over-exposed areas and blurry details in noisy regions. Therefore, it is necessary to construct a mobile HDR image dataset to facilitate the research on mobile HDR imaging, such as more effective denoising, alignment, fusion and the joint tasks of them.

Ablation Study on Network. In order to validate the effectiveness of different components in our Joint-HDRDN network, we evaluate the following variants of our model:

- Baseline. We replace our pyramid cross-attention alignment module with the attention feature extractor adopted by HDR-Transformer and AHDRNet, while keeping the merging network unchanged. That is, the baseline shares the same components as HDR-Transformer but has fewer CA-ViT blocks (12 vs. 18).

- w/o Attention Transfer. This variant removes the attention transfer mechanism adopted in our pyramid cross-attention alignment module.

- w/o Attention Fusion. This variant removes the attention fusion module and directly stacks the aligned features as the fused features.

Table 4. Quantitative results of the ablation studies.

Method	PSNR-μ(raw)	PSNR-l(raw)	HDR-VDP-2
Baseline	38.41	35.26	45.42
w/o Attention Transfer	38.81	36.00	45.68
w/o Attention Fusion	38.71	36.20	45.72
Our full model	38.92	36.37	45.96

Table 4 lists the quantitative results of our ablation study. Compared with the baseline, which shares the same merging network as our full model but removes the alignment module, the full model achieves 0.51dB and 1.11dB advantages in PSNR-μ and PSNR-l, respectively, demonstrating the effectiveness of our proposed pyramid cross-attention alignment module. The attention transfer mechanism is proposed to alleviate the difficulties in aligning with over-exposed reference regions. As shown in the table, if we remove the attention transfer from our alignment module, the performance will drop by 0.37dB in PSNR-l, validating the roles of attention transfer mechanism. The attention fusion module is also useful since the PSNR-μ will drop by 0.21dB if we remove it.

6. Conclusion

We established, for the first time to our best knowledge, a real-world mobile HDR image dataset, namely Mobile-HDR, to facilitate researches on mobile HDR imaging. Different from the existing HDR image datasets, which were mostly collected in daytime with DSLR cameras, our dataset was collected by mobile phone cameras under different lighting conditions and scenes, which contained stronger noises and larger over-exposed areas. We consequently developed a new HDR image reconstruction network, namely Joint-HDRDN, which employed a novel pyramid cross-attention alignment module to perform HDR fusion and denoising jointly. Extensive experiments validated the effectiveness of our proposed dataset and model.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 6

[2] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 7–12. IEEE, 2000. 2

[3] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 1, 2, 3

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 1

[5] Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. In *2009 IEEE International conference on computational photography (ICCP)*, pages 1–7. IEEE, 2009. 2

[6] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen*, pages 277–284, 2006. 2

[7] 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. 1

[8] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1163–1170, 2013. 2

[9] Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28(2):84–93, 2008. 2

[10] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 1, 2, 3, 4, 7, 8

[11] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003. 2

[12] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *2006 International Conference on Image Processing*, pages 2005–2008. IEEE, 2006. 1

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. High dynamic range and super-resolution from raw image bursts. *arXiv preprint arXiv:2207.14671*, 2022. 3

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[16] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters*, 21(9):1045–1049, 2014. 1, 2

[17] Wei Li, Shuai Xiao, Tianhong Dai, Shanxin Yuan, Tao Wang, Cheng Li, and Fenglong Song. Sj-hdˆ 2r: Selective joint high dynamic range and denoising imaging for dynamic scenes. *arXiv preprint arXiv:2206.09611*, 2022. 2

[18] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 1, 2

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 463–470, 2021. 1

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5, 6

[21] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 344–360. Springer, 2022. 1, 3, 6, 7, 8

[22] 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. 6

[23] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*, volume 28, pages 161–171. Wiley Online Library, 2009. 1, 2

[24] 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. 3

[25] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1219–1232, 2014. 2

[26] Fabrizio Pece and Jan Kautz. Bitmap movement detection: Hdr for dynamic scenes. In *2010 Conference on Visual Media Production*, pages 1–8. IEEE, 2010. 2

[27] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution hdr deghosting with cnn. In *European Conference on Computer Vision*, pages 497–513. Springer, 2020. 2

[28] K Ram Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, and R Venkatesh Babu. A fast, scalable, and reliable deghosting method for extreme exposure fusion. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019. 2

[29] K Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled

from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4875–4885, 2021. 3

[30] Zhiyuan Pu, Peiyao Guo, M Salman Asif, and Zhan Ma. Robust high dynamic range (hdr) imaging with complex motion and parallax. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[31] Shanmuganathan Raman and Subhasis Chaudhuri. Reconstruction of high contrast images for dynamic scenes. *The Visual Computer*, 27(12):1099–1114, 2011. 2

[32] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 267–276, 2002. 6

[33] 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. 2

[34] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[35] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. 2007. 2

[36] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for hdr images. In *Computer Graphics Forum*, volume 35, pages 139–152. Wiley Online Library, 2016. 2

[37] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5

[38] Greg Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of graphics tools*, 8(2):17–30, 2003. 2

[39] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 6, 8

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 1, 2, 4, 5, 6, 8

[41] 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. 1, 2, 6, 8

[42] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1

[43] Jinghong Zheng and Zhengguo Li. Superpixel based patch match for differently exposed images with moving objects and camera movements. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4516–4520. IEEE, 2015. 2

[44] Jinghong Zheng, Zhengguo Li, Zijian Zhu, Shiqian Wu, and Susanto Rahardja. Hybrid patching for a sequence of differently exposed images with moving objects. *IEEE Transactions on Image Processing*, 22(12):5190–5201, 2013. 2

[45] Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, volume 30, pages 405–414. Wiley Online Library, 2011. 1