SAFNet: Selective Alignment Fusion Network for Efficient HDR Imaging

Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen✉

vivo Mobile Communication Co., Ltd, China
{ltkong,libra,cokexiong,haozhang,guhong,jinwei.chen}@vivo.com

Abstract. Multi-exposure High Dynamic Range (HDR) imaging is a challenging task when facing truncated texture and complex motion. Existing deep learning-based methods have achieved great success by either following the alignment and fusion pipeline or utilizing attention mechanism. However, the large computation cost and inference delay hinder them from deploying on resource limited devices. In this paper, to achieve better efficiency, a novel Selective Alignment Fusion Network (SAFNet) for HDR imaging is proposed. After extracting pyramid features, it jointly refines valuable area masks and cross-exposure motion in selected regions with shared decoders, and then fuses high quality HDR image in an explicit way. This approach can focus the model on finding valuable regions while estimating their easily detectable and meaningful motion. For further detail enhancement, a lightweight refine module is introduced which enjoys privileges from previous optical flow, selection masks and initial prediction. Moreover, to facilitate learning on samples with large motion, a new window partition cropping method is presented during training. Experiments on public and newly developed challenging datasets show that proposed SAFNet not only exceeds previous SOTA competitors quantitatively and qualitatively, but also runs order of magnitude faster. Code and dataset is available at `https://github.com/ltkong218/SAFNet`.

Keywords: HDR imaging · Selective alignment fusion · Large motion

1 Introduction

Human eyes are capable to perceive a broad range of illumination in natural scenes, but camera sensors suffer from limited dynamic range due to inherent hardware properties, *i.e.*, the sensor's *thermal noise* and *full well electron capacity* [6]. The most common way to capture High Dynamic Range (HDR) image is to take a series of low dynamic range (LDR) photos at different exposures, and then merge them into an HDR image with increase realism [5].

If there is no motion or the LDR images are well aligned, existing image fusion methods can already produce faithful results [5, 25, 26, 34, 47]. Nevertheless, dynamic objects and camera motion usually appear in shooting scenes, which results in undesirable misalignment between LDR inputs. Directly applying previous methods will yield ghosting artifacts. To deal with dynamic

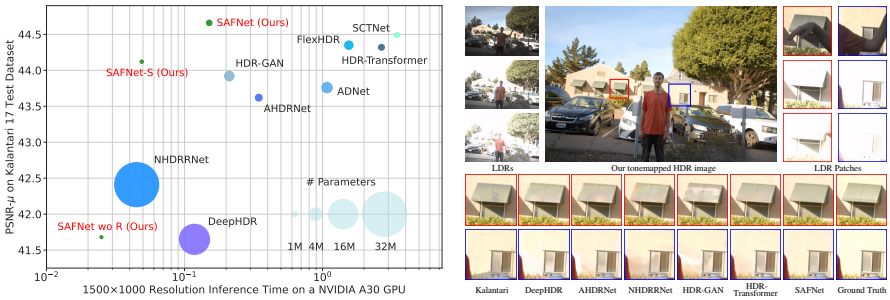

Fig. 1: Comparison on Kalantari 17 test dataset [14]. Proposed SAFNet achieves state-of-the-art HDR imaging accuracy while with fast inference speed and small model size.

scenarios, traditional methods try to align the LDR input images [1, 11, 15] or rejecting misaligned pixels [8, 12, 17, 31] before the fusion process. However, accurate alignment between LDR inputs with large motion and severe saturation is extremely challenging, while rejecting pixels in misaligned areas will cause insufficient content in moving regions.

As for deep learning based multi-exposure HDR imaging methods, main solutions can be summarized into two paradigms. The first class follows the alignment and fusion pipeline [14, 32, 33, 41], where cross-exposure motion is firstly estimated, then HDR fusion coefficients or final HDR results are generated based on aligned LDR inputs and context features. However, estimating optical flow between LDR frames under severe saturation and occlusion is error-prone. Some solutions [2, 4, 43] design special cross-exposure motion estimation networks for better alignment, but their computational complexity also increases. Differently, the second category bypasses explicit alignment, and proposes end-to-end deep networks with diverse attention mechanisms for fully spatial-/channel-wise feature interaction [3, 24, 37, 40, 44, 45]. There are also some methods that combine above paradigms for mutual promotion [23, 43]. However, as shown in Figure 1, recent research improves HDR reconstruction accuracy by proposing increasingly complex attention mechanisms, whose large inference delay and computation cost have hindered them from deploying on power constrained devices.

To achieve better efficiency, we propose a novel Selective Alignment Fusion Network (SAFNet) for multi-frame HDR reconstruction. Different from above design concepts, we observe that not all regions in the non-reference LDR images are worthy of precise alignment. For example, if some regions in the non-reference LDR inputs are over-/under-exposed or corresponding to well-exposed texture of the reference LDR frame, these areas can be directly discarded. On the other hand, if some regions in the non-reference LDR frames contain valuable texture that is missed in the reference LDR image, accurate alignment and fusion in these regions can promote final reconstruction quality. Fortunately, motion estimation in regions with distinct texture is much easier than that are saturated [16, 38]. By holding above proposition, our SAFNet performs valuable area selection and flow estimation in selected regions simultaneously, which can skip the tough yet

error prone motion estimation in worthless areas, while focus the model's learning capabilities on more meaningful things. In practice, proposed SAFNet follows the successful pyramidal pipeline in optical flow networks [19,35,39]. Specifically, we use gradually refined one-channel selection probability masks to denote valuable regions during coarse-to-fine flow estimation. These masks are finally adopted to reweight fusion coefficients for flow aligned LDR inputs and generate high quality HDR result by explicit fusion operation [14]. At last, for high frequency detail enhancement, a lightweight refine module is introduced at input resolution. Thanks to the valuable information in previous optical flow, selection masks and initial HDR prediction, we find that simply employing several dilated residual blocks can already achieve SOTA accuracy while with much higher efficiency. Tied to our two-stage deep network, a new window partition cropping method is presented when optimizing SAFNet, which can benefit long-distance texture aggregation and short-distance detail refinement simultaneously.

In addition to progressively advanced HDR algorithms, datasets also play an important role for evaluation. Though the dataset developed by Kalantari *et al.* [14] has largely facilitated the research for multi-exposure HDR imaging, only three test samples are challenging enough for visual comparison. A recent work [40] proposes a new HDR dataset with enriched scene motion and content. However, motion magnitude and saturation ratio in their datasets are relatively small, restricting its evaluative ability. In order to study the performance gap between different algorithms in challenging cases, we propose a new multi-exposure HDR dataset with enhanced motion range and saturated regions, that clearly distinguishes the quantitative and qualitative HDR reconstruction results among different approaches. Finally, we do experiments on the Kalantari 17 dataset [14] as well as our developed Challenge123 dataset. As shown in Figure 1, the proposed SAFNet not only sets new state-of-the-art accuracy but also runs order of magnitude faster than recent Transformer-based competitors [24, 40]. Main contributions of this paper can be summarized as follows:

- We propose a novel SAFNet for multi-frame HDR imaging, that jointly refines valuable region masks and cross-exposure motion in selected regions, and then explicitly fuses a high quality result with much better efficiency.
- We provide a new challenging multi-frame HDR deghosting dataset with enhanced motion and saturation for ease of analysis.
- Experiments on public and newly developed datasets show that our SAFNet outperforms previous SOTA methods and runs order of magnitude faster.

2 Related Work

Traditional Methods. Traditional multi-exposure HDR methods mainly utilize pixel rejection or motion registration techniques. The first class focuses on aligning LDR inputs globally and then discarding misaligned pixels before image fusion for deghosting. To generate error map for pixel rejection, Grosch *et al.* [9] leverage color difference between the aligned images, Pece *et al.* [31] employ the median threshold bitmap of the LDR inputs, Jacobs *et al.* [12] propose weighted

intensity variance analysis. Besides, Zhang *et al.* [47] and Khan *et al.* [17] calculate gradient-domain weight maps or probability maps of the LDR inputs, respectively. Lee *et al.* [22] and Oh *et al.* [30] detect moving areas by utilizing rank minimization. However, pixel rejection approach abandons useful texture in moving regions, producing unpleasing reconstruction quality.

The other motion registration-based methods rely on densely aligning the no-reference LDR inputs to the reference frame prior to merging them. Bogoni *et al.* [1] calculate optical flow as motion vectors for full image alignment. Kang *et al.* [15] transfer the LDR inputs into luminance domain according to exposure time for improving flow accuracy. Sen *et al.* [36] introduce a patch based energy minimization method that optimizes alignment and HDR fusion at the same time. Additionally, Hu *et al.* [11] promote image alignment by propagating brightness and gradient information iteratively in a coarse-to-fine manner. However, optimizing energy function for motion estimation usually drops into local minimum. Also, their slow speed is unsuitable for real-time applications.

Deep Learning Approaches. Early deep learning multi-exposure HDR algorithms follow traditional alignment and fusion pipeline. Kalantari *et al.* [14] pioneer learning-based multi-frame HDR reconstruction by proposing a paired LDR-HDR dataset and developing a convolutional neural network (CNN) to fuse LDR inputs after flow alignment. Wu *et al.* [41] instead adopt image-wide homography to perform background alignment, while leaving the complex foreground motions to be handled by the CNN. Despite noteworthy performance improvement over traditional methods, both [14] and [41] suffer from misalignment in the presence of both large motion and severe saturation. Subsequent methods improve alignment by building more powerful cross-exposure motion estimation modules [2, 4] or perform feature alignment with attention mechanism [23, 43]. However, the large computation cost and running time hinders their development on mobile devices.

Yan *et al.* address some limitations of the predecessors by introducing a spatial attention module [44], and further constructing a non-local block to improve global consistency [45]. Niu *et al.* [29] leverage Generative Adversarial Network (GAN) to synthesize realistic content which is missing in the LDR inputs. Furthermore, Xiong *et al.* [42] decompose HDR imaging into ghost-free image fusion and ghost-based image restoration. Ye *et al.* [46] propose a progressive feature fusion network that compares and selects appropriate LDR regions to generate high quality result. Above attention-based and region selective HDR algorithms can facilitate deghosting in fusion stage. However, their results fall behind current state-of-the-arts due to the limited long-range texture aggregation ability.

Recently, transformers have shown better ability to capture long-range dependency than CNN due to their multi-head self-attention mechanism. Song *et al.* [37] separate LDR inputs into ghost and non-ghost regions, and then selectively apply either transformer or CNN to perform HDR reconstruction. Liu *et al.* [24] integrate vision transformer with convolution to explore both local and global relationship and obtain remarkable results. Furthermore, Yan *et al.* [43] propose a HyHDRNet consisting of a content alignment subnetwork

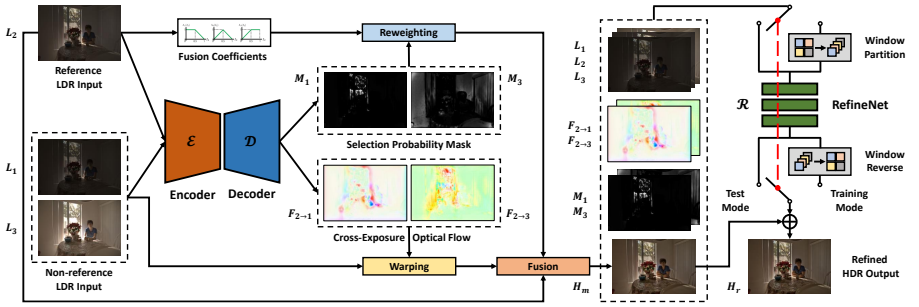

Fig. 2: Overall architecture of our SAFNet. It contains a pyramid encoder, a coarse-to-fine decoder, and a refinement subnetwork. The linked switch selects path including window partition and window reverse during training, while skip them in evaluation.

and a transformer-based fusion subnetwork for performance improvement. Tel *et al.* [40] introduce a SCTNet which integrates both spatial and channel attention modules into the transformer-based network to enhance semantic consistency. Nevertheless, transformer-based methods suffer from large inference delay. Moreover, their patch-based prediction manner is unable to aggregate cross-patch texture produced by large motion, which is common in high resolution imagery.

3 Proposed Method

Overview. Given three LDR images L_1, L_2, L_3 from a dynamic scene with different exposures as input, our goal is to generate a ghost-free HDR image H_r with consistent scene structure as the reference image L_2. Figure 2 depicts overall architecture of the proposed SAFNet, containing three subnetworks, *i.e.*, the pyramid encoder $\mathcal{E}$, the coarse-to-fine decoder $\mathcal{D}$, and the detail refine module $\mathcal{R}$. SAFNet first performs an extraction phase to retrieve a pyramid of features from each input frame L_i by the encoder $\mathcal{E}$. Then, it jointly refines selection probability masks M_1, M_3 together with cross-exposure optical flow $F_{2\to1}, F_{2\to3}$ in selected regions by the decoder $\mathcal{D}$. Furthermore, a high quality HDR image H_m is explicitly merged by flow aligned LDR inputs and selection masks reweighted fusion coefficients. Finally, our SAFNet generates a refined HDR image H_r by the refine network $\mathcal{R}$ based on LDR inputs L_i, cross-exposure motion $F_{2\to i}$, selection probability masks M_i and merged HDR image H_m. In addition to network architecture, training loss function and developed window partition cropping method are also elaborated in this section.

Pyramid Encoder. Like previous methods [14,40,43,44], we first map the LDR frames L_i to the HDR linear domain by using gamma correction as follows:

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

where t_i denotes the exposure time of LDR image L_i, γ is the gamma parameter, which is set to 2.2. By concatenating L_i and H_i along the channel dimension, we obtain three 6-channel tensors $X_i = [L_i, H_i]$ as network input.

Inspired by the success of pyramidal flow estimation architectures [19, 20, 35, 39], our encoder network $\mathcal{E}$ extracts multi-scale pyramid features to better estimate cross-exposure optical flow in the challenging large motion and heavy saturation cases. Purposely, the encoder is built by a block of two 3×3 convolutions in each pyramid level, with strides 2 and 1, respectively. The parameter shared encoder extracts 4 levels of pyramid features, counting 8 convolution layers, each followed by a PReLU activation [10]. With gradually decimated spatial size, it keeps the feature channels to 40 among all 4 scales, generating pyramid features $\phi_1^k, \phi_2^k, \phi_3^k$ in level k ($k = 1, 2, 3, 4$) for LDR inputs L_1, L_2, L_3, separately.

Coarse-to-Fine Decoder. To deal with the large displacement challenge for motion estimation, we follow the successful pyramid optical flow networks [19, 35, 39], that adopt coarse-to-fine warping strategy and predicts easier residual flow at each scale. After extracting meaningful pyramid features, the decoder $\mathcal{D}^k$ iteratively refines cross-exposure optical flow by backward warping pyramid features ϕ_1^k, ϕ_3^k to generate $\tilde{\phi}_1^k, \tilde{\phi}_3^k$ according to $F_{2\to1}^k$ and $F_{2\to3}^k$, respectively, where $\mathcal{D}^k$ means this parameter shared decoder $\mathcal{D}$ is called in level k. However, unlike traditional optical flow task, cross-exposure motion estimation is much more challenging due to co-existence of large motion and severe saturation. Instead of designing complex flow estimation network for overall improvement as previous [2,4,43], we observe that not all regions of $F_{2\to1}, F_{2\to3}$ are worthy of accurate prediction. Therefore, besides cross-exposure optical flow $F_{2\to1}^{k-1}, F_{2\to3}^{k-1}$, the decoder network $\mathcal{D}^k$ further predicts two selection probability masks M_1^{k-1}, M_3^{k-1} to denote valuable regions of $F_{2\to1}^{k-1}, F_{2\to3}^{k-1}$ during coarse-to-fine flow estimation, respectively. Specifically, M_1^{k-1} and M_3^{k-1} are one-channel tensors exported by sigmoid function whose elements range from 0 to 1. Different from previous cascading HDR reconstruction pipeline [14], *i.e.*, first optical flow then fusion, our jointly refined M_1, M_3 and $F_{2\to1}, F_{2\to3}$ can benefit each other. First, M_1, M_3 can inform the decoder to focus on estimating $F_{2\to1}, F_{2\to3}$ in the identified areas. In turn, better estimated $F_{2\to1}, F_{2\to3}$ can aggregate valuable pyramid features from non-reference frames to promote further region identification and residual flow estimation. Therefore, accuracy and efficiency are both improved. In summary, input and output of the coarse-to-fine decoder can be formulated as:

$$[F_{2\to1}^{k-1}, F_{2\to3}^{k-1}, M_1^{k-1}, M_3^{k-1}] = \mathcal{D}^k([F_{2\to1}^k, F_{2\to3}^k, M_1^k, M_3^k, \tilde{\phi}_1^k, \phi_2^k, \tilde{\phi}_3^k]), \quad (2)$$

where $\mathcal{D}^k$ ($k = 1, 2, 3, 4$) denotes the iterative refinement in level k, $[\cdot]$ means feature concatenation. The initial value of $F_{2\to1}^4, F_{2\to3}^4, M_1^4, M_3^4$ are all set to 0, while the final prediction by $\mathcal{D}^1$ are written as $F_{2\to1}, F_{2\to3}, M_1, M_3$.

Concretely, the decoder $\mathcal{D}$ is consist of a block of five 3×3 convolutions and one 4×4 deconvolution, with strides 1 and 1/2, respectively. A PReLU [10] follows each convolution layer. Feature channels of intermediate layers of $\mathcal{D}$ are all set to 120. Following the success of efficient flow estimation in [19], the middle three convolutions of $\mathcal{D}$ are group convolution with group number equal to 3, which are separated by channel shuffle operation. Figure 3 shows structure details of the decoder that is shared among all pyramid levels.

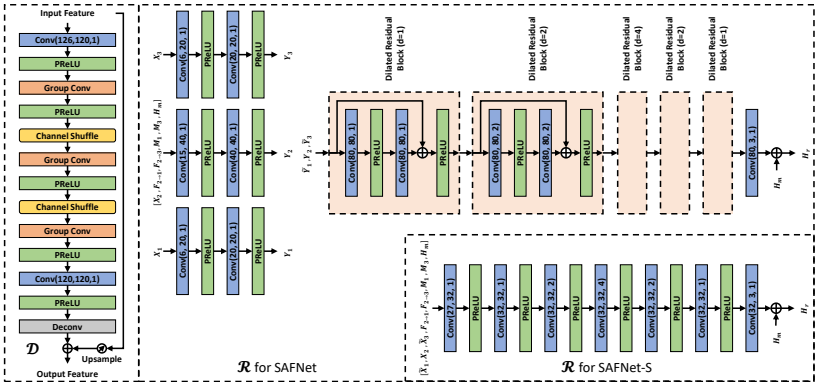

Fig. 3: Details of the decoder $\mathcal{D}$ and the refine network $\mathcal{R}$ for SAFNet and SAFNet-S. Arguments of 'Conv' from left to right are input channels, output channels and dilation. All convolutions have 3×3 kernel size. Stride is equal to dilation for each 'Conv'.

Explicit HDR Fusion. The final goal of predicted optical flow $F_{2\to1}, F_{2\to3}$ and selection probability masks M_1, M_3 is to merge a high quality HDR image H_m as close as possible to the ground truth H_{gt}. As shown in Figure 2, there are two preliminary steps before the HDR fusion procedure. At first, we use estimated optical flow $F_{2\to1}, F_{2\to3}$ to align the non-reference linear domain images H_1, H_3, and generate warped images of $\tilde{H}_1, \tilde{H}_3$ as follows:

$$\tilde{H}_1 = w(H_1, F_{2\to1}),\ \tilde{H}_3 = w(H_3, F_{2\to3}), \tag{3}$$

where w means backward warping [35, 39]. Secondly, predicted selection probability masks M_1, M_3 are employed to reweight initial fusion coefficients for ghost-suppressed HDR synthesis. Considering that the predicted optical flow $F_{2\to1}, F_{2\to3}$ by SAFNet are relatively accurate only in regions where M_1, M_3 contain a relatively large selection probability. Therefore, to eliminate ghosting artifacts when fusing unrelated textures, we multiply the initial fusion coefficients of $\tilde{H}_1, \tilde{H}_3$ with their selection probability masks M_1, M_3, respectively. Meanwhile, the unselected parts of initial fusion coefficients of $\tilde{H}_1, \tilde{H}_3$ are transferred to the reference image H_2 for normalization. Formulaically, proposed reweighted fusion coefficients can be computed by:

$$\begin{aligned} W_1 = \Lambda_1 \odot M_1, \quad W_3 = \Lambda_3 \odot M_3, \\ W_2 = \Lambda_2 + \Lambda_1 \odot (1 - M_1) + \Lambda_3 \odot (1 - M_3), \end{aligned} \tag{4}$$

where $\Lambda_1, \Lambda_2, \Lambda_3$ are initial HDR fusion coefficients for inputs H_1, H_2, H_3, that are defined in Figure 4. W_1, W_2, W_3 are the reweighted fusion coefficients of SAFNet. $\odot$ means element-wise multiplication. Given optical flow aligned input images and selection masks reweighted fusion coefficients, we can merge a high quality HDR image explicitly by:

$$H_m = W_1 \odot \tilde{H}_1 + W_2 \odot H_2 + W_3 \odot \tilde{H}_3. \tag{5}$$

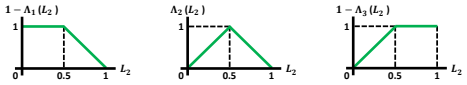

Fig. 4: Functions to define initial HDR fusion coefficients.

Refine Network. Explicit fusion approach in Eq. 5 can already fuse a high quality HDR image. However, it can not synthesize textures that are truncated or occluded in all input frames. To compensate for missing content and remove potential ghost, a refine network $\mathcal{R}$ is introduced in Figure 2, whose details are shown in Figure 3. Specifically, $\mathcal{R}$ is a fully convolutional network that works at original input resolution to enhance high frequency details. First, three independent feature extractors, each including two 3×3 convolutions, are adopted to extract local features Y_1, Y_2 and Y_3 from inputs X_1, $[X_2, F_{2\to1}, F_{2\to3}, M_1, M_3, H_m]$ and X_3, respectively. Then, aligned shallow features of $\tilde{Y}_1, \tilde{Y}_3$ are obtained by backward warping Y_1, Y_3 according to flow $F_{2\to1}, F_{2\to3}$, separately. Finally, the concatenated feature of $[\tilde{Y}_1, Y_2, \tilde{Y}_3]$ is forwarded to an aggregation module, containing five dilated residual blocks and one convolution, to estimate residual details and yield a refined HDR image H_r as our final prediction.

Loss Function. Since HDR images are typically viewed after tonemapping, we use μ-law function to map the image from linear domain to the tonemapped domain as follows:

$$T(H) = log(1 + \mu H)/log(1 + \mu), \quad \mu = 5000, \tag{6}$$

where H denotes the HDR image in linear domain, μ is the compression parameter. Following methods [24, 40, 43], we employ the weighted $\mathcal{L}_1$ loss and perceptual loss $\mathcal{L}_p$ [13] between our refined output H_r and the ground truth H_{gt} to supervise HDR reconstruction of SAFNet by:

$$\mathcal{L}_r = \mathcal{L}_1(T(H_r), T(H_{gt})) + \alpha\mathcal{L}_p(T(H_r), T(H_{gt})), \tag{7}$$

where $\mathcal{L}_p$ measures the distance on multi-scale features extracted by a pre-trained VGG-16 network, α is the weighting parameter set to 0.01. Additionally, we add an auxiliary brightness reconstruction loss $\mathcal{L}_m$ to guide the learning of alignment and fusion for the merged HDR image H_m as:

$$\mathcal{L}_m = \mathcal{L}_1(T(H_m), T(H_{gt})) + \mathcal{L}_c(T(H_m), T(H_{gt})), \tag{8}$$

where $\mathcal{L}_c$ is the census loss [18, 21, 28] calculating the soft Hamming distance between census-transformed image patches of size 7×7. In summary, our total training loss is the combination of $\mathcal{L}_r$ and $\mathcal{L}_m$, that can be expressed as:

$$\mathcal{L} = \mathcal{L}_r + \beta\mathcal{L}_m, \tag{9}$$

where β is the trade-off coefficient, which is set to 0.1.

Window Partition Cropping. Recent multi-frame HDR methods [24, 40, 43] crop 128×128 image patches during training, that can generate enough data with diverse saturation and occlusion for sufficient learning. However, their relatively small crop size will block long-range texture aggregation in large motion

Fig. 5: Visual example of LDR input L_2, optical flow from L_2 to L_1 and saturated regions of L_2 in proposed dataset.

Table 1: Statistics comparison among different multi-exposure HDR datasets.

Statistics (Avg)	Kalantari [14]	Tel [40]	Ours
Motion Magnitude	20.1	16.2	128.7
Saturation Ratio	0.061	0.073	0.201

cases. To deal with this problem, we propose a new window partition cropping method during optimization, that is bound to our two-stage SAFNet. As shown in Figure 2, the merged H_m is generated on large patches of 512×512 to better aggregate long-range inter-frame textures. On the other hand, the refined H_r is predicted on small patches of 128×128 to better synthesize local details. We unify above two different cropping sizes by window partition and reverse operations, where the additional size is first shifted to the batch dimension and then shifted back. In test stage, window partition and reverse are discarded.

4 Proposed Dataset

The existing labeled multi-exposure HDR datasets [7, 14, 40] have facilitated research in related fields. However, results of recent methods [2, 24, 40, 43] tend to be saturated due to their limited evaluative ability [14, 40]. We attribute this phenomenon to most of their samples having relatively small motion magnitude between LDR inputs and relatively small saturation ratio of the reference image. To probe the performance gap between different algorithms, we propose a new challenging multi-exposure HDR dataset with enhanced motion range and saturated regions. Our proposed Challenge123 dataset follows the same collection pipeline as [14], but use a vivo X90 Pro+ phone equipped with Sony IMX 989 sensor. There are 96 training samples and 27 test samples in our dataset, whose construction details can be found in our supplementary material.

To highlight the differences of our dataset over existing ones, we use a pre-trained flow network [19] to estimate optical flow from L_2 to L_1 and calculate saturated area ratio of L_2 for every test sample in each dataset. Figure 5 depicts one visual example of this statistic approach. By averaging motion magnitude and saturation ratio over all test samples in each HDR dataset, we summarize the statistics comparison of different datasets in Table 1. As can be seen, motion magnitude and saturation ratio of proposed Challenge123 dataset are about 7× and 3× larger than existing HDR datasets [14,40], respectively. It is worth noting that our dataset is complementary to existing ones [14,40], which aims to widen the performance gap between different algorithms for ease of analysis, even if it is less natural than them. For example, the challenging regions of [14] where the man is waving his arms contain more than 150 pixels displacement, which is similar as the average motion magnitude of our dataset.

Table 2: Quantitative comparison with methods on Kalantari 17 test dataset [14]. For each item, the best result is **boldfaced**, and the second best is underlined. Time and FLOPs are measured on one NVIDIA A30 GPU under 1500×1000 resolution.

Method	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP2	Time (s)	Params (M)	FLOPs (T)
Sen [36]	40.80	38.11	0.9808	0.9721	59.38	-	-	-
Kalantari [14]	42.67	41.23	0.9888	0.9846	65.05	-	-	-
DeepHDR [41]	41.65	40.88	0.9860	0.9858	64.90	0.118	16.61	1.304
AHDRNet [44]	43.63	41.14	0.9900	0.9702	64.61	0.345	1.52	2.170
NHDRRNet [45]	42.41	41.43	0.9877	0.9857	61.21	**0.045**	31.56	0.421
HDR-GAN [29]	43.92	41.57	0.9905	0.9865	65.45	0.211	2.56	1.455
ADNet [23]	43.87	41.69	0.9925	0.9885	65.56	1.078	2.96	4.648
FlexHDR [2]	44.35	42.60	0.9931	0.9902	66.56	1.550	2.12	-
HDR-Transformer [24]	44.32	42.18	0.9916	0.9884	66.03	2.673	1.22	1.886
HyHDRNet [43]	44.64	42.47	0.9915	0.9894	66.05	-	-	-
SCTNet [40]	44.49	42.29	0.9924	0.9887	66.65	3.466	0.99	1.547
SAFNet-S (Ours)	44.12	42.50	0.9928	0.9907	66.75	0.049	**0.57**	**0.146**
SAFNet (Ours)	**44.66**	**43.18**	**0.9932**	**0.9917**	**66.93**	0.151	1.12	0.976

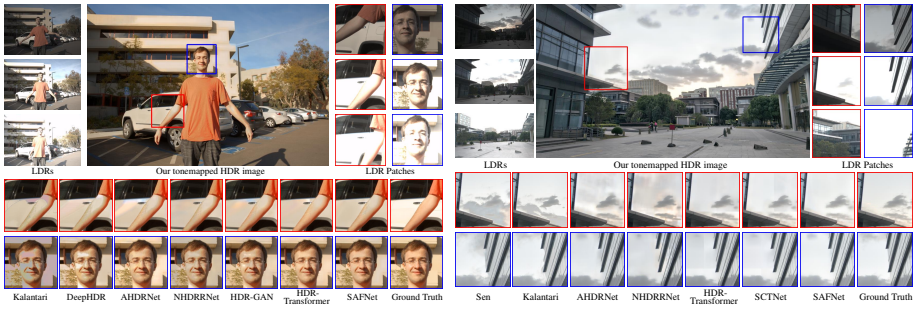

Fig. 6: Visual comparison on Kalantari [14] (left) and Challenge123 (right) test sets.

5 Experiments

Datasets. We first evaluate proposed SAFNet on the conventional Kalantari 17 dataset [14] including 74 training samples and 15 test samples. Then, we compared our algorithm with recent SOTA methods on the proposed Challenge123 dataset consist of 96 training scenes and 27 test scenes. Finally, we analyze the generalization ability of diverse approaches on the Sen's dataset [36].

Evaluation Metrics. For quantitative comparison, we calculate five commonly used metrics, *i.e.*, PSNR-μ, PSNR-l, SSIM-μ, SSIM-l and HDR-VDP2 [27], where -μ denotes tonemapped domain and -l means linear domain.

Implementation Details. We implement proposed method in PyTorch, and use Kalantari 17 [14] or our Challenge123 training set to train SAFNet from scratch for separate comparison. Proposed network is optimized by Adam algorithm with β_1=0.9, β_2=0.999 and ϵ=1e−8 for total 10,000 epochs. The optimization is conducted with total batch size 4 on 2 NVIDIA A30 GPUs, whose learning rate is initially set to 2e−4 and gradually decays to 1e−5 following a cosine attenuation schedule. During training, we augment the samples by random cropping, flipping, rotating and reversing channel order to prevent overfitting.

Proposed window partition cropping method is also employed with input resolution of 512×512 and 128×128 for the two-stage SAFNet. For better efficiency, optical flow and selection masks are predicted at 1/2 input resolution and then upsampled back. The overall training takes less than 24 hours.

Comparison on Kalantari 17 dataset. We compare proposed algorithm with 11 well-known multi-exposure HDR methods on Kalantari 17 test dataset [14], whose quantitative results are summarized in Table 2. As can be seen, proposed SAFNet achieves state-of-the-art performance on all five metrics consistently. It is worth noting that our method surpasses the second-best result [2] by 0.58 dB and 0.0015 in terms of PSNR-l and SSIM-l, because of our explicit HDR fusion process in linear domain as Eq. 5. Figure 1 and Figure 6 qualitatively compare SAFNet with 6 famous HDR algorithms. It is obvious that our method can generate more realistic background with much fewer ghosting artifacts.

To compare execution efficiency of different solutions, we download their official code and evaluate the running time, model parameters and computation complexity on a machine equipped with one NVIDIA A30 GPU under 1500×1000 resolution. Inference time is averaged over 100 iterations. In Table 2, FlexHDR [2], HDR-Transformer [24] and SCTNet [40] achieve similar PSNR-μ accuracy as ours. However, running speed of SAFNet is about **10**×, **18**× and **23**× faster than them, respectively. We attribute the excellent efficiency of SAFNet to its coarse-to-fine fully convolutional deep architecture without any complex attention mechanism. In regard to computation complexity, proposed SAFNet saves about 37%, 48% and 55% multiply-add operations than advanced SCTNet [40], HDR-Transformer [24] and AHDRNet [44], separately. For comparison with the efficient NHDRRNet [45] which also leverages the pyramidal encoder-decoder architecture, we build a small SAFNet-S by reducing feature channels of the refine module from 80 to 32 and replacing residual blocks with convolutions as shown in Figure 3. As listed in Table 2, SAFNet-S has similar running time as NHDRRNet [45], but obtains significant performance improvement. In conclusion, proposed SAFNet sets new records of speed-accuracy trade-off on multi-exposure HDR imaging task, which is depicted in Figure 1 intuitively.

Evaluation on Challenge123 dataset. For evaluation on the developed Challenge123 dataset, we retrain 4 advanced methods [24, 40, 44, 45] on our training set with their official implementations, and also compare two alignment-based algorithms [14, 36]. Due to the 24GB memory limitation for NVIDIA A30 GPU, transformer-based methods [24, 40] can only be optimized up to 256×256 resolution. Therefore, we first train all approaches on 256×256 patches under the same data augmentation method and learning schedule for fair comparison. After training, we summarize the accuracy of different algorithms on our test set in the middle part of Table 3. It can be concluded that our SAFNet consistently performs well on all five metrics, exceeding HDR-Transformer [24], SCTNet [40], AHDRNet [44] and NHDRRNet [45] by 0.24, 0.29, 0.50 and 3.12 dB on PSNR-μ, separately, verifying our superior multi-exposure HDR architecture.

Since our dataset contains larger motion magnitude than Kalantari 17 [14], training crop size of AHDRNet [44], NHDRRNet [45] and our SAFNet are ex-

Table 3: Quantitative comparison on Challenge123 test set. The middle and bottom parts are trained on patches of 256×256 and 512×512, respectively.

Method	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP2
Sen [36]	37.11	27.80	0.9729	0.9687	51.93
Kalantari [14]	37.83	29.62	0.9707	0.9705	51.32
AHDRNet [44]	40.44	28.13	0.9877	0.9703	54.58
NHDRRNet [45]	37.82	26.75	0.9769	0.9632	53.38
HDR-Tran. [24]	40.70	28.72	0.9881	0.9731	54.63
SCTNet [40]	40.65	28.73	0.9882	0.9721	54.35
SAFNet (Ours)	40.94	28.93	0.9885	0.9740	54.84
AHDRNet [44]	40.61	28.33	0.9880	0.9708	54.97
NHDRRNet [45]	37.44	26.31	0.9762	0.9596	53.51
SAFNet (Ours)	**41.88**	**29.73**	**0.9897**	**0.9784**	**55.07**

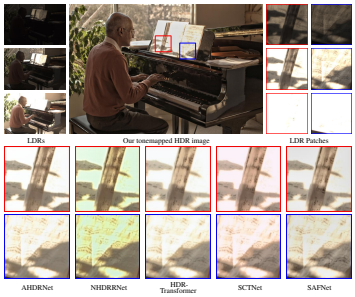

Fig. 7: Generalization comparison on the Sen's dataset [36].

tended to 512×512 for further comparison on long-range texture aggregation. In the bottom of Table 3, our SAFNet still outperforms all the others on all metrics, and our performance advantages are further amplified in challenging cases. Increasing patch size from 256 to 512, proposed SAFNet improves most, AHDRNet [44] improves smaller than ours, while NHDRRNet [45] even drops in some metrics. It indicates that our SAFNet is more robust and capable to learn from challenging large motion training samples. Transformer-based solutions [24, 40] still fall behind our SAFNet, since their patch-based prediction manner can not aggregate cross-patch moving texture in high-resolution photography. Furthermore, in the top of Table 3, alignment-based HDR fusion approach [14] behaves well in PSNR-l, confirming the significance of alignment for large motion. Differently, our SAFNet exceeds [14] on all metrics, due to our stronger registration and fusion ability. More analysis can be found in our supplementary.

Figure 6 visually compares them on a challenging sunset scene, including both camera and clouds motion. We can observe that early flow and patch based methods [14,36] generate distorted edges since misalignment in saturated regions. Attention-based networks [44,45] synthesize blurry outputs due to lack of motion compensation. Transformer-based solutions [24, 40] suffer from block artifacts because of limited aggregation scope. In contrast, our SAFNet can fuse more pleasing HDR images. More results can be found in our supplementary.

Generalization Ability. To compare generalization capability of recent HDR imaging methods, we test 5 algorithms on the unsupervised Sen's dataset [36],

Table 4: Ablation on first fusion stage. $F_{2\to i}$ and M_i are output components of the decoder. Accuracy is measured on H_m.

$F_{2\to i}$	M_i	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l
✓	✗	33.69	36.30	0.9568	0.9701
✗	✓	40.69	37.08	0.9834	0.9772
✓	✓	**41.68**	**39.61**	**0.9851**	**0.9808**

Table 5: Ablation on second refinement stage. $F_{2\to i}$, M_i and H_m are input of the refine network. Accuracy is measured on H_r.

$F_{2\to i}$	M_i	H_m	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l
✗	✗	✗	43.63	41.67	0.9922	0.9902
✓	✗	✗	43.73	41.88	0.9924	0.9902
✓	✓	✗	43.87	42.04	0.9925	0.9904
✓	✓	✓	**44.59**	**43.15**	**0.9929**	**0.9911**

Table 6: Ablation on window partition cropping. S1 and S2 are input resolution for two stages. Accuracy is measured on H_r.

S1	S2	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l
128×128	128×128	44.59	43.15	0.9929	0.9911
512×512	512×512	44.54	43.09	0.9930	0.9914
512×512	128×128	**44.66**	**43.18**	**0.9932**	**0.9917**

where all methods are trained on our proposed Challenge123 dataset fairly. As shown in Figure 7, NHDRRNet [45] and SCTNet [40] suffer from unnatural color problem, while AHDRNet [44] and HDR-Transformer [24] synthesize more hazy textures compared to proposed SAFNet, especially in the shaded areas.

Ablation on First Fusion Stage. In the first stage, SAFNet explicitly merges an initial HDR image H_m as in Eq. 5. To verify the effectiveness of optical flow $F_{2\to i}$ and selection masks M_i predicted by the decoder, we do ablations on Kalantari 17 dataset [14] by selectively removing them in the progressive refinement procedure. In Table 4, removing selection masks M_i will result in significant performance degradation, since precise alignment for entire image is extremely challenging under heavy saturation and complex motion. On the other hand, removing optical flow $F_{2\to i}$ will cause a smaller but also noticeable accuracy loss, because of lost ability for long-range texture aggregation. Figure 8 visually compares these ablation experiments. As can be seen, H_m w/o M_i looks more twisted, while H_m w/o $F_{2\to i}$ lacks texture in moving regions. By jointly refining selection masks and optical flow in selected regions, proposed approach can concentrate on finding and fusing more valuable textures for HDR reconstruction more efficiently. Figure 9 depicts two visual examples of selection masks and optical flow predicted by SAFNet. We can observe that our network can find over-exposed wall in the reference LDR image and estimate relatively precise optical flow in selected regions to aggregate valuable cross-exposure textures. As for unselected black regions, flow estimation of SAFNet is relatively free without worry about negative impacts, such as potential ghosting artifacts.

Ablation on Second Refinement Stage. In the second stage, SAFNet compensates high frequency details by the refine network based on LDR inputs and first stage outputs, *i.e.*, optical flow $F_{2\to i}$, selection masks M_i and merged HDR image H_m. To explore the value of different inputs for the refinement module, we gradually remove H_m, M_i and $F_{2\to i}$ from the inputs, and carry out ablations on Kalantari 17 dataset [14]. As listed in Table 5, optical flow, selection masks and merged HDR image all contain valuable while complementary information for guiding the refinement procedure, where the best result is achieved when using

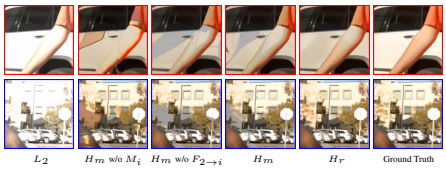

Fig. 8: Qualitative results of ablation on selection mask, optical flow and refinement.

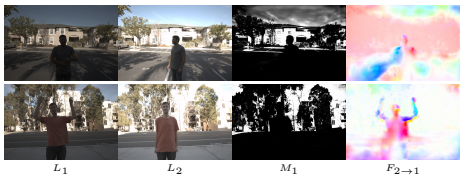

Fig. 9: Visualization of selection mask M_1 and optical flow $F_{2\to 1}$.

all these input components. It is worth noting that H_m contributes most to the performance, confirming the significance of region selective HDR fusion in the first stage. Figure 8 depicts two visual examples to demonstrate the effectiveness of our refine subnetwork, which can not only compensate for missing contextual details but also correct distorted scene structure.

Ablation on Window Partition Cropping. To confirm the effectiveness of our window partition cropping method, we do ablations on different input resolutions when optimizing SAFNet, whose results are summarized in Table 6. It can be seen that when employing proposed window partition cropping approach to train SAFNet on different input resolutions, we can yield the best performance. We attribute the reason to that there is trade-off between training patch size and HDR accuracy. Larger patch size can create more long-range aggregation training samples for merging H_m, while smaller input patches can generate more challenging samples with diverse occlusion and saturation cases for refining H_r.

6 Conclusion

In this paper, we present a novel SAFNet for efficient and accurate multi-exposure HDR imaging. By jointly refining valuable area selection masks and optical flow in selected regions, it can focus on finding and aggregating more useful LDR textures and finally merge a high quality HDR image explicitly. Based on diverse features exported from the first fusion stage, a lightweight refinement module is introduced to compensate for missing details. Moreover, to better optimize our two-stage SAFNet, a new window partition cropping method is proposed. Experiments on conventional and newly developed challenging datasets demonstrate that our algorithm not only outperforms recent SOTA methods quantitatively and qualitatively, but also runs order of magnitude faster than advanced transformer-based solutions.

References

1. Bogoni, L.: Extending dynamic range of monochrome and color images through fusion. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. vol. 3, pp. 7–12 vol.3 (2000)
2. Catley-Chandar, S., Tanay, T., Vandroux, L., Leonardis, A., Slabaugh, G., Pérez-Pellitero, E.: Flexhdr: Modeling alignment and exposure uncertainties for flexible hdr imaging. IEEE Transactions on Image Processing **31**, 5923–5935 (2022)
3. Chen, J., Yang, Z., Chan, T.N., Li, H., Hou, J., Chau, L.P.: Attention-guided progressive neural texture fusion for high dynamic range image restoration. IEEE Transactions on Image Processing **31**, 2661–2672 (2022)
4. Chung, H., Cho, N.I.: Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12760–12769 (2023)
5. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. p. 369–378. SIGGRAPH '97 (1997)
6. Fossum, E.R., Hondongwa, D.B.: A review of the pinned photodiode for ccd and cmos image sensors. IEEE Journal of the Electron Devices Society **2**(3), 33–43 (2014)
7. 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. vol. 9023, p. 90230X (Mar 2014)
8. Gallo, O., Gelfandz, N., Chen, W.C., Tico, M., Pulli, K.: Artifact-free high dynamic range imaging. In: 2009 IEEE International Conference on Computational Photography (ICCP). pp. 1–7 (2009)
9. Grosch, T.: Fast and robust high dynamic range image generation with camera and object movement. In: IEEE Conference of Vision, Modeling and Visualization (2006)
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
11. Hu, J., Gallo, O., Pulli, K., Sun, X.: Hdr deghosting: How to deal with saturation? In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1163–1170 (2013)
12. Jacobs, K., Loscos, C., Ward, G.: Automatic high-dynamic range image generation for dynamic scenes. IEEE Computer Graphics and Applications **28**(2), 84–93 (2008)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision – ECCV 2016. pp. 694–711 (2016)
14. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4) (2017)
15. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. ACM Trans. Graph. **22**(3), 319–325 (2003)
16. Kearney, J.K., Thompson, W.B., Boley, D.L.: Optical flow estimation: An error analysis of gradient-based methods with local optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-9**(2), 229–244 (1987)
17. Khan, E.A., Akyuz, A.O., Reinhard, E.: Ghost removal in high dynamic range images. In: 2006 International Conference on Image Processing. pp. 2005–2008 (2006)

18. Kong, L., Jiang, B., Luo, D., Chu, W., Huang, X., Tai, Y., Wang, C., Yang, J.: Ifrnet: Intermediate feature refine network for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
19. Kong, L., Shen, C., Yang, J.: Fastflownet: A lightweight network for fast optical flow estimation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 10310–10316 (2021)
20. Kong, L., Yang, J.: Fdflownet: Fast optical flow estimation using a deep lightweight network. In: 2020 IEEE International Conference on Image Processing (ICIP) (2020)
21. Kong, L., Yang, J.: Mdflow: Unsupervised optical flow learning by reliable mutual knowledge distillation. IEEE Transactions on Circuits and Systems for Video Technology (2022)
22. Lee, C., Li, Y., Monga, V.: Ghost-free high dynamic range imaging via rank minimization. IEEE Signal Processing Letters **21**(9), 1045–1049 (2014)
23. Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., Liu, S.: Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 463–470 (2021)
24. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: Computer Vision – ECCV 2022. pp. 344–360 (2022)
25. Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z.: Deep guided learning for fast multi-exposure image fusion. IEEE Transactions on Image Processing **29**, 2808–2819 (2020)
26. Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., Zhang, L.: Robust multi-exposure image fusion: A structural patch decomposition approach. IEEE Transactions on Image Processing **26**(5), 2519–2532 (2017)
27. 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 Trans. Graph. **30**(4) (jul 2011)
28. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
29. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.H.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. IEEE Transactions on Image Processing **30**, 3885–3896 (2021)
30. Oh, T.H., Lee, J.Y., Tai, Y.W., Kweon, I.S.: Robust high dynamic range imaging by rank minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(6), 1219–1232 (2015)
31. Pece, F., Kautz, J.: Bitmap movement detection: Hdr for dynamic scenes. In: 2010 Conference on Visual Media Production. pp. 1–8 (2010)
32. Prabhakar, K.R., Agrawal, S., Singh, D.K., Ashwath, B., Babu, R.V.: Towards practical and efficient high-resolution hdr deghosting with cnn. In: Computer Vision – ECCV 2020. pp. 497–513 (2020)
33. 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). pp. 1–8 (2019)

34. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
35. 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 (CVPR) (2017)
36. 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**(6) (2012)
37. Song, J.W., Park, Y.I., Kong, K., Kwak, J., Kang, S.J.: Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In: Computer Vision – ECCV 2022. pp. 288–304 (2022)
38. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2432–2439 (2010)
39. 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 (CVPR) (2018)
40. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12836–12845 (2023)
41. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Computer Vision – ECCV 2018. pp. 120–135 (2018)
42. Xiong, P., Chen, Y.: Hierarchical fusion for practical ghost-free high dynamic range imaging. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 4025–4033 (2021)
43. Yan, Q., Chen, W., Zhang, S., Zhu, Y., Sun, J., Zhang, Y.: A unified hdr imaging method with pixel and patch level. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22211–22220 (2023)
44. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
45. 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**, 4308–4322 (2020)
46. Ye, Q., Xiao, J., Lam, K.m., Okatani, T.: Progressive and selective fusion network for high dynamic range imaging. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 5290–5297 (2021)
47. Zhang, W., Cham, W.K.: Gradient-directed multiexposure composition. IEEE Transactions on Image Processing **21**(4), 2318–2323 (2012)