# Robust High Dynamic Range (HDR) Imaging with Complex Motion and Parallax

Zhiyuan Pu[1][0000−0002−3072−1306], Peiyao Guo[1][0000−0003−2887−3463], M. Salman Asif[2][0000−0001−5993−3903], and Zhan Ma[1(✉)][0000−0003−3686−4057]

[1] Nanjing University, Nanjing, China
{zhiyuanpu,peiyao}@smail.nju.edu.cn, mazhan@nju.edu.cn
[2] University of California, Riverside, CA, USA
sasif@ece.ucr.edu

**Abstract.** High dynamic range (HDR) imaging is widely used in consumer photography, computer game rendering, autonomous driving, and surveillance systems. Reconstructing ghosting-free HDR images of dynamic scenes from a set of multi-exposure images is a challenging task, especially with large object motion, disparity, and occlusions, leading to visible artifacts using existing methods. In this paper, we propose a Pyramidal Alignment and Masked merging network (PAMnet) that learns to synthesize HDR images from input low dynamic range (LDR) images in an end-to-end manner. Instead of aligning under/overexposed images to the reference view directly in pixel-domain, we apply deformable convolutions across multiscale features for pyramidal alignment. Aligned features offer more flexibility to refine the inevitable misalignment for subsequent merging network without reconstructing the aligned image explicitly. To make full use of aligned features, we use dilated dense residual blocks with squeeze-and-excitation (SE) attention. Such attention mechanism effectively helps to remove redundant information and suppress misaligned features. Additional mask-based weighting is further employed to refine the HDR reconstruction, which offers better image quality and sharp local details. Experiments demonstrate that PAMnet can produce ghosting-free HDR results in the presence of large disparity and motion. We present extensive comparative studies using several popular datasets to demonstrate superior quality compared to the state-of-the-art algorithms.

## 1 Introduction

Human visual system has astounding capabilities to capture natural scenes with high dynamic range [1]. In recent years, significant efforts have been made to develop specialized high dynamic range (HDR) imaging sensors, such as using beam splitters [2, 3] or spatially varying exposed pixels [4]. Most common approaches for HDR imaging still rely on capturing and fusing multi-exposure images with cost-efficient, low dynamic range (LDR) sensors.

The multi-exposure fusion schemes input a sequence of LDR images captured at different exposures and apply a variety of computational methods to construct

ghosting-free HDR images. Images with different exposures can be captured in two possible options: (1) Using a single camera by adjusting its exposure over time to capture a set of images. (2) Using a camera array (e.g., in a multi-camera system) in which each camera is set to a different exposure to capture a set of images simultaneously. Images captured by the first approach often contain object motion, while the parallax effects are inevitable for multi-camera, multi-exposure setup alternatively. Thus, effectively removing artifacts caused by motion or parallax is the main challenge for high-quality, ghosting-free HDR imaging.

Over past several years, numerous efforts have been made to reduce the ghosting artifacts for HDR image generation. For example, the popular exposure bracketing-based weighted fusion is enhanced with motion detection and displaced pixel rejection [5, 6] to alleviate motion-induced artifacts. Its performance, however, heavily depends on the accuracy of motion detection algorithm. On the other hand, alignment-before-merging schemes have been proposed to align input LDR images to a reference view, and then merge them altogether for HDR image reconstruction [7]. Inspired by recent advancements in deep neural networks, a large amount of learning-based approaches have also been introduced for HDR imaging. The method in [8] performs optical flow-based image alignment followed by a convolutional neural network (CNN)-based merging process. However, aligning images in pixel domain is often prone to the noisy or saturated pixels-induced misalignment, which leads to visible artifacts in final synthesized presentation. End-to-end learning-based approaches such as [9, 10] without implicitly alignment directly feed LDR images into a network to reconstruct HDR images, failing to deal with scenarios with complex motion or large disparity.

In this work, we present a robust HDR imaging system along the alignment-before-merging direction, where *alignment* and *merging* network models are carefully designed to efficiently resolve the ghosting problem that arises due to temporal motion or spatial disparity-induced displacement. In *alignment* network, we use *feature-domain* processing to replace existing pixel-domain solutions, where a deformable convolution-based network is applied on three input LDR images to generate multiscale features for subsequent *pyramidal alignment*. Aligned features are then fed into a *merging network* for synthesizing the final HDR output. The merging process includes dilated dense blocks with squeeze-and-excitation (SE) attention modules and adaptive mask-based weighting by which feature redundancy and misalignment are efficiently removed. This scheme preserves local details and provides better image quality. Such feature-domain pyramidal alignment and masked merging networks (PAMnet) are trained in an end-to-end manner. Our experiments demonstrate that the proposed PAMnet can produce ghosting-free HDR images using inputs with complex motion and parallax. We compare our method against popular algorithms in [9, 10, 8, 11, 12], on various public test datasets, with superior reconstruction quality.

The main contributions are summarized below:

– We propose a deformable convolution-based pyramidal alignment network that uses multiscale image features for offset generation.

- We employ an attention optimized dense network to suppress misaligned features and fully utilize feature information for subsequent effective fusion using context-adaptive masks in the merging network.
- The pyramidal alignment and masked merging in feature domain can efficiently capture the complex displacements across LDR inputs induced by either temporal motion or spatial disparity for the HDR output with better image quality and richer local details. Extensive experiments and comparisons to the state-of-the-art algorithms have validated the superior efficiency of our proposed PAMnet, across a variety of test datasets.

## 2 Related Work

In this section, we will briefly review the existing approaches for multi-exposure based HDR reconstruction and deep learning methods for image registration related to this study.

### 2.1 Motion Handling Methods in HDR Reconstruction

Early works deal with camera motion and object motion to reconstruct ghosting-free HDR images. Previous approaches can be categorized into two classes depending on how to deal with object motion. The first class is based on motion detection. They detect moving pixels in the images which are rejected for final weighted HDR fusion, assuming that the images have been globally registered. The key to these methods is accurate motion detection. Yan *et al.* [5] formulate the object motion using a sparse representation. Lee *et al.* [6] propose to detect motion via rank minimization. However, these algorithms heavily depend on the effectiveness of motion detection and can not fully exploit the information of inputs.

The other class relies on the region alignment. Alignment-before-merging methods first align images of different exposures to the reference image, then merge them altogether to reconstruct the HDR image. The alignment is achieved through optical flow or patch match. Bogoni [7] registers local motion through estimating an optical flow field between each source image and a common reference image. Hu *et al.* [13] and Sen *et al.* [11] use patch-based methods to jointly optimize the alignment and reconstruction. These patch-based methods provide high robustness, but are time-consuming and may fail when there are large motions or large over-exposed regions in the reference image.

Recently, deep CNNs have offered significant performance improvement for many image reconstruction tasks. Kalantari *et al.* [8] firstly introduce neural networks in an alignment-before-merging pipeline to generate HDR images. Wu *et al.* [9] and Yan *et al.* [14] employ deep auto-encoder networks to translate multiple LDR images into a ghosting-free HDR image. Metzler *et al.* [15] and Sun *et al.* [16] jointly train an optical flow encoder and a CNN decoder to hallucinate the HDR content from a single LDR image. Choi *et al.* [17] suggest to reconstruct HDR videos using interlaced samples with joint sparse coding. Prabhakar *et*

*al.* [18] propose a deep network that can handle arbitrary inputs, where they first align input images using optical flow and refinement network, then merge aggregated features. Yan *et al.* [10] utilize spatial attention mechanism to guide HDR merging, however, its principle is close to that of the motion detection based methods. The spatial attention cannot fully exploit the characteristics of image features. Instead, we propose a network with multiscale feature-based pyramidal alignment which is more flexible and robust in handling motion and disparity.

### 2.2 Multi-camera HDR Reconstruction

Alternatively, multi-camera system can be utilized for HDR imaging, in which two or more cameras are fixed on a rig with different exposure settings. Park *et al.* [19] utilize multi-exposed LDR images from a stereo camera for HDR merging, where they estimated the depth map with the help of superpixel and hole-filling, then fused the HDR image using a depth weighting map. Selmanovic *et al.* [20] generate stereoscopic HDR video using an HDR and an LDR camera where HDR content of the LDR view was enhanced with the reference of captured HDR data. Popovic *et al.* [21] produce panoramic HDR videos using a circular multi-camera system. Adjacent cameras share pixels with similar perspective but different exposures. Most multi-camera HDR solutions employ depth-based warping for HDR reconstruction and can generate HDR images of multiple views, while occlusion and overexposure often make accurate depth estimation difficult.

### 2.3 Deep Registration Networks

Special network structures have been developed to solve the alignment task, which allow complex inputs and are not limited to traditional alignment pipeline steps [22]. The spatial transformer network [23] predicts hyper parameters for global spatial transformations but cannot solve local object motions. Bert *et al.* [24] use filters generated dynamically based on the input to learn transformations such as translation and rotation. However, limited kernel size restricts the ability of dynamic filters in handling large motions. The deformable convolution [25] enhances the standard convolution with flexible offsets and enables free deformable receptive fields. For modeling unknown transformations (e.g., motion induced occlusion, etc), we employ the deformable convolution for feature alignment.

## 3 Proposed Method

This section describes our network architecture in detail. Given a set of LDR images $(L_1, L_2, ..., L_k)$ that contain object and camera motion and are captured with different exposures, our goal is to reconstruct a ghosting-free HDR image $\widehat{H}$ at a specific reference view. We choose the middle exposure image as the reference image, that is, the image with the least number of underexposed or

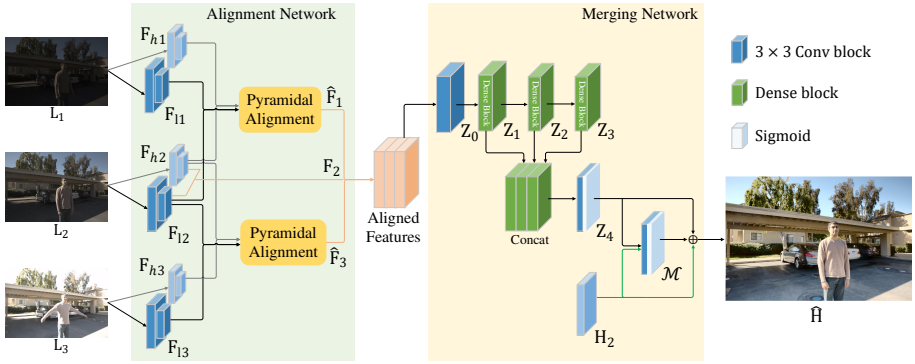



**Fig. 1. PAMnet.** The overall model includes a pyramidal alignment network and a merging network. Given input images with different exposures $L_i$, the alignment network first extracts image features at different scales ($F_{li}$, $F_{hi}$) and aligns them to the reference view. $H_i$ denotes the gamma corrected $L_i$. Reference image features $F_2$ and the aligned features $\widehat{F}_i$ are then fed into the merging network to reconstruct HDR image $\widehat{H}$.

overexposed pixels. Using an image group $(L_1, L_2, L_3)$ with three exposures as input, we set the middle exposure image $L_2$ as the reference image. We first convert the LDR images to corresponding HDR representations using gamma correction. Such gamma corrected images are closer to what we have perceived with our eyes [11]. The mapping operation can be written as

$$H_i = L_i^{\gamma}/t_i, \quad \text{for } i = 1, 2, 3, \tag{1}$$

where $\gamma = 2.2$ [26], $t_i$ is the exposure time of the $i^{th}$ image $L_i$, and $H_i$ denotes the mapped HDR domain image after gamma correction from the $L_i$.

As suggested in [8], LDR images can be used to detect the noisy or saturated regions (pixels), while the corresponding gamma corrected samples measure the content deviations from the reference image. As a result, we feed $L_i$ and $H_i$ together into our network. Note that image pre-alignment is not required in our approach. Pixel values of $L_i$, $H_i$ and $\widehat{H}$ are all normalized to $[0, 1]$.

### 3.1 Approach Overview

Our network employs deformable convolution for image feature alignment, then merges the aligned features using dilated dense blocks with attention optimization. The deformable convolution estimates content-based offsets for adaptive receptive fields. As shown in Fig. 1, our network is composed of two subnetworks: feature alignment network and merging network.

The alignment network first extracts image features at different scales in LDR and HDR domain using convolution layers. Then reference and non-reference image pyramidal feature pairs are fed into separate pyramidal alignment modules.

The modules take multi-scale image features in the HDR domain for offsets computation. Non-reference LDR and HDR domain image features are then aligned to the reference image features using corresponding offsets.

The merging network concatenates aligned image features in LDR and HDR domain as the input, and reconstructs HDR image by exploiting dilated dense residual blocks with SE connection and masked merging. Dilation rate $d$ is enlarged in dense blocks to learn nonlocal information followed by the SE attention [27] to remove redundant information and alleviate misalignment. Finally, the HDR image is reconstructed through a weighted fusion using adaptive mask.

### 3.2 Pyramidal Alignment

Given the input $(L_i, H_i),\ i = 1, 2, 3$ , the alignment network first extracts image features of different scales in LDR and HDR domain $(F_{li}^s, F_{hi}^s),\ i = 1, 2, 3,\ s = 1, 2$ , where $s$ denotes the number of scales. Considering the trade-off of network efficiency and capacity, pyramidal alignment module at 2 scales $(max(s)s = 2)$ is enough for strong performance, although a pyramidal alignment module with larger $s$ has a stronger ability in dealing with large motions.

We use the deformable convolution [28] to align image features in a coarse-to-fine manner. Let $w_k$ denotes convolution weight of the $k$-th location and let $p_k \in \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$ denotes pre-specified offset. Set $K = 9$ for a $3 \times 3$ kernel convolution layer with dilation rate 1. For pixel $x\,(p_0)$ at position $p_0$, the pixel value after deformable convolution alignment is:

$$y(p_0) = \sum_{k=1}^{K} w_k \cdot x\,(p_0 + p_k + \Delta p_k) \cdot \Delta m_k \ . \tag{2}$$

where $\Delta p_k$ and $\Delta m_k$ are the learnable offset and modulation scalar for the $k$-th location. $\Delta m_k$ and $\Delta p_k$ are calculated with the same features, the computation of $\Delta m_k$ is omitted in the following description for simplicity.

As shown in Fig. 2, we use image features in the HDR domain for offset computation which reduces the impact of different exposures. Learnable offsets are predicted from concatenated reference and $i$-th image features in HDR domain, then a deformable convolution layer aligns the source image features to the reference features:

$$\Delta p^s = ConvM\,([F_{hi}^s, F_{h2}^s]) \ . \tag{3}$$

$$\widehat{F}_i^s = Dconv\,([F_{li}^s, F_{hi}^s], \Delta p^s) \ . \tag{4}$$

where $ConvM$ denotes convolution layers, $\widehat{F}_i^s$ is the aligned feature of $i$-th input of scale $s$, $[\cdot,\cdot]$ is the concatenation operation, and $DConv$ is the deformable convolution. Let $\uparrow^2$ denotes the bilinear upsampling of scale factor 2, $\Delta p^s$ and $\Delta p^{s+1}$ refer to offset of scale $s$ and $s+1$ separately. After obtaining the aligned feature of $i+1$-th scale, we further refine the alignment on a upper scale with deformable convolution:

$$\Delta p^s = ConvM\left(F_{hi}^s, F_{h2}^s, \Delta p^{s+1} \uparrow^2\right) \ . \tag{5}$$

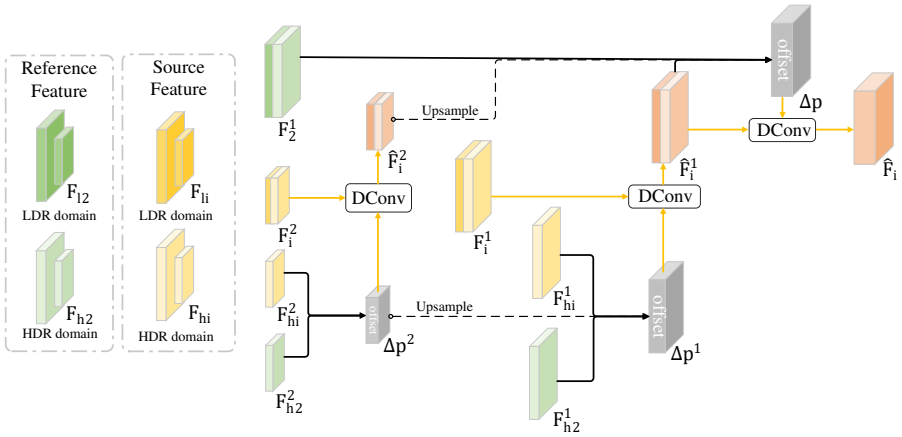



**Fig. 2.** Pyramidal alignment module architecture. The gray dotted frame on the left side of the figure shows the input features (yellow boxes indicate source features and green boxes indicate reference features), and the right side is the detailed alignment module structure. The alignment network uses HDR domain image features to compute offset, then align multi-scale features with the corresponding offsets. The aligned features $\widehat{F}_i^s$ and the final aligned feature after refinement $\widehat{F}_i$ are indicated by orange boxes.

Following the pyramidal feature alignment, an additional deformable convolution is introduced to improve details in the final aligned image feature $\widehat{F}_i$:

$$\Delta p = ConvM\left(\left[F_2^1, F_i^1, F_i^2 \uparrow^2\right]\right) \ . \tag{6}$$

$$\widehat{F}_i = Dconv\left(F_i^1, \Delta p\right) \ . \tag{7}$$

Where $\Delta p$ denotes the offset for refinement and $\widehat{F}_i$ is the final aligned feature. Previous works have validated the effect of coarse-to-fine spatial pyramidal alignment approach [29–31]. Taking advantage of the feature-based offset and pyramidal alignment, our alignment network can successfully handle parallax and complex motions, improving alignment accuracy with sub-pixel refinement. Besides, features in the HDR domain have higher brightness consistency. Consequently, we compute offsets with features in the HDR domain for better performance and less computation.

### 3.3 Merging Network

Inspired by the success of previous methods [10, 32], we utilize residual dense blocks with SE connection [27] for feature merging. As shown in Fig. 1, the merging network takes the concatenated aligned features $F = [\widehat{F}_1, F_2, \widehat{F}_3]$ as input. $Z_0$ is obtained after a convolution layer, then feature maps $Z_1$, $Z_2$, $Z_3$ are generated by feeding $Z_0$ into 3 residual dense blocks. Concatenated feature maps

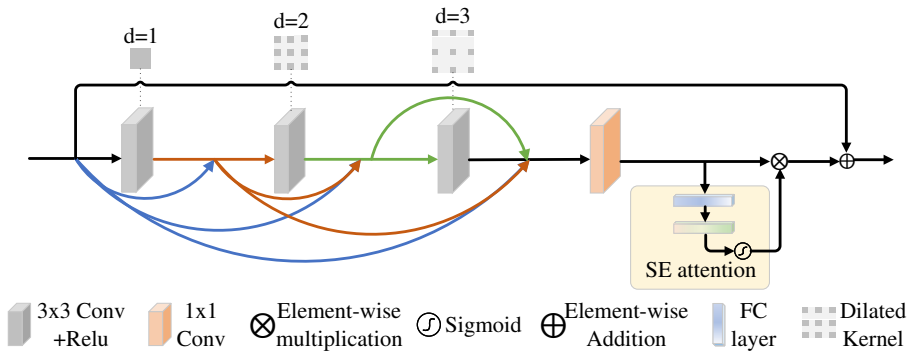



**Fig. 3.** Residual dense block with three dilated layers. squeeze-and-excitation (SE) attention module (the yellow box) is employed to remove redundant information and misalignment in channels.

produce $Z_4$ after several convolution layers and the sigmoid function. Finally, the HDR image $\widehat{H}$ is reconstructed with $Z_4$, $H_2$ and corresponding mask $M_{merge}$:

$$\mathcal{M} = \texttt{sigmoid}(ConvM\left([Z_4, H_2]\right)) \ , \tag{8}$$

$$M_{refine}, M_{merge} = \texttt{split}(\mathcal{M}) \ , \tag{9}$$

$$\widehat{H} = M_{merge} \cdot (M_{refine} \cdot Z_4) + (1 - M_{merge}) \cdot H_2 \ . \tag{10}$$

The `split` operation splits the mask $\mathcal{M}$ in the shape of $(N, 4, H, W)$ into a 3-channel mask $M_{refine}$ and 1-channel mask $M_{merge}$, of which the shape are $(N, 3, H, W)$ and $(N, 1, H, W)$ respectively.

Since dense connections may cause information redundancy, we apply SE connections in residual dense blocks which helps to remove redundancy. In addition, growing dilation rates are set in dense blocks to get larger receptive field for hallucinating details and misalignment elimination, as shown in Fig. 3.

### 3.4 Loss Function

Since HDR images are displayed after tone mapping, optimization in the tone mapped domain produces results with fewer artifacts in the dark regions than optimization in the HDR domain. We employ the $\mu$-law for tone mapping as supposed in [8], which is formulated as:

$$T(H) = \frac{\log(1 + \mu H)}{\log(1 + \mu)} \ . \tag{11}$$

where $\mu$ is set to 5000. Denoting ground truth HDR image and predicted HDR image as $H$ and $\widehat{H}$, the loss function can be defined as:

$$L(H, \widehat{H}) = \|T(H) - T(\widehat{H})\|_1 + \alpha\|\nabla T(H) - \nabla T(\widehat{H})\|_2 + \beta\|T(H) - T(Z_4)\|_1. \tag{12}$$

where $\alpha = 10, \beta = 0.5$ , $\nabla T(H)$ denotes the gradient magnitude of image $T(H)$, $\|\cdot\|_1$ and $\|\cdot\|_2$ denote $\ell_1$ and $\ell_2$ norm, respectively.

# 4 Experiments

## 4.1 Datasets

**Training Dataset.** We train on Kalantari's dataset [8] which contains 74 samples for training and 15 samples for testing. Each sample includes ground truth HDR images and three LDR images with exposure biases of $\{-2, 0, +2\}$ or $\{-3, 0, +3\}$. The dataset has both indoor and outdoor scenes and all images are resized to $1000 \times 1500$.

**Testing Dataset.** Testing is performed on Kalantari's testset [8] which has 15 scenes with ground truth and a dataset without ground truth [11]. To verify the model's ability in handling parallax, we also test our model on the Middlebury dataset [33] which consists of sets of images of different views with three different exposures. We use scenes from Middlebury 2005 and Middlebury 2006 for testing, and choose image sets of 3 illuminations and 2 different reference views from each scene.

## 4.2 Implementation Details

Given training data, we first crop them into $256 \times 256$ patches with a stride of 128 to expand training set size. The crop is conducted on LDR images and corresponding HDR label. Random flipping, noise, and 90 degrees rotation are applied on generated patches to avoid over-fitting. We use Adam optimizer [34] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate $= 10^{-4}$ and set batch size as 8. We perform training for 160 epochs. In order to better train the deformable convolution based alignment module, we employ the learning rate warmup trick where the warmup epoch is set to 10. We implement our model using PyTorch [35] on NVIDIA GeForce GTX 1080 GPU, and we decrease the learning rate by a factor of 4 every 70 epochs.

## 4.3 Analysis of Single-camera Case

We compare the proposed model with existing state-of-the-art methods on two datasets captured with single camera. We perform quantitative evaluations on

**Table 1.** Quantitative comparison on Kalantari's Testset [8]. Red color indicates the best performance and blue color indicates the second-best result.

| Methods | PSNR-L | PSNR-$\mu$ | SSIM-L | SSIM-$\mu$ | HDR-VDP-2 |
|---|---|---|---|---|---|
| AHDR [10] | 41.1782 | 43.7013 | 0.9857 | 0.9905 | 62.0521 |
| Wu [9] | 41.6593 | 41.9977 | 0.9860 | 0.9878 | 61.7981 |
| Kalantari [8] | 41.2200 | 42.7177 | 0.9829 | 0.9889 | 61.3139 |
| Sen [11] | 38.3425 | 40.9689 | 0.9764 | 0.9859 | 60.3463 |
| Ours | 41.6452 | 43.8487 | 0.9870 | 0.9906 | 62.5495 |


**Fig. 4.** Visual comparison using "Parking" from Kalantari's testset [8]. **Left**: input LDR images; **Upper Right**: HDR image reconstructed using proposed PAMnet; **Lower Right**: Zoomed-in patches of LDR images and HDR images. We choose the medium-exposure image as the reference. We show the results of the state-of-the-art HDR imaging algorithms, AHDR [10], Wu [9], Kalantari [8], Sen [11], and HDRNet [12]. The proposed PAMnet can produce high-quality HDR images even there are background saturation and large foreground motions.

Kalantari's testset [8] and qualitative assessments on dataset without ground truth [11]. We compare our model with the patch-based method [11], the single image enhancement method (HDRnet) [12], the flow alignment based method with a CNN merger [8], the UNet-based method [9] and the attention-guide method (AHDR) [10]. Note that we use a PyTorch [35] implementation of HDRnet which is trained on the same dataset as [12]. For other methods, we utilize the code and trained models provided by the authors for testing comparison.

We use metrics such as PSNR, SSIM, and HDR-VDP-2 for quantitative comparison. We compute PSNR and SSIM for images in linear domain (PSNR-L and SSIM-L) and images after the $\mu$-law tone mapping (PSNR-$\mu$, SSIM-$\mu$). We also compute HDR-VDP-2 [36] for quantitative comparison. Quantitative evaluation on Kalantari's testset [8] can be found in Table 1. All values are averaged on 15 test scenes. The proposed PAMnet has better numerical performance than other methods. For the sake of fairness, the HDRnet [12] is not included in the quantitative comparison because it produces enhanced LDR images. Fig. 4 and Fig. 5 compare our method with existing state-of-the-art methods. Sample of Fig. 4

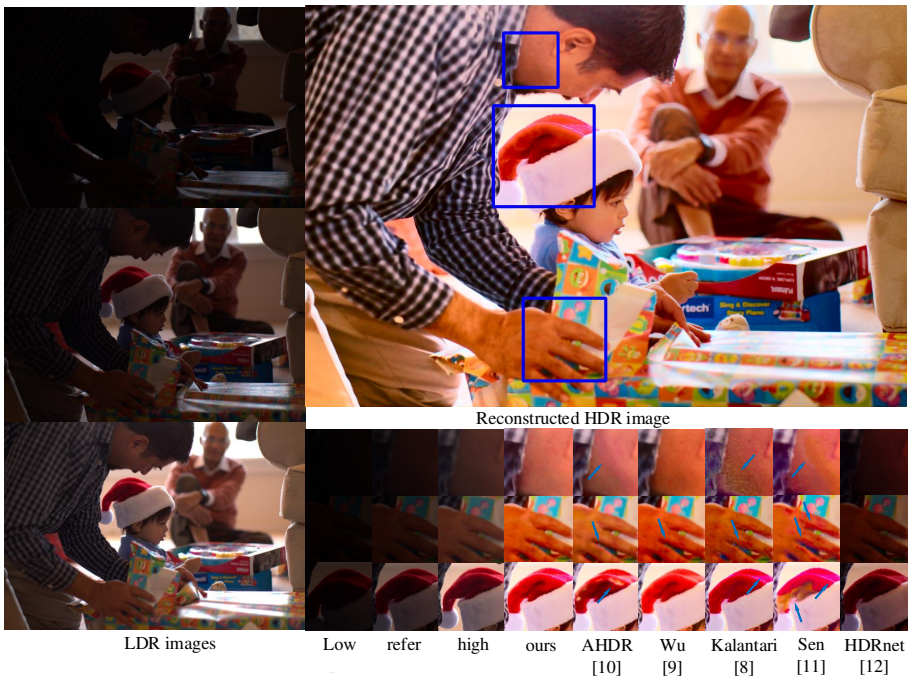
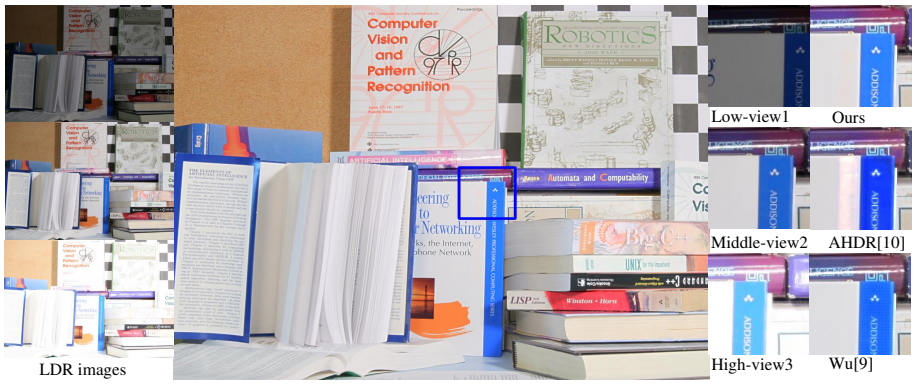


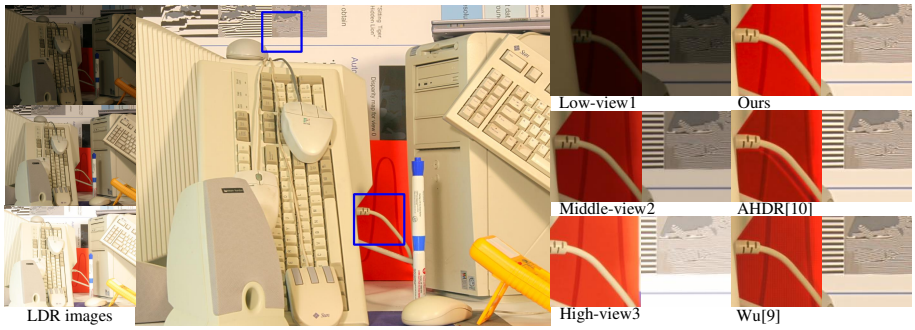

**Fig. 5.** Visual comparison using "Santas Little Helper" from Sen's dataset [11]. **Left**: input LDR images; **Upper Right**: HDR image reconstructed using proposed PAMnet; **Lower Right**: Zoomed-in patches of LDR images and HDR images. Our PAMnet produces HDR images with less noise and artifacts.

contains saturated background and large foreground motion. AHDR [10] and method of Wu *et al.* [9] produce ghosting artifacts in occluded region. Method of Kalantari *et al.* [8] leaves artifacts caused by optical flow alignment. Patch-based method (Sen *et al.* [11]) cannot find the right corresponding patches in the saturated region and generates line dislocation and color blocks. HDRnet [12] can't recover details in the saturated region and perturbs the tone of the image. With task-oriented pyramidal alignment, our network can produce high-quality HDR images even there is large motion in the inputs. For samples with underexposed objects in the reference image (as shown in Fig. 5), most methods can't reconstruct artifacts-free face and hand. Method of Sen *et al.* [11] and Kalantari *et al.* [8] produce results with local abnormal color and noise. HDRnet [12] enhances the underexposed region but can't hallucinate all details in the dark areas. Local color inconsistency appears around the hat in results generated by AHDR [10] because the spatial attention mask can only suppress unwanted pixels rather than find corresponding useful pixels.




(a) Books (illumination 1)


(b) Doll (illumination 2)




(c) Computer (illumination 1)

**Fig. 6.** Visual comparison on the Middlebury dataset [8]. **Left**: the input LDR images; **Middle**: the HDR image; **Right**: Zoomed-in patches of LDR images and HDR images. Our PAMnet can handle parallax (see zoomed-in patches) and produce ghosting-free HDR images.

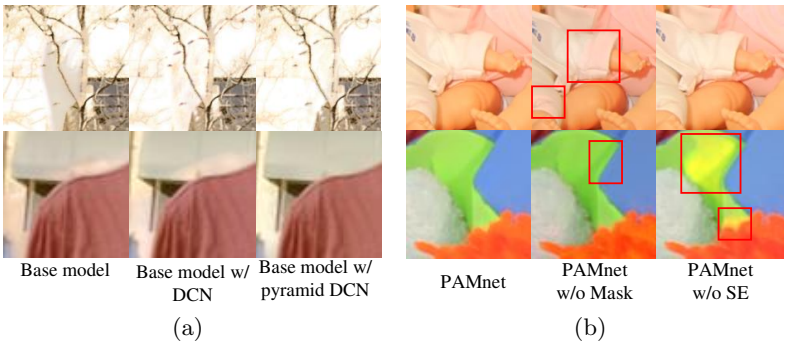



**Fig. 7.** Visual comparison of network variants via modularized switch.

### 4.4 Analysis of Multi-camera Case

To validate the model's ability to handle parallax, we performed evaluation on Middlebury testset [33]. For each scene, we select a set of 3 images with different exposures as inputs. We choose the two models with the best quantitative performance on Kalantari's testset [8] to compare with ours on the Middlebury dataset [33]. Results on three scenes with different environment illumination are shown in Fig. 6. AHDR [10] suffers ghosting artifacts because the attention-based network which suppresses unhelpful pixel values before merging is not suitable for handling large disparity. Method of Wu *et al.* [9] produces gridding effect in the fused results, which can be observed more obviously in a zoom-in view of patches in Fig. 6. Experimental results validate the superiority of our PAMnet which can handle large parallax and produce ghosting-free HDR images.

## 5 Ablation Study

This ablation study demonstrates the effectiveness of pyramidal feature alignment, masked merging, and usage of SE connection. Quantitative comparisons of the network variants are shown in Table 2.

**Table 2.** Quantitative comparison of different models.

| Methods | PSNR-L | PSNR-$\mu$ | SSIM-L | SSIM-$\mu$ |
|---|---|---|---|---|
| Base model | 40.1189 | 41.9602 | 0.9841 | 0.9891 |
| Base model w/ DCN | 40.8987 | 42.3182 | 0.9863 | 0.9896 |
| Base model w/ pyramidal DCN | 41.4254 | 43.4087 | 0.9865 | 0.9900 |
| PAMnet w/o SE | 41.5191 | 43.5479 | 0.9865 | 0.9903 |
| PAMnet w/o mask | 41.4764 | 43.6738 | 0.9866 | 0.9901 |
| PAMnet | 41.6452 | 43.8487 | 0.9870 | 0.9906 |

**Deformable Convolution-based Pyramidal Alignment** The deformable convolution based feature alignment can better mitigate the ghosting problem caused by motion and parallax. We remove the pyramidal alignment, SE connection and masked merging from our PAMnet as the base model. As shown in Fig. 7 (a), model with deformable convolution generates fewer ghosting artifacts comparing with the base model. The model with pyramidal feature alignment can generate ghosting-free HDR images, and its ability to handle large motions is stronger than the model with only single scale feature alignment.

**Masked Merging and SE Connection** Though model with pyramidal feature alignment can handle large motions, inaccurate alignment may introduce extra artifacts to the fused image. To suppress the misaligned features, we employ the SE attention [27] to the dense blocks of our network. The masked merging also helps to generate better results. As shown in Fig. 7 (b), the model without SE connection produces abnormal color in the second row while unnatural shadows and highlights arise in images generated by the model without masked merging (rectangle region). The full model with SE connection and masked merging can discard redundant information and inaccurate alignment, producing results with richer details and fewer artifacts.

## 6 Conclusion

In this paper, we propose a learned HDR imaging method that can handle complex object motion and large camera parallax. With pyramidal alignment and masked merging in feature domain, our method can produce high-quality HDR images in various scenarios having saturation, parallax, and occlusion in input data. Experiments validate that our model performs well on multi-exposure frames captured by both single-camera and multi-camera systems. We compare our method with existing state-of-the-art approaches on publicly available datasets and observe that our method offers significant improvement both objectively and subjectively. Our current work is exemplified using a fixed number of LDR images, but it is possibly extended to support an arbitrary number of images.

## 7 Acknowledgement

We are grateful for the constructive comments from anonymous reviewers. The corresponding author is Dr. Zhan Ma (mazhan@nju.edu.cn).

## References

1. Ledda, P., Santos, L.P., Chalmers, A.: A local model of eye adaptation for high dynamic range images. In: Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa. (2004) 151–160

2. Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., Brendel, H.: Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In: Digital Photography X. Volume 9023., International Society for Optics and Photonics (2014) 90230X
3. Tocci, M.D., Kiser, C., Tocci, N., Sen, P.: A versatile hdr video production system. ACM Transactions on Graphics (TOG) **30** (2011) 1–10
4. Nayar, S.K., Mitsunaga, T.: High dynamic range imaging: Spatially varying pixel exposures. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662). Volume 1., IEEE (2000) 472–479
5. Yan, Q., Sun, J., Li, H., Zhu, Y., Zhang, Y.: High dynamic range imaging by sparse representation. Neurocomputing **269** (2017) 160–169
6. Lee, C., Li, Y., Monga, V.: Ghost-free high dynamic range imaging via rank minimization. IEEE signal processing letters **21** (2014) 1045–1049
7. Bogoni, L.: Extending dynamic range of monochrome and color images through fusion. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. Volume 3., IEEE (2000) 7–12
8. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36** (2017) 144–1
9. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 117–132
10. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1751–1760
11. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. ACM Trans. Graph. **31** (2012) 203–1
12. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36** (2017) 1–12
13. Hu, J., Gallo, O., Pulli, K., Sun, X.: Hdr deghosting: How to deal with saturation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1163–1170
14. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Transactions on Image Processing **29** (2020) 4308–4322
15. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep optics for single-shot high-dynamic-range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 1375–1385
16. Sun, Q., Tseng, E., Fu, Q., Heidrich, W., Heide, F.: Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 1386–1396
17. Choi, I., Baek, S.H., Kim, M.H.: Reconstructing interlaced high-dynamic-range video using joint learning. IEEE Transactions on Image Processing **26** (2017) 5353–5366
18. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: 2019 IEEE International Conference on Computational Photography (ICCP), IEEE (2019) 1–8

19. Park, W.J., Ji, S.W., Kang, S.J., Jung, S.W., Ko, S.J.: Stereo vision-based high dynamic range imaging using differently-exposed image pair. Sensors **17** (2017) 1473
20. Selmanovic, E., Debattista, K., Bashford-Rogers, T., Chalmers, A.: Enabling stereoscopic high dynamic range video. Signal Processing: Image Communication **29** (2014) 216–228
21. Popovic, V., Seyid, K., Pignat, E., Çogal, Ö., Leblebici, Y.: Multi-camera platform for panoramic real-time hdr video construction and rendering. Journal of Real-Time Image Processing **12** (2016) 697–708
22. Villena-Martinez, V., Oprea, S., Saval-Calvo, M., Azorin-Lopez, J., Fuster-Guillo, A., Fisher, R.B.: When deep learning meets data alignment: A review on deep registration networks (drns). arXiv preprint arXiv:2003.03167 (2020)
23. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. (2015) 2017–2025
24. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: Advances in Neural Information Processing Systems. (2016) 667–675
25. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 764–773
26. Poynton, C.: Digital video and HD: Algorithms and Interfaces. Elsevier (2012)
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7132–7141
28. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. arXiv preprint arXiv:1811.11168 (2018)
29. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4161–4170
30. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8934–8943
31. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019) 0–0
32. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 2472–2481
33. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2007) 1–8
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. (2017)
36. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Transactions on graphics (TOG) **30** (2011) 1–14