Journal of Imaging Science and Technology® 67(5): 1–19, 2023.

Multi-Attention Guided SKFHDRNet For HDR Video Reconstruction

Ehsan Ullah and Marius Pedersen▲
Colourlab, Department of Computer Science, NTNU, Gjøvik, Norway
E-mail: marius.pedersen@ntnu.no

Kjartan Sebastian Waaseth and Bernt-Erik Baltzersen
DvNor, Nagra Kudelski, Oslo, Norway

Abstract. *We propose a three stage learning-based approach for High Dynamic Range (HDR) video reconstruction with alternating exposures. The first stage performs alignment of neighboring frames to the reference frame by estimating the flows between them, the second stage is composed of multi-attention modules and a pyramid cascading deformable alignment module to refine aligned features, and the final stage merges and estimates the final HDR scene using a series of dilated selective kernel fusion residual dense blocks (DSKFRDBs) to fill the over-exposed regions with details. The proposed model variants give HDR-VDP-2 values on a dynamic dataset of 79.12, 78.49, and 78.89 respectively, compared to Chen et al. ["HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset," Proc. IEEE/CVF Int'l. Conf. on Computer Vision (IEEE, Piscataway, NJ, 2021), pp. 2502–2511] 79.09, Yan et al. ["Attention-guided network for ghost-free high dynamic range imaging," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (IEEE, Piscataway, NJ, 2019), pp. 1751–1760] 78.69, Kalantari et al. ["Patch-based high dynamic range video," ACM Trans. Graph. 32 (2013) 202–1] 70.36, and Kalantari et al. ["Deep hdr video from sequences with alternating exposures," Computer Graphics Forum (Wiley Online Library, 2019), Vol. 38, pp. 193–205] 77.91. We achieve better detail reproduction and alignment in over-exposed regions compared to state-of-the-art methods and with a smaller number of parameters.* © *2023 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2023.67.5.050409]

1. INTRODUCTION

There is a difference and mismatch of dynamic range information when capturing a physical scene. This means that more visual information is available in the scene than what can be captured and reproduced as conventional camera system's capabilities are limited in simultaneously covering the wide range of luminance in a single exposure. Additionally, a large part of the digital content currently used is stored and captured using 8-bit integer values, offering $2^8 = 256$ distinct levels. These device-referred formats such as JPEG, PNG, TIFF, etc. are constructed according to the limitations of display devices and accommodate according to the capabilities of the imaging device with minimum care for loss of visual information that the imaging device cannot display [1].

High Dynamic Range (HDR) video can be created through reconstruction using single or multiple Low Dynamic Range (LDR) frames captured using conventional cameras by alternating the exposure of each frame using software solutions or using specialized single-shot HDR cameras. HDR reconstruction using single exposure is further divided into three unique sub-problems of decontouring (Daly and Feng [2], Song et al. [3], Luzardo et al. [4], Mukherjee et al. [5], tone expansion Banterle et al. [6, 7], De Simone et al. [8], Masia et al. [9]) and filling of details in over-exposed regions from its adjacent non-exposed pixels [10, 11]. Time-sequential multi-exposure techniques are another way to capture HDR images, by taking a sequence of images with different exposures. Although an LDR sensor may record only a small portion of the whole luminance range of a scene at any given time, it has a functional range with the potential to include the entire luminance range by adjusting the exposure of each capture. The images are further combined to generate an image with a higher dynamic range. For video, one can obtain alternate exposures between subsequent video frames, this had resulted in multi-exposure techniques for video. In the case of HDR reconstruction of video, the problem of frame alignment to compensate for camera and object motion arises. This is often solved by methods that rely on pixel-level alignment with optical flow [12–15]. Recently, several learning-based methods have been used for reconstructing HDR video. Refs. [13–15] addresses the problem of HDR reconstruction by using Convolutional Neural Network(CNN) with optical flow to learn the HDR video reconstruction. Wu et al. [16] aligned LDR frames by performing homography, which is a non-flow-based approach. Yan et al. [17] applied attention mechanism for content alignment and gave importance to only those features that are similar to the reference image and excluded regions with motion and severe saturation. Later, they introduced a non-local neural network [18]. Despite these appraoches it still remains a big challenge to reconstruct ghost-free HDR videos from sequences with alternating exposures.

In this paper, we introduce a learning-based approach to address the issue of HDR video reconstruction with two

▲ IS&T Member.

Received June 11, 2023; accepted for publication Aug. 23, 2023; published online Oct. 6, 2023. Associate Editor: Gabriele Gianini.

1062-3701/2023/67(5)/050409/19/$25.00

alternating exposures. The goal is to obtain ghost-free videos with good detail preservation. Our approach has three main stages, the first stage performs alignment of neighboring frames to the current frame by estimating the flows between them, recovering a large part of missing details from the input LDR images, and the second stage is composed of multi-attention modules and a Pyramid Cascading Deformable (PCD) alignment module [19] to refine previously aligned features by performing a sophisticated feature alignment. The final stage performs merging by estimating the final HDR scene based on a series of Dilated Selective Kernel Fusion Residual Dense Blocks (DSKFRDBs) with global residual learning strategy [17, 20] that allows the network to fill the over-exposed regions with rich details. The entire network is trained in an end-to-end fashion to reconstruct HDR video. We employ L_1 and a combined L_1MS–SSIM [21] loss function to minimize the error between the reconstructed and original HDR frames.

The major contributions of our work for HDR video reconstruction are as follows:

- Introduction of multi-attention (particularly using a selective kernel fusion module) blocks with the goal of proper image alignment by extracting rich information spatially, channel-wise, and giving attention to the scale of the content in the input frames.
- For effective HDR video reconstruction, we employ robust DSKFRDBs in the merge network for recovering details in over- and under-exposed regions.
- Our proposed model has fewer network parameters than previous learning-based techniques.
- Model training is performed using L_1 and a combined L_1MS–SSIM loss to guide the optimization algorithm by learning more refined network weight parameters for HDR video reconstruction.

Our proposed multi-attention selective kernel fusion HDR network (SKFHDRNet) method showed a fair improvement over existing techniques and makes it possible to use LDR frames in HDR video reconstruction.

2. RELATED WORK

Different approaches have been proposed for hardware-based HDR video acquisition and computationally-based HDR reconstruction. Nayar and Mitsunaga [22] and Nayar et al. [23] proposed different types of per-pixel changeable optical density masks that were used to vary the spatial exposure to capture the scene at different exposures. Others [24–26] were able to successfully capture a wider range of HDR video through internal/external beam-splitters. The sensor's dynamic range capabilities are improved by [27], while some sensors calculate the logarithm of the irradiance in the analog domain using the logarithmic response of a sensor [28, 29].

Many single-exposure computationally-based inverse tone mapping operators made efforts to solve the issue by applying separate expansion to pixels that are classified as saturated recovering details in over-exposed regions [6, 30–34]. Didyk et al. [35] decomposed video frame components into diffuse, reflections, and light sources using a semi-manual classifier (Zhang and Brainard [10] and Xu et al. [11]) to perform pixel-level image processing. A dithering-based approach was proposed that adds noise to mask banding artifacts due to quantization [2, 5]. More recently, several methods have employed deep learning strategies for single-exposure HDR image reconstruction. Eilertsen et al. [36] used a CNN-based encoder and decoder architecture reconstructing colors, intensities, and details in saturated regions. By merging bracketed LDR images, Endo et al. [37] indirectly recreated an HDR image from a single LDR input. Liu et al. [38] developed three deep networks for dequantization, linearization, and hallucination of missing details in over-exposed regions.

Kang et al. [12] proposed the first HDR video reconstruction algorithm for sequences with alternating exposures using optical flow. Mangiat and Gibson [39] improved the approach by Kang et al. [12] using a block-based motion estimation method coupled with a refinement stage. In follow-up work, Mangiat and Gibson [40] proposed to filter regions with a large motion to reduce blocking artifacts. Kalantari et al. [41] proposed a patch-based optimization system to synthesize the missing exposures at each frame. Gryaditskaya et al. [42] improved the method of Kalantari et al. [41] by adaptively adjusting the exposures. Li et al. [43] proposed the HDR video reconstruction problem as maximum *a posteriori* estimation. Kalantari and Ramamoorthi [14] addressed the drawbacks of their previous approach [13] by proposing to use CNNs to learn the HDR video reconstruction process. Eilertsen et al. [44] improved the temporal stability of CNNs by introducing a regularization approach that encourages the network to produce consistent results for consecutive frames in a video. Yan et al. [17] proposed an attention-guided deep neural network with an attention mechanism for frame alignment for HDR imaging. Kim et al. [45] addressed the reconstruction of (ultra high definition) UHD HDR videos by simultaneously working on the content super-resolution and inverse tone-mapping and introducing GAN (Generative Adversarial Network) based architecture with multiple subnets for specific tasks. The super-resolution and inverse tone-mapping (SR-ITM) framework is further extended by utilizing information at multi-scale to enhance the network's local receptive fields. The approach involves downsampling image features at various scales, enabling to catch complex image patterns from pixels using varied local receptive field sizes [46]. Chen et al. [47] suggested a deep learning pipeline composed of adaptive global color mapping, local enhancement, and highlight generation. For adaptive global color mapping, they introduced a color condition block that extracts global image priors and adapts them to different images. Beside that, ResNet was used as their network architecture and a GAN model for local enhancement and highlight generation, respectively. Similarly, GAN-based framework for HDR video reconstruction from LDR sequences with

Figure 1. Representation of three consecutive frames with two alternate exposures of the carousel firework scene in Ref. [26] HDR dataset. Each frame in three consecutive frame input contains few missing contents with the presence of noise in frame $F_i - 1$ and $F_i + 1$ in the darker region due to acquisition with low exposure whereas F_i, which was taken with high exposure, lacks details in over-saturated and bright regions. The missing content of a final HDR image has to be reconstructed from neighboring frames with alternating exposures. For our full model we also used the neighboring frames $F_i - 2$ and $F_i + 2$ as well.

alternating exposures was adopted by Anand et al. [48]. Yang et al. [49] introduced a multimodal learning framework for reconstructing HDR videos based on three components. One component to align the frames; the second, a fusion component based on confidence guided multimodal fusion, and the last component to suppress flicker. Yang et al. [50] proposed a lightweight-efficient network based on structural re-parameterization, and a motion alignment loss to reduce motion artifacts. Cogalan et al. [51] proposed a CNN method for HDR image and video reconstruction that works for both for single-shot acquisition with spatially-interleaving exposures and for multi-shot acquisition with spatially-interleaving and temporally-alternating exposures. Their method used optical flow and is stated to work well for non-linear motion as well. Liu et al. [52] focused on optical flow estimation for LDR images of different exposures, and they proposed an unsupervised approach that incorporates a model-based algorithm and a data-driven deep network. Martorell and Buades [53] proposed a variational temporal approach to optical flow estimation that has data and spatial smoothness terms, as well as a temporal smoothness term and to match pixels from different frames. Jiang et al. [54] introduced a tri-exposure quad-bayer sensors. With a larger number of exposure sets uniformly distributed over each frame, providing robustness to noise and spatial artifacts. Ref. [55] produced high-dynamic range (HDR) video using dual-exposure sensors, which capture differently exposed and spatially interleaved half-frames in a single shot, eliminating the need for exposure alignment. Neural networks are employed for denoising, deblurring, and upsampling tasks and optical flow is utilized for precise warping. Recently, Chen et al. [15] came up with a two-stage coarse-to-fine framework for HDR video reconstruction. Their first stage aligns images using optical flow and blending in the image space. Their second stage performs more sophisticated alignment fusion for HDR video using deformable convolution [56] in PCD module as well as performing fusion temporally.

However, most single exposure-based techniques are not built to handle videos and cannot handle noise in the dark regions while hallucinating only smaller saturated regions. Similarly, solving the issue of frame alignments and temporal aspects of HDR video reconstruction through single attention is challenging, and recent models with optical flow have a large number of parameters and struggle on examples with large motions.

3. MULTI-ATTENTION GUIDED SKFHDRNet FOR HDR VIDEO RECONSTRUCTION

Given an input LDR video/sequential frames $\{I|i = 1, \ldots n\}$ with alternating exposures $\{t|i = 1, \ldots n\}$, the Multi-Attention SKFHDRNet reconstructs a high-quality HDR video $\{H|i = 1, \ldots n\}$. Similar to [13–15], input frames in linear and LDR domain are stacked and passed to the network for HDR video reconstruction shown in Figure 1.

3.1 *Data Preprocessing*

Similar to the work of [13–15] the camera response function of the input frames I_i is assumed to be known. As in Refs. [14, 15], we replace the camera response function of the input images with a fixed gamma curve as:

$$F_i = \text{lin}_i(I_i) => (I_i t_i)^{1/\gamma}, \quad (1)$$

where γ is set to 2.2 and lin_i is a function that transfers the image I_i from the linear HDR domain into LDR domain at exposure t_i. Similarity transforms that include rotation, translation, and isometric scaling are applied to globally align adjacent frames to simplify the learning process of our proposed model.

Real-world cameras often produce noisy images and are difficult to calibrate. It is necessary for the training dataset to represent these limitations of conventional camera systems to enable the learning-based model to perform and generalize effectively on scenes captured with conventional consumer cameras. Refs. [13–15] imitate the flaws of common consumer cameras by introducing noise and altering the tone of the synthetic images in their synthetic training dataset for ensuring the generalizability of their proposed network during inference time. Image acquisition through conventional digital cameras usually contains noisy pixels in dark regions. Then the information from those darker regions of the image should be taken from the high-exposure image which has more details in that region.

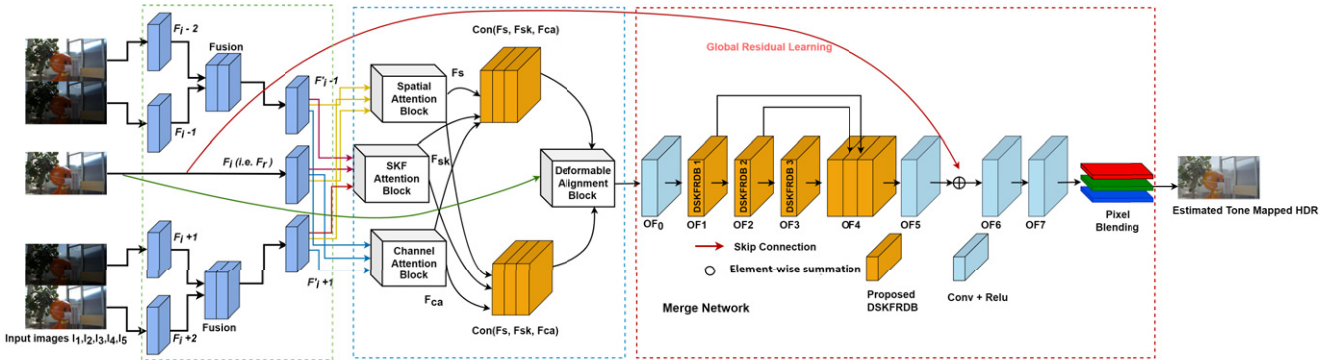

Figure 2. Visualization of the network architecture of our proposed multi-attention SKFHDRNet for HDR video reconstruction with two alternating exposures.

The input LDR synthetic training dataset usually has the same amount of noise for both exposures. Using the dataset directly without modification, the content of the high-exposure image in the dark regions will be unused, which eventually produces noisy results in real scenes [14]. Similar to Kalantari and Ramamoorthi [14] and Chen et al. [15], zero-mean Gaussian noise was added to the input LDR images with low exposure making the models use the information in the dark regions of a clean high exposure image. The zero-mean Gaussian noise was specifically applied to the images in the linear domain. The intention was to magnify the noise in the dark regions after transforming the image into the LDR domain. To account for noise variation similar to [14, 15], random Gaussian noise range using the standard deviation between $10\times^{-3}$ and 3×10^{-3} and the tone of the reference image was perturbed with $\gamma = \exp(d)$ function, where d is randomly selected from the range $[-0.7, 0.7]$ for simulation of an inaccurate camera response function. Cropped patches of size 256×256 were given as input to the proposed model along with random horizontal/vertical flipping and rotation.

3.2 *Pipeline*

In Figure 2, the multi-attention SKFHDRNet comprises two primary sub-networks. These sub-networks are designed to align and recover missing content in the reference (center) frame using attention modules, incorporating spatial, channel, and attention through adaptive kernel selection and fusion mechanisms. The multi-attention blocks focus solely on the relevant features related to the center frame. To achieve this, neighboring frame features are fused with the reference frame, and the resulting features are passed through the multi-attention blocks to extract missing content from surrounding frames in relation to the center frame. Furthermore, to enhance temporal coherence and alignment, the aligned features are passed through the PCD [19] alignment module. These refined features are then fed into the merge network, which is composed of a series of DSKFRDBs. These DSKFRDBs with dilation convolutions helps in recovering details due to over-exposure and motion of objects by enlarging the receptive field and ultimately estimating high-quality HDR video.

Motivated by the work of Ledig et al. [20] and Yan et al. [17], global residual learning strategy was adopted by adding the shallow reference frame feature F_r to OF_5 before reconstructing the final HDR frame. Our proposed method predicts blending weights (see Section 4) and produces a 15 channel output. The input images are averaged using their blending weights to obtain the final HDR_i image at frame i.

3.3 *Image Alignment Using Optical Flow*

We adopted the optical flow network of Chen et al. [15] for efficient frame alignment. Alignment of frames is done in the initial phase of learning-based techniques with the reference frame L_i. Flows are estimated for neighbouring frames L_{i-1} and L_{i+1}, in relation to the reference frame, L_i. The nearby frames L_{i-1} and L_{i+1} are then warped with the help of two estimated flows to set a series of aligned images L_{i-1}, i and L_{i+1}, i in relation to the reference frame L_i for efficient treatment of non-rigid motion and the inaccuracies introduced by global alignment.

3.4 *Multi-Attention Guided Feature Alignment*

The attention-guided blocks were given five 6-channels input frames in linear and LDR domain F_i, where $i = 1, 2, 3, 4, 5$. First neighbouring input frames $F_i - 2$, $F_i - 1$ and $F_i + 1$, $F_i + 2$ were concatenated and fused (see Fig. 2) before passing to the attention blocks.

3.4.1 *Channel Attention*

We make use of channel attention proposed by Woo et al. [57] to take advantage and exploit dependencies among features across channels. The architecture of the channel attention network is represented in Figure 3.

In the channel attention blocks, spatial information is collected from the feature maps through both average and max-pooling operations, resulting in two sets of features F_{avg} and F_{max} (refer to Eq. (2)). These two sets of features are then fed into a shared Multi-Layer Perceptron (MLP) network with one hidden layer for providing attention-guided weights for each channel, represented as $W \in R^{C\times1\times1}$. The MLP's hidden layer parameter size is set to $R^{C/r\times1\times1}$, where r (reduction ratio) is utilized to reduce and control the size

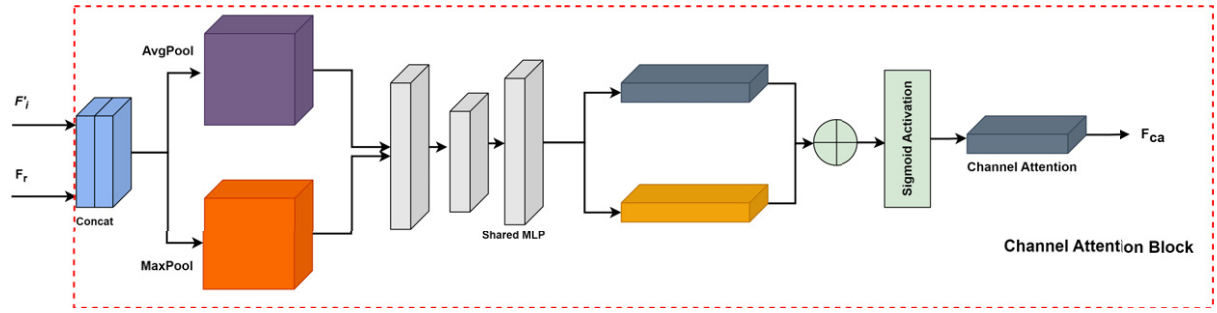

Figure 3. The channel attention sub-module uses a combination of max and average pooling, alongside a shared MLP network.

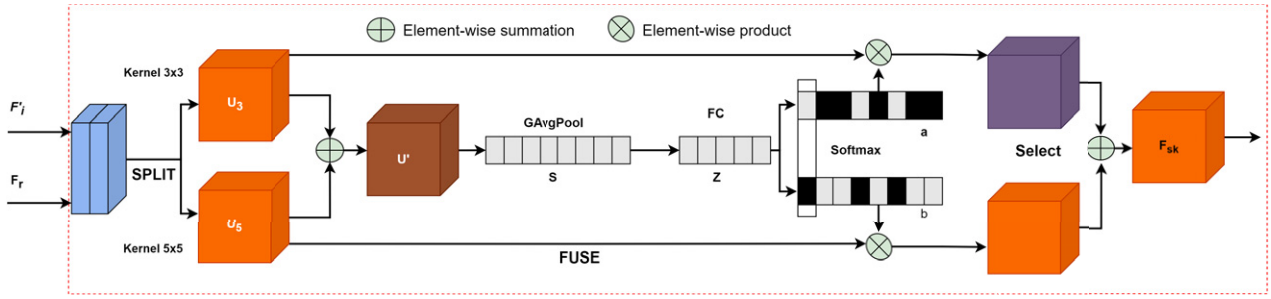

Figure 4. Represents selective kernel fusion attention block involving three main operations specifically, split, fuse and select.

of parameters in the hidden layer. Finally, the output feature vectors from the shared network (MLP) corresponding to the F_{avg} and $F_{\max}$ features are combined using element-wise summation.

$$\begin{aligned} A_i &= \sigma(MLP(F_{\text{avg}}(F_{ir}))) + MLP(F_{\max}(F_{ir})) \\ &= \sigma(W_1(W_0((F_i, F_r)F_{\text{avg}})) + W_1(W_0((F_i, F_r)F_{\max}))), \end{aligned} \tag{2}$$

where σ denotes the sigmoid function, $W_0 \in R^{C/r \times C}$ and $W_1 \in R^{C \times C/r}$ represent MLP layer weights, and F_{ir} is the concatenated feature by fusing F_i and F_r respectively. The estimated attention maps are point-wise multiplied to attend the features of the non-reference frames via Eq. (3):

$$F_i'' = A_i \circ F_i', \quad (i = 1, 3), \tag{3}$$

where $\circ$ denotes the point-wise multiplication between A_i and F_i', $(i = 1, 3)$. Attention-guided features $F_i'' - 1$ and $F_i'' + 1$ are concatenated and fused with the reference frame F_r to get the final stack of channel attention-guided feature F_{ca} using Eq. (4).

$$F_{sk} = \text{Concat}(F_i'' - 1, F_r, F_i'' + 1), \tag{4}$$

3.4.2 *Soft Attention Using Selective Kernel Fusion*

We utilize the work proposed by [58] as an adaptive soft attention technique. This method involves employing multiple kernels with varying receptive field sizes to effectively capture information from objects of different scales within the input. The selective kernel fusion block consists of three main operations: splitting, fusing, and selecting, as depicted in Figure 4.

3.4.3 *Split*

Through split operation, the incoming features F_i', F_r of size $H' \times W' \times C'$ are transformed to U_3 and U_5 features based on the receptive field sizes of 3×3 and 5×5 and applying efficient depthwise convolutions [59], followed by ReLU activation function performing convolution with dilation size of 2.

3.4.4 *Fuse*

Fuse module adaptively controls the information flow of different scales of the two branches that have different receptive fields into the activation functions in the upcoming layer.

The data from the two branches is combined via element-wise summation. Following this, global average pooling is applied to incorporate global information and produce channel-wise statistics represented as $S \in R^C$ (see Eq. (5)).

$$S = F_{gp}(U') = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U'(i, j), \tag{5}$$

Moreover, the feature vector obtained from global average pooling is then fed into a fully connected layer to enable accurate and adaptive feature selection, resulting in $Z \in R^{d \times 1}$. Additionally, a dimensionality reduction parameter is incorporated in Eq. (6) for improving the

efficiency of the attention block.

$$Z = F_{fc}(S) = \delta((WS)), \tag{6}$$

where δ is the ReLU function and $W \in R^{d} \times^{C}$ represent fully connected (fc) layer parameters.

$$d = \max(C/r, L), \tag{7}$$

where C represents channel and d represents reduction ratio which is controlled by parameter r for modifying the parameter size of the fully connected layer and $L = 32$ represent the minimal value of variable d.

3.4.5 *Select*

The last step involves the adaptive selection of informative content from the guided feature descriptor Z by applying a channel-wise softmax operator, as described in Eq. (8). Focusing on different scales of valuable information.

$$a = \text{softmax}(Z), \quad b = \text{softmax}(Z) \tag{8}$$

The softmax-based attention-guided feature maps are multiplied with U_3 and U_5 features which was retrieved previously through split process and then summed to obtain the final attention-guided feature map using Eq. (9).

$$A_i = a \cdot U_3 + b \cdot U_5, \tag{9}$$

where A_i represents soft attention-guided features that are then pointwise multiplied with non-reference features F_i' using Eq. (10).

$$F_i'' = A_i \circ F_i', \quad (i = 1, 3), \tag{10}$$

Attention guided features $F_i'' - 1$ and $F_i'' + 1$ are concatenated and fused with the reference frame F_r to get selective kernel fusion based soft attention guided features F_{sk} by using Eq. (11).

$$F_{sk} = \text{Concat}(F_i'' - 1, F_r, F_i'' + 1), \tag{11}$$

3.4.6 *Spatial Attention*

We also utilize the findings of [17] to acquire spatial attention maps for the non-reference frames as depicted in Fig. 2. Fused features F_i', $i = 1, 3$ of the non-reference images are introduced to the convolutional attention module $a_i(\cdot)$, $i = 1, 3$ along with the reference frame feature map F_r, obtaining attention maps A_i, $i = 1, 3$ for the non-reference frames using Eq. (12).

$$A_i = a_i(F_i', F_r), \quad (i = 1, 3). \tag{12}$$

The predicted attention maps are used to attend to the features of the non-reference images via Eq. (13):

$$F_i'' = A_i \circ F_i', \quad (i = 1, 3), \tag{13}$$

where $\circ$ denotes the point-wise multiplication between A_i and F_i', $(i = 1, 3)$. The F_i'' denotes the feature maps with attention guidance. The reference feature map F_r (i.e. F_i) and the attention-guided features of the non-reference images $F_i'' - 1$ and $F_i'' + 1$ are stacked and fused to get the final 64 channel attention-guided feature map F_s.

3.5 *Refined Deformable Feature Alignment*

Recently, for the task of video super-resolution, researchers [56] introduced deformable convolution, which has been effectively employed by [19] and [60]. The fundamental idea behind deformable alignment is to predict an offset using an offset prediction module defined by Eq. (14). This module employs general convolutional layers and takes two features as input, our fused features F_s, F_{ca}, and F_{sk}, along with a reference frame feature map F_i.

$$\Delta p_i - 1 = \text{func}([\text{fused}(F_s, F_{ca}, F_{sk}), F_i]) \tag{14}$$

After acquiring the learned offset, the fused multi-attention guided features F_s, F_{ca}, and F_{sk} can be sampled and aligned to the reference frame F_i using deformable convolution introduced by [56] using Eq. (15):

$$\tilde{F}_i = DF\text{Conv}(\text{fused}(F_s, F_{ca}, F_{sk}), \Delta p_i - 1). \tag{15}$$

The overall structure of the PCD alignment module is represented in Figure 5 where the alignment is performed at multiple scales between the fused refined features and the reference frame. The final HDR video reconstruction is optimized by implicit learning capabilities of deformable convolution offsets for this alignment process.

3.6 *Merge Network for HDR Image Reconstruction*

The primary goal of the merge network is to reconstruct a high-quality HDR frame using attention-guided aligned features. This network is designed to identify and eliminate any alignment artifacts that may still be present in the registered images and to restore missing content in the over and under-exposed regions, resulting in the final HDR image.

We introduce the selective kernel fusion network, which is based on a residual dense network architecture, similar to the approach presented in Ref. [61]. Our merge network comprises convolution layers and DSKFRDBs with the incorporation of skip connections, as illustrated in Figure 6.

The merge network takes the stacked features from the PCD alignment module. The merge network first applies a conv layer to produce 64-channel feature maps. These feature maps are then passed to three DSKFRDBs outputting three corresponding feature maps OF_1, OF_2 and OF_3. All three feature maps are then concatenated to get OF_4. Then convolution operations are applied for extracting more relevant information from all the three merged feature maps produced from DSKFRDBs to get OF_5.

3.6.1 *Global Residual Learning With the Reference Features*

Motivated by the work of [17, 20], global residual learning strategy was adopted by adding the shallow reference frame feature F_r to OF_5 where the representation of the original reference information is integrated before reconstructing the

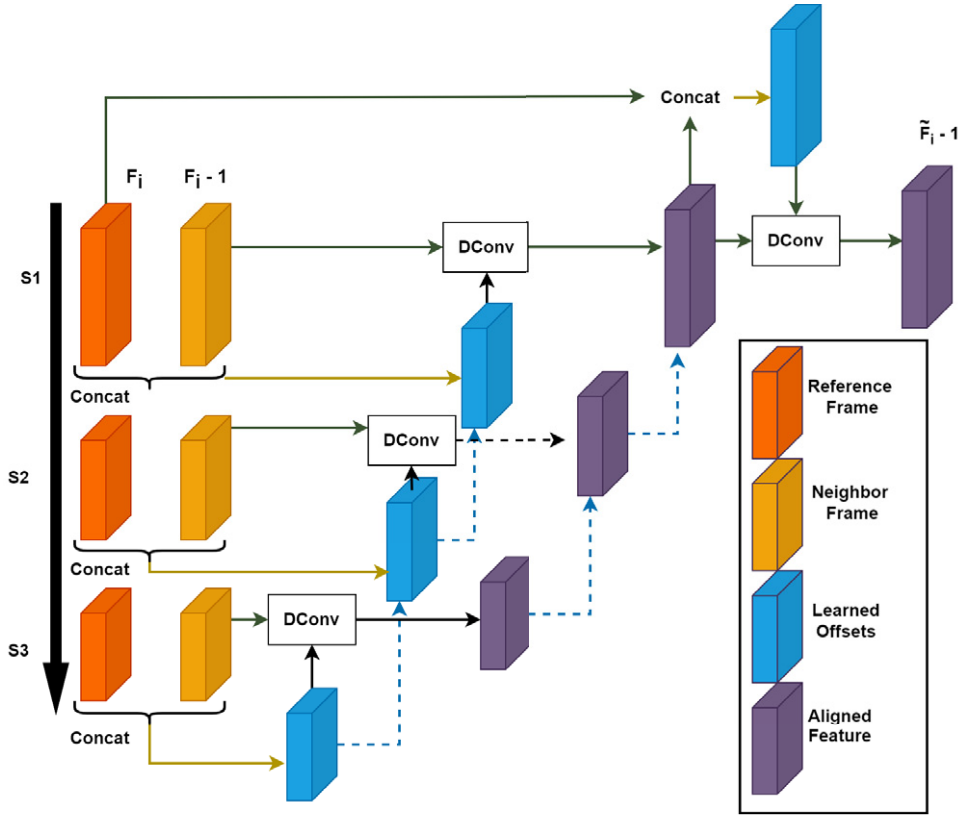

Figure 5. Represents architecture of PCD [19] alignment module.

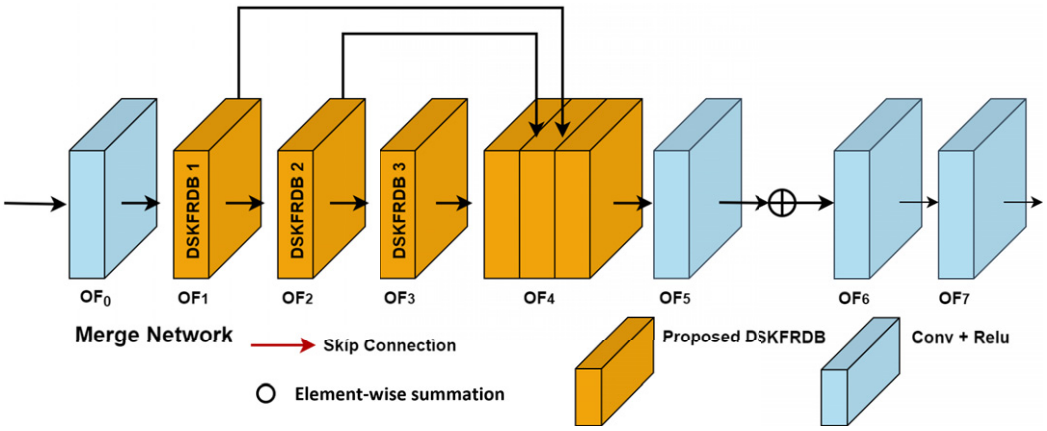

Figure 6. Represents merge network composed of series of dilated selective kernel fusion residual dense blocks with skip connections.

final HDR image from OF_5 to optimize the accuracy of the model.

$$OF_6 = OF_5 + F_r, \tag{16}$$

The final feature map $OF6$ contains almost all the ingredients for reconstructing the final HDR image without ghosting artifacts with details recovered in over and under-exposed regions with large motion. The final HDR image is estimated in the HDR domain after two convolution layers followed by activation funtion.

3.6.2 *Dilated Selective Kernel Fusion Residual Dense Block*

The merge network requires a larger receptive field for hallucinating details since the reconstruction of some local regions of the HDR images cannot receive enough information from the LDR images due to the occlusion of moving objects and saturation. Therefore, we used a DSKFRDB having two branches with dilation. The proposed DSKFRDB, which is represented in Figure 7, perform final HDR video reconstruction by adaptive feature selection using two different receptive fields using the split, fuse, and select strategy with dense concatenation based

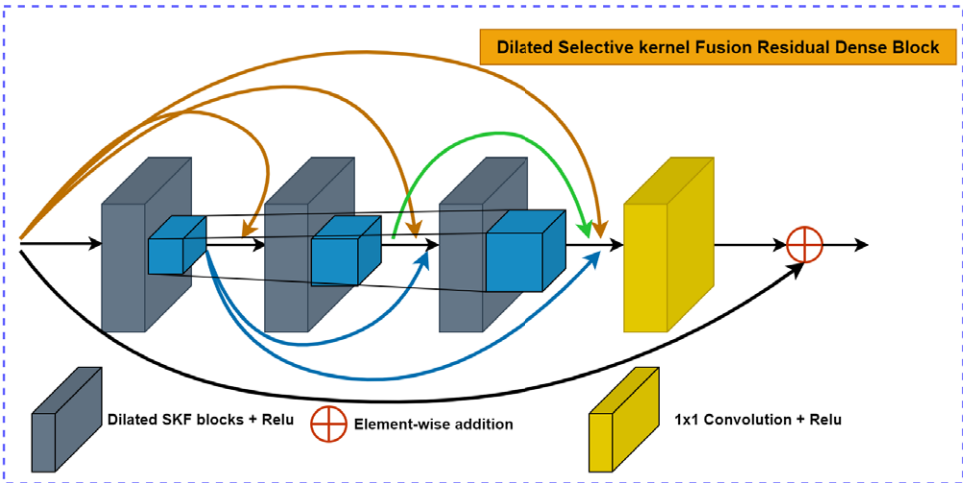

Figure 7. Illustration of a three-layer dilated selective kernel fusion residual dense block structure following the residual dense block strategy of [61] as a framework.

skip-connections where the input for each layers is the concatenation of all feature maps from preceding layers.

4. PIXEL BLENDING

To our full multi-Attention SKFHDRNet, we provided five 6-channel input images in both LDR and linear domains making a 30 channel input. Then, for these five images, our network predicted the blending weights and produced a 15 channel output. To effectively utilize the information in each color channel, we estimated blending weights for each color channel in a manner similar to the methods proposed by [41, 62]. The five input images are averaged using their blending weights to get the final HDR image HDR_i at frame i by using Eq. (17).

$$HDR_i = \frac{w_1 L_i - 1 + w_2 \hat{L}_i - 1 + w_3 L_i + w_4 \hat{L}_i + 1 + w_5 L_i + 1}{\sum_{K=1}^{5} wk}, \tag{17}$$

where, wk is the estimated blending weight for each image.

5. LOSS FUNCTION

Following the works of [14, 15, 36] the linear HDR images are transformed into log domain for boosting the pixel values in the dark regions of the image. Directly applying the loss function on the images in the linear HDR domain will produce inaccuracies by underestimating the error in the pixel values of the dark regions. We specifically employ the differentiable μ-law function using Eq. (18):

$$T_i = \frac{\log(1 + \mu HDR_i)}{\log(1 + \mu)}, \tag{18}$$

where HDR_i represent linear HDR frame with the pixel values in range of [0, 1]. The parameter μ is set to 5000 to control the rate of compression range. The model parameters are updated by minimizing the L_1 distance between the estimated, $\hat{T}_i$, and ground truth, T_i, HDR frames in the log domain with Eq. (19):

$$E = \|\hat{T}_i - T_i\|_1. \tag{19}$$

5.1 L_1*MS–SSIM Loss Function*

According to [21], MS–SSIM preserves the contrast in high-frequency regions better than the other loss functions. On the other hand, L_1 preserves colors and luminance and error are weighted equally regardless of the local structure but does not produce quite the same contrast as MS–SSIM. To capture the best characteristics of both error functions, [21] propose a combined L_1MS–SSIM loss function which is represented by Eq. (20):

$$L_{\text{mix}} = \alpha L_{MS-SSIM} + (1 - \alpha) G^M_{\sigma\, G} \cdot L_1, \tag{20}$$

where α is empirically set to 0.84 with point-wise multiplication between $G^M_{\sigma\, G}$ and L_1. $G^M_{\sigma\, G}$ which represents the computation of mean and standard deviations with a Gaussian filter. We adopted the work of [21] to optimize the training of our model. The parameters or weights of the networks are modified using these computed gradients continuously until convergence.

6. IMPLEMENTATION DETAILS

PyTorch framework was used to implement the Multi-Attention SKFHDRNet model architecture. We integrated the flow network implemented by [15] using Pytorch into our pipeline for HDR video reconstruction. End-to-end training is done for both optical flow and multi-attention SKFHDRNet. The technique used by [63] is used to initialize the initial weights of the network parameters. Using ADAM with default settings of $\beta 1 = 0.9$ and $\beta 2 = 0.999$ with a learning rate of 0.0001, to solve the optimization problem. Mantiuk et al. [64] approach was used for tone-mapping the results. Given training images, we randomly crop the images

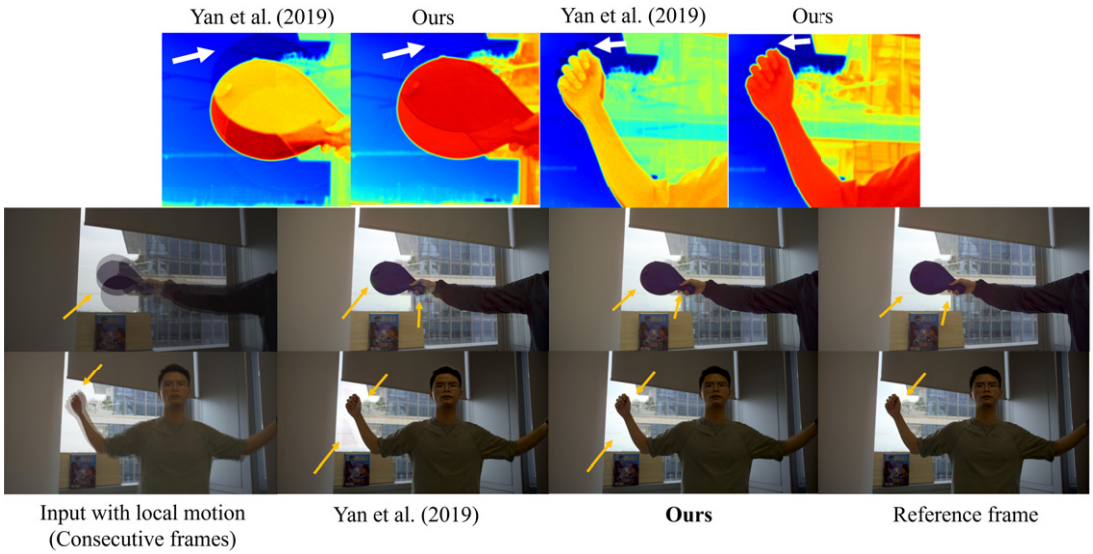

Figure 8. Visualization of the model's outputs having consecutive frames as an input with local motion and a conv layer feature maps after passing through attention modules.

of size 256 × 256 for training. The model was trained for 20 epochs on two NVIDIA Tesla V100 32 Gb of NTNU cluster [65].

7. EXPERIMENT RESULTS

We conducted experiments and performed an evaluation on synthetic test HDR scenes and real-world dataset (dynamic and static scenes from [15], under CC BY-NC-SA 4.0 license) to verify the effectiveness of the proposed method. All models are visually compared, and the predicted HDR frame is evaluated in terms of multiple image quality metrics. We specifically used μ-law tone-mapped PSNR, HDR–VDP2 [40] and HDR-VQM [66] (HDR-VQM for full model comparison). We followed the HDR-VQM design of [15] to assess the quality of HDR videos. Additionally, all models were evaluated based on color difference error between estimated and ground truth HDR using CIEDE2000 [67]. All visual results in the experiment are tone-mapped using Mantiuk et al. [64] tone-mapping method.

7.1 *Evaluation of Baseline Models*

We performed our initial comparisons with [17] in the case of no optical flow and no pixel blending where the model estimated a 3-channel final HDR image. This was specifically done to check and compare the effectiveness of our proposed attention modules against [17] AHDRNet. The proposed attention modules effectiveness is represented in Figure 8 indicating better performance in frame alignment against reference frame with less ghosting artifacts in comparison to AHDRNet [17] attention module. In Fig. 8, this is seen especially in the hand and racket with fewer ghosting artifacts.

Similarly, the robustness of our proposed DSKFRDBs in filling rich details of over-exposed regions is illustrated in Figure 9 against AHDRNet [17]. Our proposed DSKFRDBs enable the model to produce results with rich details while achieving more accurate content in over-saturated areas. This can be seen by the proposed model having less color difference in the highlights compared to AHDRNet.

Similarly, the zoomed regions of the CAROUSEL FIREWORKS frame represented in Figure 10 show poor performance of Yan et al. [17] AHDRNet. It struggles in reducing ghosting artifacts due to large motion which ultimately introduces higher color difference errors, which can be seen in color difference maps of the images. However, our proposed Multi-Attention SKFHDRNet model performed better alignment in case of large motions and produced a smaller color difference error in relation to the ground truth HDR frame.

Our baseline model SKFHDRNet performed fairly well in case of the static dataset. From the visual results, the Yan et al. [17] model struggled to recover details in over-exposed regions which are illustrated in the zoomed regions of static dataset scene in Figure 11. Multi-attention SKFHDRNet recovers much of the missing information in the over-exposed regions with a small color difference error as shown in Fig. 11. This indicates that using DSKFRDBs in the merge network for filling missing content in the over-exposed regions works better compared to the dilated residual dense block of [17].

Quantitative results in terms of μPSNR and HDR–VDP2 are represented in Table I. Our multi-attention SKFHDRNet showed better performance in terms of visual results as well as image/video quality metrics, where the values are higher than Yan et al. in all datasets for both μPSNR and HDR–VDP2. This indicates our multi-attention modules efficiency which guides more relevant features from the neighbouring frames in relation to the reference frame and robustness of our DSKFRDBs in merge network in filling missing content in the over-exposed regions.

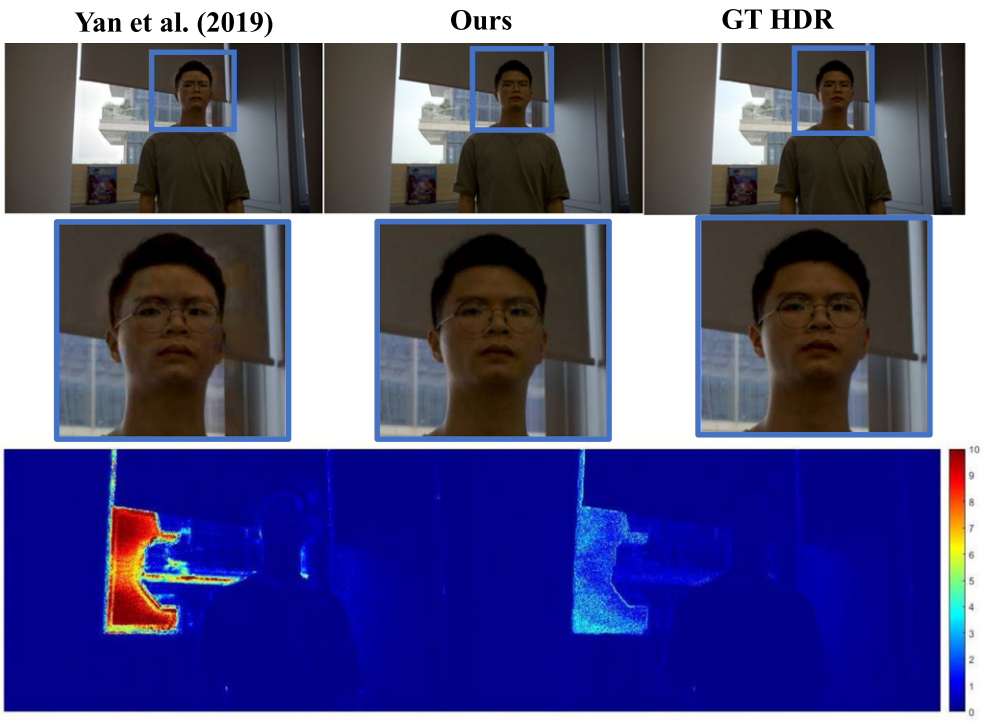

Figure 9. Dynamic scene Ground Truth (GT) test sample and its estimated HDR scene of AHDRNet [17] and our proposed multi-attention SKFHDRNet. The top shows the full image, the middle images are a zoomed in area, and the bottom show the CIEDIE2000 color difference map.

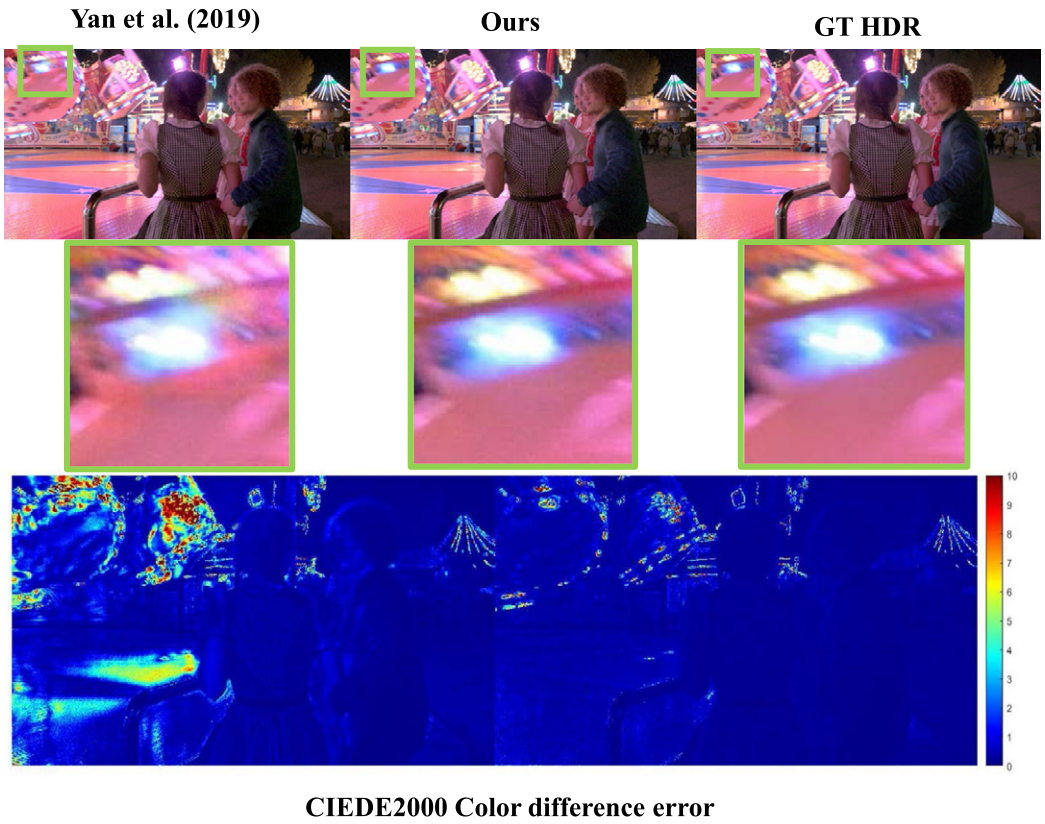

Figure 10. Represents visual and color difference error results of baseline models on synthetic dataset (CAROUSEL FIREWORKS) scene.

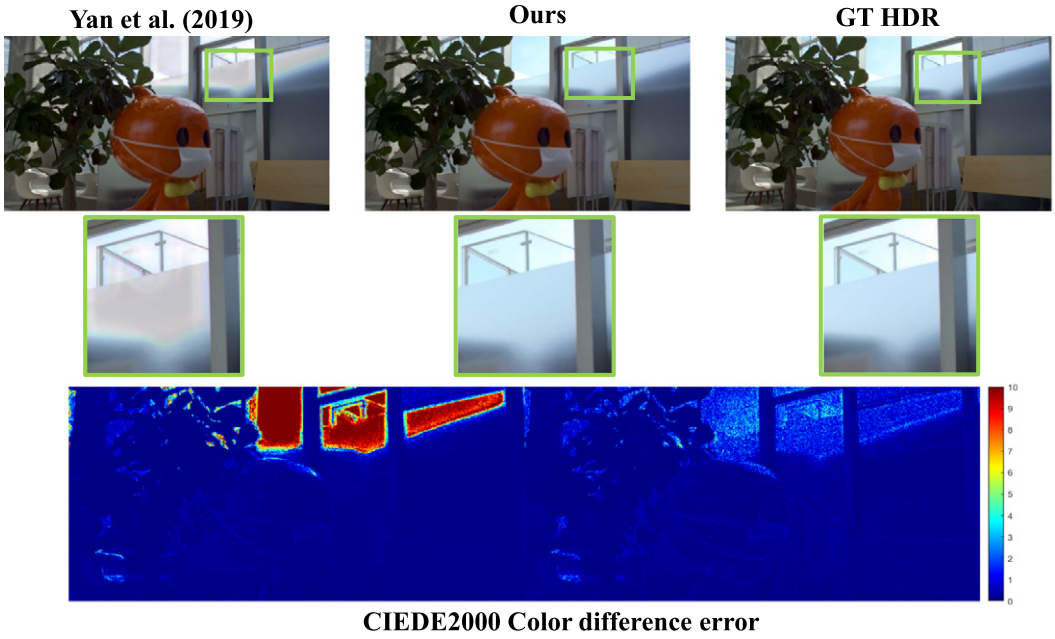

Figure 11. Represents visual and color difference error results on the static dataset scene.

Table I. Quantitative results of our baseline Multi-attention SKFHDRNet and Yan et al. [17] AHDRNet on test datasets are represented. Bold text indicates the better among models.

Models	μPSNR	HDR–VDP–2
Model performance on synthetic dataset		
Yan et al. [17]	28.78	63.56
Multi-attention SKFHDRNet	**32.11**	**65.65**
Model performance on dynamic dataset		
Yan et al. [17]	34.68	68.42
Multi-attention SKFHDRNet	**40.77**	**73.81**
Model performance on static dataset		
Yan et al. [17]	33.06	69.81
Multi-attention SKFHDRNet	**36.76**	**71.34**

7.2 *Per Frame Objective Metric Results Visualization of Our Baseline Model Without Optical Flow and Pixel Blending*

Figure 12 represents our baseline model performance in relation to Yan et al. [17] AHDRNet on all the three datasets. Blue violin plots represent [17] model and orange violin plots represent our baseline Multi-Attention SKFHDRNet. The data points represent per frame image quality metrics results specifically, μPSNR and HDR–VDP–2. The median is represented by (the red point), and the first and third quartile are represented by the black bar where the lower region of the bar represent first quartile and the upper region of the black bar represents the third quartile. Our baseline model predicted better per frame quality metrics' results considering the median in a violin plot which is higher than Yan et al. [17] AHDRNet on all three datasets. From the results, an intersection between data points can be clearly seen, especially in case of synthetic and dynamic datasets. This represents the performance of models on low and high exposure samples. The model shows higher performance for samples with low exposure, which is represented mostly in the third quartile region of the violin plot above the median. Samples with center frame having high exposure are represented below the median red point in the first quartile region of violin plot. It is worth noting that the proposed model is able to generate higher values in the synthetic dataset, as seen in HDR-VDP-2, the lowest values are approximately the same between the two models, but the proposed model has higher maximum values. In μPSNR we see a shift from a bottom heavy distribution to values being increased. For the static dataset, we see a similar behaviour for HDR-VDP-2, with a larger concentration of values being towards the top end, while in μPSNR it is in general, a shift upwards. Lastly for the dynamic dataset, the proposed model shifts the values upwards for μPSNR with more values concentrated towards the higher end, while in HDR-VDP-2 the values have a larger spread with more values being above the highest values in AHDRNet. In general, our model performance based on μPSNR and HDR–VDP–2 was higher than Yan et al. [17] AHDRNet.

7.3 *Evaluation of Our Full Model*

We compared our full model performance with [13, 14, 17] and [15] along with its individual networks CoarseNet and RefineNet. We re-implemented Yan et al. [17] method for alternating-exposure HDR video reconstruction and used the already trained Chen et al. [15] network parameters for comparison. For Kalantari et al. [13] and Kalantari and Ramamoorthi [14], we took the results of the model from [15] since the same datasets are used for comparison.

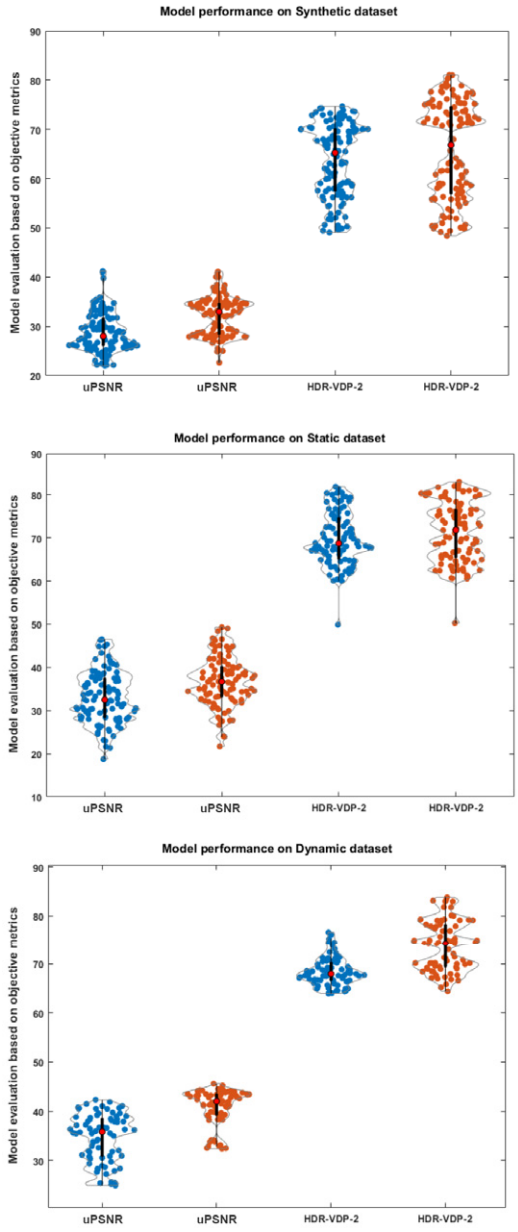

Figure 12. Per frame representation of image quality objective metric results on all three datasets using violin plot of our baseline architecture (orange) against Yan et al. [17] AHDRNet (blue).

All models are visually compared and the predicted HDR image is evaluated in terms of multiple image quality metrics. We specifically used μ-law tone-mapped PSNR, HDR–VDP2 [40] and HDR-VQM [66]. Additionally, all models were evaluated based on the color difference error between estimated and ground truth HDR using CIEDE2000 [67].

7.4 *Synthetic Dataset for Training*

Following the work of [13–15], we used 13 HDR video scenes from [26] and eight downsampled video scenes of resolution 1280×720 from [68] for training purposes. Furthermore, we also used a high-quality Vimeo-90K [69] dataset as training samples similar to [15] due to the limited size of the training HDR video dataset.

7.5 *Evaluation on Synthetic Dataset*

Our proposed multi-attention SKFHDRNet with re-implemented AHDRNet [17] is evaluated on a synthetic test dataset which is composed of two HDR videos (i.e., POKER FULLSHOT and CAROUSEL FIREWORKS) of [26] HDR dataset with random Gaussian noise added to low-exposure images like [15].

Figure 13 illustrates the model performance on POKER FULLSHOT HDR scene. From the visual results, the color difference error is more prominent in Yan et al. [17] AHDRNet estimated HDR image. The reconstructed scene is noisy and the color difference map shows error across the scene. Similarly, there is higher color difference error in the saturated regions specifically in the edges and the curtain of the table in the scene reconstructed by [15] where some pixels are still over-saturated which is detected by the CIEDE-2000 color difference metric. However, the reconstructed HDR scenes of our model variants have less over-saturated pixels in the edges and the curtain on the table. This indicates DSKFRDB's robustness to filling rich details in the over-exposed regions with 50% less model parameters compared to Chen et al. [15] and providing better performance in accuracy.

Quantitative results using HDR-VDP2, HDR-VQM and $\mu PSNR$ of our multi-attention SKFHDRNet variants on the synthetic dataset are presented in Table II.

Our multi-attention SKFHDRNet showed better performance on all three image and video quality metrics. This indicates our multi-attention modules' efficiency regarding noise reduction and filling details in over-exposed regions.

7.6 *Evaluation on Real World Static Dataset*

We test our multi-attention SKFHDRNet variants on a static dataset that is composed of random global motions. Random translation was performed for each frame in the range of [0, 5] pixels. For all methods, no pre-alignment is done on input frames similar to Chen et al. [15] to evaluate their robustness to input with inaccurate global alignment. The Yan et al. [17] model produce results with noise and the error is captured and visualised in Figure 14 in the color difference error map. While Chen et al. [15] model produce results with out noise in the reconstructed frame but showed higher color difference error in the over-saturated regions in the scene which can be seen in the color difference error maps represented in Fig. 14. Our model variants produce better performance in case of noise and filling rich details in over-saturated regions producing smaller color difference error.

Similarly [15] model struggle to perform proper alignment in the zoomed and highlighted regions in Figure 15. The straight lines are distorted in the highlighted region of [15] reconstructed frame. In case of Yan et al. [17] model, apart from distortions in the straight lines, there are also more prominent color fringe patterns in the highlighted and zoomed region shown in Fig. 15. However, our proposed model variants showed better performance with reduced distortion and without prominent color fringe patterns in the

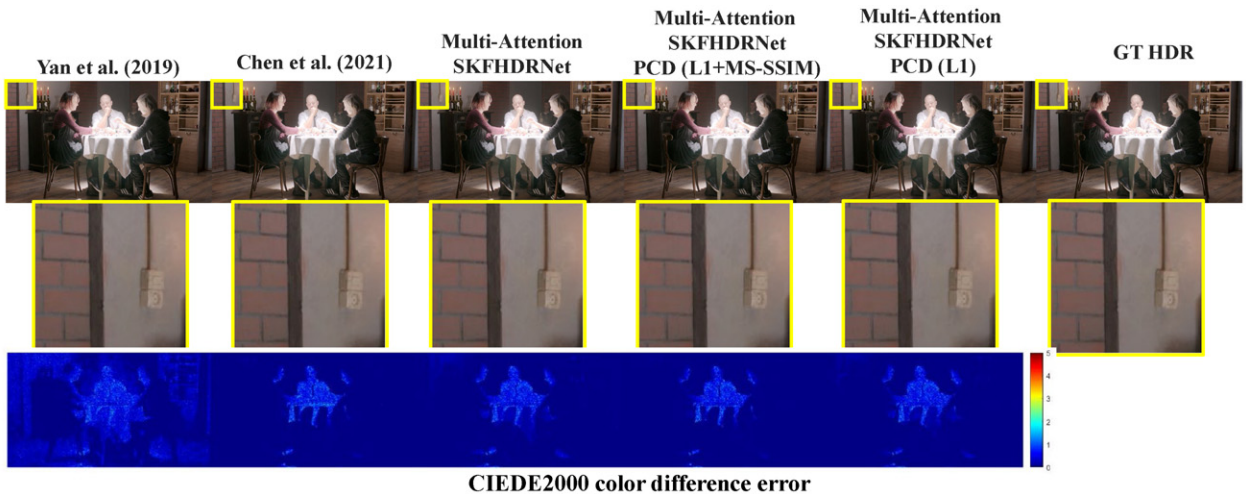

Figure 13. Visual and color difference error results on the synthetic dataset.

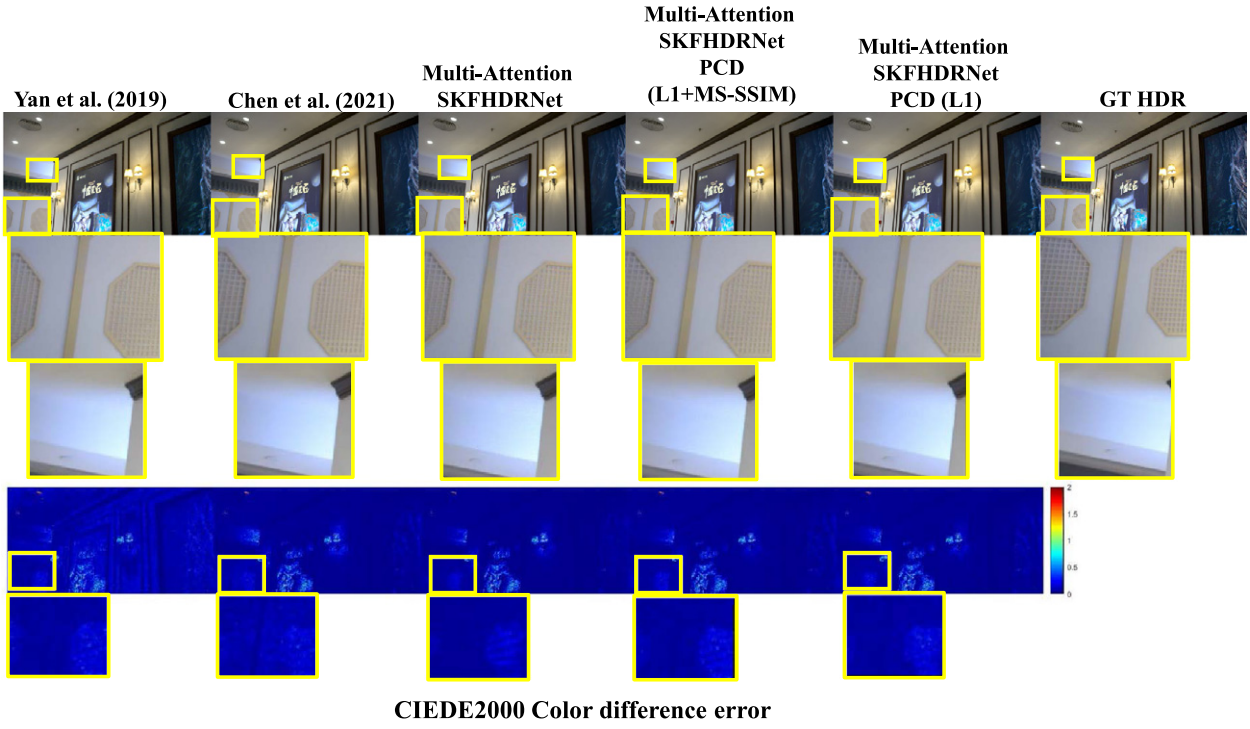

Figure 14. Represents visual and color difference error results on the static dataset.

highlighted region of reconstructed frame and the error is recorded by CIEDE2000 color difference error maps.

Our multi-attention SKFHDRNet variants performed better than Yan et al. [17] AHDRNet and [13, 14] learning-based methods using objective image and video quality metrics represented in Table II. Our models also performed better than the [15] single models (CoarseNet and RefineNet). However, the model of Chen et al. [15] showed slightly better results compared to our multi-attention SKFHDRNet variants based on image/video quality metrics.

Our proposed model showed comparable results on static scenes in comparison to prior work with half the size of network parameters than [15] full model, which can be seen in Table II.

7.7 *Evaluation on Real World Dynamic Dataset*

The dynamic dataset contains large local motions, making it challenging for the models to perform well in these cases. Figure 16 visualizes the results of our multi-attention SKFHDRNet variants along with [17] and [15] models. All of our models clearly show high performance in large local motion regions in the dynamic dataset scene, apart from our model variant SKFHDRNet having L_1 and MS-SSIM loss, which can be seen in the zoomed region of the dynamic dataset scene in Fig. 16. The arrow pointing to

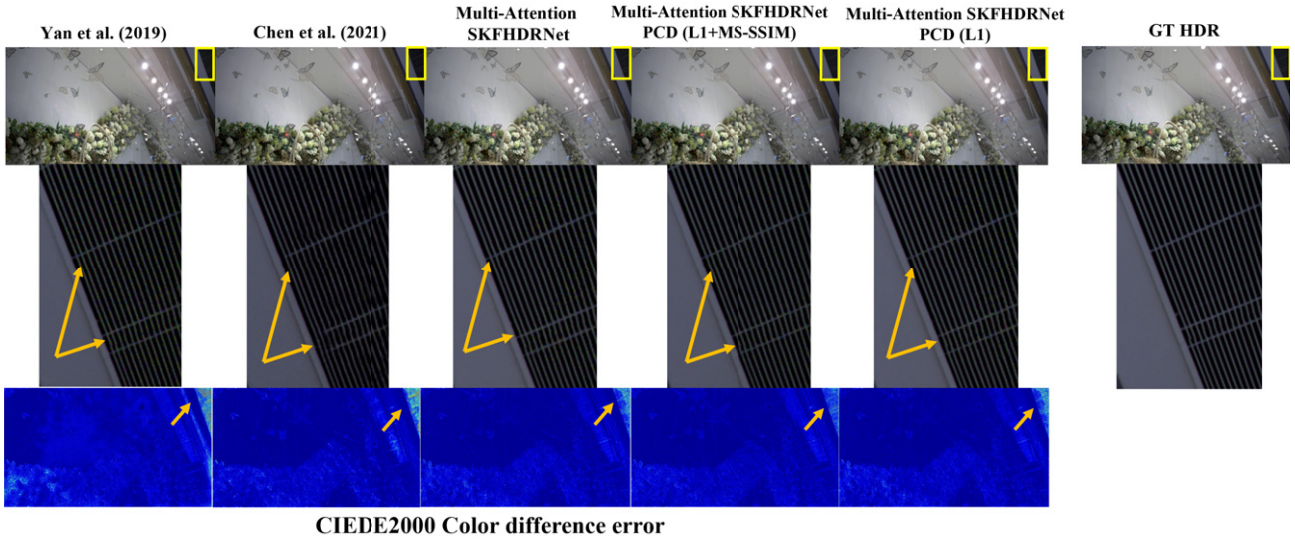

Figure 15. Represents visual and color difference error results on the static dataset.

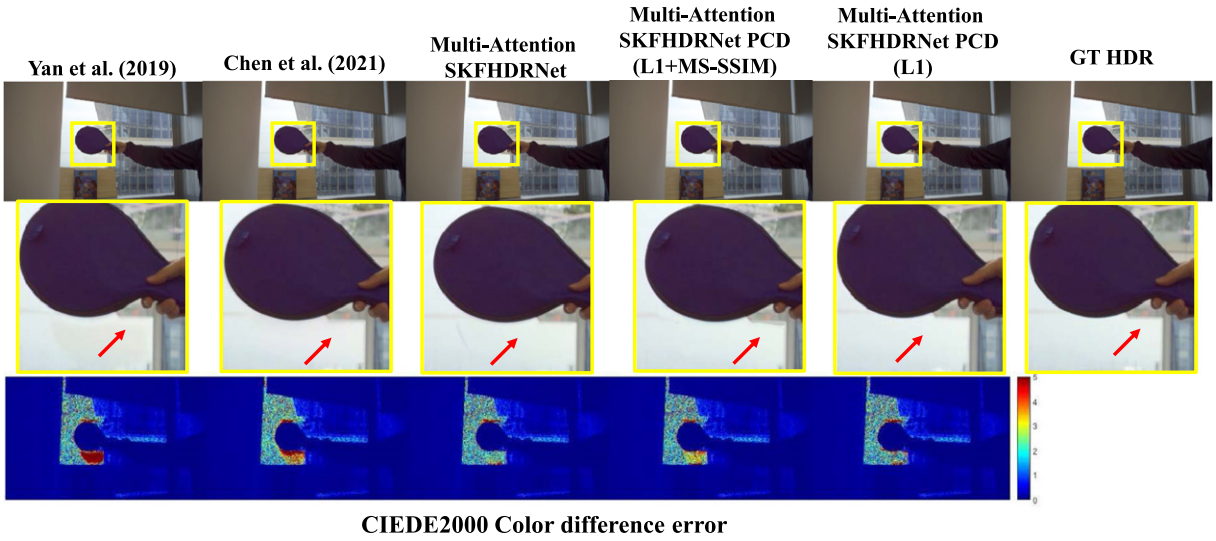

Figure 16. Represents visual and color difference error results on the dynamic dataset.

Table II. Quantitative results of our multi-attention SKFHDRNet variants on all three datasets. The best model is represented with red text, the second best model is represented by blue text, and the third best model is represented by green text.

	Synthetic dataset			Static dataset			Dynamic dataset		
Models	μPSNR	HDR–VDP2	HDR–VQM	μPSNR	HDR–VDP2	HDR–VQM	μPSNR	HDR–VDP2	HDR–VQM
Kalantari et al. [13]	37.53	59.07	84.51	40.02	71.89	76.22	41.72	70.36	85.33
Yan et al. [17]	36.49	71.01	69.68	38.05	74.73	64.33	42.76	78.69	87.27
Kalantari and Ramamoorthi [14]	37.48	70.67	84.57	39.88	74.13	73.84	44.72	77.91	87.16
Chen et al. [15] CoarseNet	39.25	70.81	–	40.62	74.51	–	44.43	77.74	–
Chen et al. [15] RefineNet	39.69	70.95	–	37.61	75.30	–	43.70	78.97	–
Chen et al. [15] Full model	40.34	71.79	85.71	41.18	76.15	78.84	45.46	79.09	87.40
Ours SKFHDRNet (L_1)	39.48	71.42	82.22	40.20	75.23	74.05	45.43	79.12	88.44
Ours SKFHDRNet (PCD + L_1 + MS–SSIM)	39.92	71.68	86.04	40.47	75.36	75.67	44.96	78.49	86.96
Ours SKFHDRNet (PCD + L_1)	39.94	71.95	84.05	40.62	75.32	75.42	45.53	78.89	87.80

regions is where we can see the ghosting artifacts and blur in the reconstructed scene of [15] results. Similarly, there is ghosting artifact of whole racket in the reconstructed scene of [17] results. This shows our multi-attention and PCD module effectiveness regarding feature alignment of neighbouring frames in to the reference frame. The color difference error maps also show large deviation in color information from the original HDR image in the motion regions of the estimated HDR frames of [17] and [15] models.

The performance of our proposed model variants was better than Yan et al. [17] AHDRNet, [13, 14], and Chen et al. [15] learning-based methods using objective image/video quality metrics on dynamic dataset represented in Table II. Our models also showed better performance than the [15] single models (CoarseNet and RefineNet).

This again indicates our model's DSKFRDB fusion block effectiveness in filling the missing content in large over-exposed regions with local motion (see results in Table II).

7.8 *Per Frame Objective Metric Results Visualization of Our Full Architecture*

Figure 17 represents violin plots of our multi-attention SKFHDRNet variants specifically, multi-attention SKFHDRNet with L_1 loss, multi-attention SKFHDRNet with L_1 loss and PCD alignment module, multi-attention SKFHDRNet with L_1MS–SSIM loss along with PCD alignment module. The performance of the mentioned model is compared to [15] network.

Fig. 17 represents violin plot where the blue violin plots represent our multi-attention SKFHDRNet with L_1 loss. The orange violin plot represents multi-attention SKFHDRNet with L_1 loss and PCD alignment module. The yellow violin plot represents multi-Attention SKFHDRNet with L_1MS–SSIM loss function and PCD alignment module. Purple violin plots represent Chen et al. [15] model results. Our model variants produce consistent or in some cases showed better results from [15] full model considering μPSNR and HDR-VDP2 per frame image quality results. By looking at the median point in red, the performance of all models looks almost equivalent. However, in some cases, like the result of our multi-attention SKFHDRNet with PCD and L_1 loss (orange) in terms of HDR–VDP–2 image quality metric, produce better results by considering the median (red point) point of a violin plot. It is also worth noting that for the dynamic dataset (bottom), the Chen et al. [15] model produces higher minimum values than the others for HDR-VDP-2, but the others have slightly higher maximum HDR-VDP-2 values. A similar behaviour can also be seen on the synthetic dataset (top). Overall, the behaviour of all models were similar, where all the models performed well in the case of HDR test scenes with a center frame under-exposed, while producing inferior results in the case of scenes with a center frame highly over-exposed with large motions.

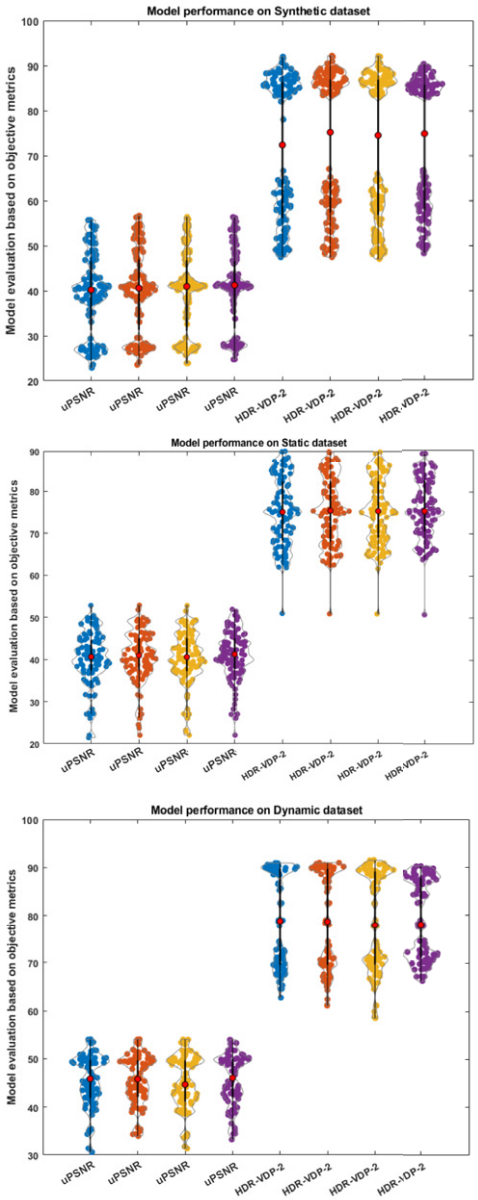

Figure 17. Per frame representation of image quality objective metric results on all three datasets using violin plot of our multi-attention SKFHDRNet variants blue violin representing our multi-attention SKFHDRNet with L_1 loss, orange representing multi-attention SKFHDRNet with L_1 loss and PCD alignment module, and yellow represents multi-attention SKFHDRNet with L_1MS–SSIM loss function and PCD alignment module against purple points of Chen et al. [15] model results.

8. NETWORK PARAMETERS ANALYSIS

The full model of Chen et al. [15] is composed of 6.1 million parameters, with 3.1M parameters for CoarseNet and 3.0M for RefineNet, while Yan et al. [17] model contained 1.9M parameters and Kalantari and Ramamoorthi [14] model had 9.0M parameters mentioned by [15]. However, our full model without the PCD module has 1.3M parameters. Our other model variants having the PCD module have 2.9M parameters providing almost similar or even surpassing performance of Chen et al. [15] model which has network parameters more than half the size of our model. However,

Table III. Performance of the proposed network.

Models	Synthetic dataset 1920 × 1080	Dynamic dataset 1476 × 753	Static dataset 1536 × 813	Network parameters
Kalantari et al. [13]	185 s	–	–	
Yan et al. [17]	0.82 s	0.46 s	0.52 s	1.9M
Kalantari and Ramamoorthi [14]	0.59 s	–	–	9.0M
Chen et al. [15] Full model	0.84 s	0.63 s	0.89 s	6.1M
Our SKFHDRNet (L_1)	1.21 s	0.68 s	0.97s	**1.3M**
Our SKFHDRNet (PCD + L_1)	2.02 s	1.15 s	1.29 s	2.9M
Our SKFHDRNet (PCD + L_1 + MS–SSIM)	2.04 s	1.16 s	1.30 s	2.9M

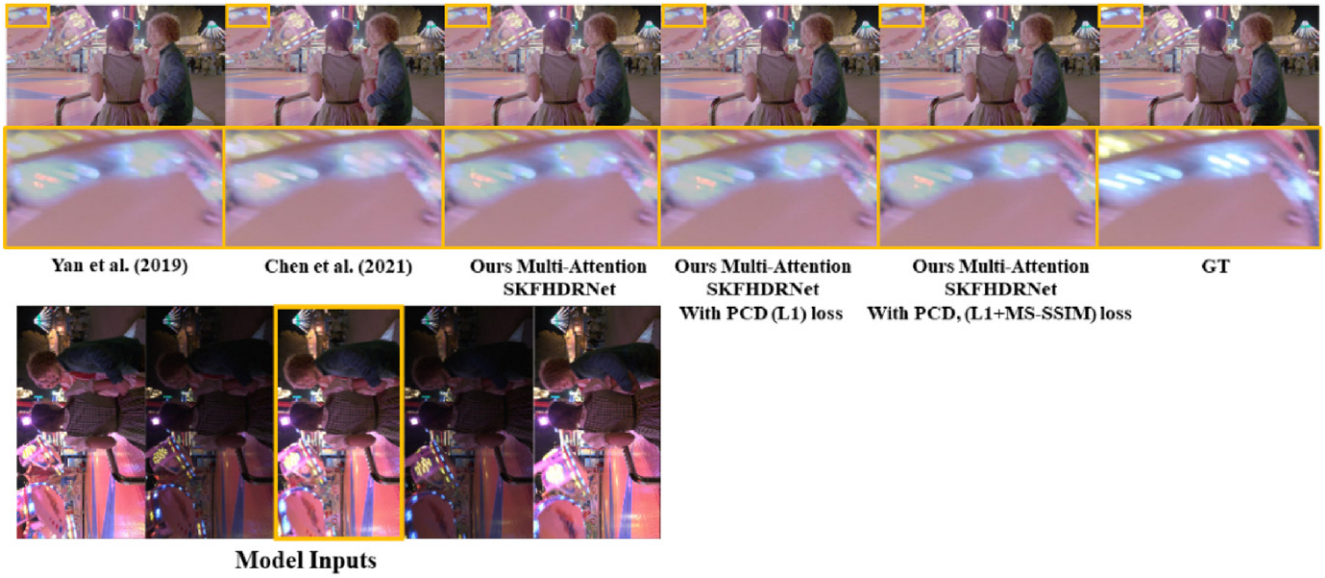

Figure 18. The top row represents estimated HDR scenes for CAROUSEL FIREWORKS scene using two alternating exposures. The bottom row shows the zoomed region where all the models introduced decolorized pixels. By looking at the model inputs, where the center (reference) frame L_i is over-exposed in the highlighted region and the missing content should be recovered from the neighboring frames with low exposure, $L_i - 2$, $L_i - 1$ and $L_i + 1$, $L_i + 2$. Because of significant displacement of objects due to large motions along with high exposure in that region none of the methods are able to properly register and reconstruct details in that region of the image, producing ghosting artifacts which can be seen from the bottom row. Therefore, our method is similar to other approaches and contains artifacts in this region.

our full model variants had a high inference time on the test images, which is represented in Table III.

9. LIMITATIONS OF OUR PROPOSED METHODOLOGY

In general, our approach performs better and produces high quality HDR video. However, some use cases were harder, and the model struggled to produce satisfactory HDR video reconstruction. One typical example of our model poor performance is observed in cases where the center (reference) frame has highly over-exposed regions and there is apparently large movement of objects during consecutive frames with large occlusion. As can be seen in Figure 18, our method results in ghosting and other distortions, such as decolorized pixels. Other methods from the results also encounter difficulties in these regions and provided estimated HDR with a similar type of artifacts.

Moreover, in cases where the center (reference) image has low exposure and the neighboring frames with high exposure contain darker pixels in the same region; this scenario makes it harder for the models to recover detail in darker regions because the information is very limited in all the frames which produce noise in those regions. This is illustrated in the zoomed region of the static dataset scene in Figure 19. However, our full model results are still considerably better than the other learning-based techniques.

10. FUTURE WORK

Considering the real-time scenarios, further research is needed to make the model more interactive by minimizing

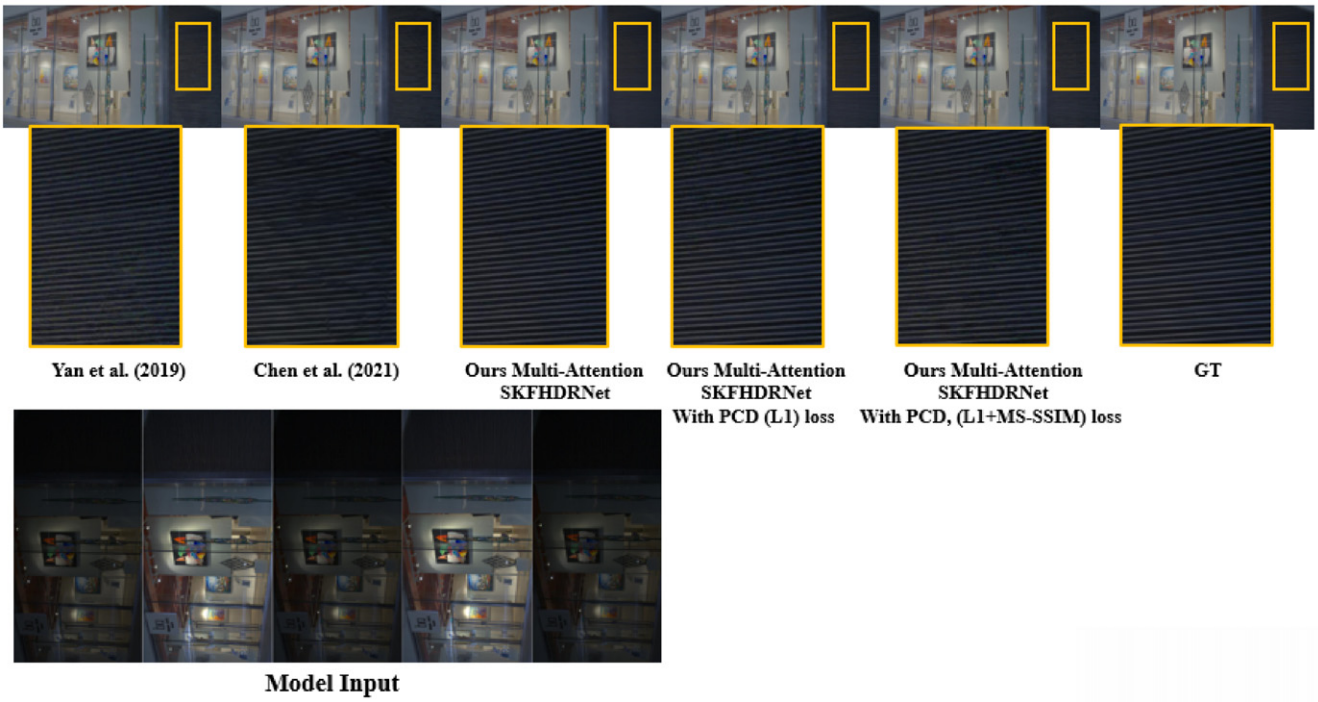

Figure 19. The top row represents the estimated HDR scenes for static scene using two alternating exposures. The bottom row shows the zoomed region where all the models introduced noise in the dark region. By looking at the model inputs, where the center (reference) frame L_i is under-exposed and the highlighted region have very dark pixels. Upon that the neighboring frames with high exposure, $L_i - 2$, $L_i - 1$ and $L_i + 1$, $L_i + 2$ also have darker pixel values in the same regions. Due to less information in the middle as well as neighbouring frames, the models produced noisy texture in those regions which is visualized in the zoomed sections in the bottom row. Therefore, our method, similar to other approaches contains artifacts in this region. However, our multi-attention SKFHDRNet variants have less noisy estimated HDR scene than the other methods.

the inference time of the model. As an example, performing HDR video estimation without an optical flow network will further reduce the model inference time.

Although our methodology showed improved performance regarding recovering details in over-exposed regions of LDR images, further improvement is required as most of the prior work similar to our proposed method showed inferior performance in recovering missing details in challenging over-exposed examples.

In the future, we will extend the evaluation by conducting a psychophysical study to evaluate model performance. Additionally, it would be interesting to modify our system to work with different types of capturing setups, for example, stereo cameras with various exposures.

11. CONCLUSION

We proposed a learning-based technique having optical flow, multi-attention, and PCD alignment modules for improved model performance regarding image alignment and ghosting artifacts. For recovering lost details in under and over-exposed regions, we merged the previously refined aligned features using a series of (DSKFRDBs) for estimating high-quality final HDR scenes. We demonstrate the performance of our method on a number of HDR test datasets containing challenging cases with over-exposed regions and large motions. Our learning-based method achieves better results in most cases than recent state-of-the-art methods with model parameters half the size of the recent state-of-the-art method.

REFERENCES

[1] R. K. Mantiuk, K. Myszkowski, and H.-P. Seidel, *High Dynamic Range Imaging* (Wiley Encyclopedia of Electrical and Electronics Engineering, 2015).

[2] S. J. Daly and X. Feng, "Bit-depth extension using spatiotemporal microdither based on models of the equivalent input noise of the visual system," Proc. SPIE **5008**, 455–466 (2003).

[3] Q. Song, G.-M. Su, and P. C. Cosman, "Hardware-efficient debanding and visual enhancement filter for inverse tone mapped high dynamic range images and videos," *2016 IEEE Int'l. Conf. on Image Processing (ICIP)* (IEEE, Piscataway, NJ, 2016), pp. 3299–3303.

[4] G. Luzardo, J. Aelterman, H. Luong, W. Philips, and D. Ochoa, "Real-time false-contours removal for inverse tone mapped HDR content," *Proc. 25th ACM Int'l. Conf. on Multimedia* (ACM, New York, NY, 2017), pp. 1472–1479.

[5] S. Mukherjee, G.-M. Su, and I. Cheng, "Adaptive dithering using curved Markov–Gaussian noise in the quantized domain for mapping SDR to HDR image," *Int'l. Conf. on Smart Multimedia* (Springer, Cham, 2018), pp. 193–203.

[6] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, "Inverse tone mapping," *Proc. 4th Int'l. Conf. on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia* (ACM, New York, NY, 2006), pp. 349–356.

[7] F. Banterle, P. Ledda, K. Debattista, A. Chalmers, and M. Bloj, "A framework for inverse tone mapping," Vis. Comput. **23**, 467–478 (2007).

[8] F. De Simone, G. Valenzise, P. Lauga, F. Dufaux, and F. Banterle, "Dynamic range expansion of video sequences: A subjective quality assessment study," *2014 IEEE Global Conf. on Signal and Information Processing (GlobalSIP)* (IEEE, Piscataway, NJ, 2014), pp. 1063–1067.

[9] B. Masia, A. Serrano, and D. Gutierrez, "Dynamic range expansion based on image statistics," Multimedia Tools Appl. **76**, 631–648 (2017).

[10] X. Zhang and D. H. Brainard, "Estimation of saturated pixel values in digital color imaging," J. Opt. Soc. Am. A **21**, 2301–2310 (2004).

[11] D. Xu, C. Doutre, and P. Nasiopoulos, "Correction of clipped pixels in color images," IEEE Trans. Vis. Comput. Graph. **17**, 333–344 (2011).

[12] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High dynamic range video," ACM Trans. Graph. (TOG) **22**, 319–325 (2003).

[13] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," ACM Trans. Graph. **36**, 1–12 (2017).

[14] N. K. Kalantari and R. Ramamoorthi, "Deep hdr video from sequences with alternating exposures," *Computer Graphics Forum* (Wiley, Hoboken, NJ, 2019), Vol. 38, pp. 193–205.

[15] G. Chen, C. Chen, S. Guo, Z. Liang, K.-Y. K. Wong, and L. Zhang, "HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 2502–2511.

[16] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," *European Conf. on Computer Vision ECCV 2018: Computer Vision – ECCV 2018* (Springer, Cham, 2018), pp. 120–135.

[17] Q. Yan, D. Gong, Q. Shi, A. v. d. Hengel, C. Shen, I. Reid, and Y. Zhang, "Attention-guided network for ghost-free high dynamic range imaging," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 1751–1760.

[18] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, and Y. Zhang, "Deep HDR imaging via a non-local network," IEEE Trans. Image Process. **29**, 4308–4322 (2020).

[19] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops* (IEEE, Piscataway, NJ, 2019), pp. 1954–1963.

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2017), pp. 105–114.

[21] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," IEEE Trans. Computational Imaging **3**, 47–57 (2016).

[22] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)* (IEEE, Piscataway, NJ, 2000), Vol. 1, pp. 472–479.

[23] pp. I–IS. K. Nayar, V. Branzoi, and T. E. Boult, "Programmable imaging using a digital micromirror array," *Proc. 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2004. CVPR 2004* (IEEE, Piscataway, NJ, 2004), Vol. 1.

[24] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," ACM Trans. Graph. (TOG) **30**, 1–10 (2011).

[25] J. Kronander, S. Gustavson, G. Bonnet, and J. Unger, "Unified HDR reconstruction from raw CFA data," *IEEE Int'l. Conf. on Computational Photography (ICCP)* (IEEE, Piscataway, NJ, 2013), pp. 1–9.

[26] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," Proc. SPIE **9023**, 279–288 (2014).

[27] T. Lulé, H. Keller, M. Wagner, and M. Böhm, "LARS II-a high dynamic range image sensor with a-si: H photo conversion layer," *1999 IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors, Nagano, Japan, Citeseer* (IEEE, Piscataway, NJ, 1999).

[28] U. Seger, U. Apel, and B. Höfflinger, "HDRC-imagers for natural visual perception," Handbook Comput. Vis. Appl. **1**, 2 (1999).

[29] S. Kavadias, B. Dierickx, D. Scheffer, A. Alaerts, D. Uwaerts, and J. Bogaerts, "A logarithmic response CMOS image sensor with on-chip calibration," IEEE J. Solid-State Circuits **35**, 1146–1152 (2000).

[30] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *Proc. 29th Annual Conf. on Computer Graphics and Interactive Techniques* (ACM, New York, NY, 2002), pp. 267–276.

[31] L. Meylan, S. Daly, and S. Süsstrunk, "The reproduction of specular highlights on high dynamic range displays," *Proc. IS&T/SID CIC14: Fourteenth Color Imaging Conf.* (IS&T, Springfield, VA, 2006), pp. 333–338.

[32] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. Whitehead, and G. Ward, "Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs," ACM Trans. Graph. (TOG) **26** (2007) 39-es.

[33] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, "Expanding low dynamic range videos for high dynamic range applications," *Proc. 24th Spring Conf. on Computer Graphics* (ACM, New York, NY, 2008), pp. 33–41.

[34] R. P. Kovaleski and M. M. Oliveira, "High-quality reverse tone mapping for a wide range of exposures," *2014 27th SIBGRAPI Conf. on Graphics, Patterns and Images* (IEEE, Piscataway, NJ, 2014), pp. 49–56.

[35] P. Didyk, R. Mantiuk, M. Hein, and H.-P. Seidel, "Enhancement of bright video features for HDR displays," *Computer Graphics Forum* (Wiley, Hoboken, NJ, 2008), Vol. 27, pp. 1265–1274.

[36] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," ACM Trans. Graph. (TOG) **36**, 1–15 (2017).

[37] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," ACM Trans. Graph. **36** (2017) 177–1.

[38] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image HDR reconstruction by learning to reverse the camera pipeline," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2020), pp. 1651–1660.

[39] S. Mangiat and J. Gibson, "High dynamic range video with ghost removal," Proc. SPIE **7798**, 779812 (2010).

[40] S. Mangiat and J. Gibson, "Spatially adaptive filtering for registration artifact removal in HDR video," *2011 18th IEEE Int'l. Conf. on Image Processing* (IEEE, Piscataway, NJ, 2011), pp. 1317–1320.

[41] N. K. Kalantari, E. Shechtman, C. Barnes, S. Darabi, D. B. Goldman, and P. Sen, "Patch-based high dynamic range video," ACM Trans. Graph. **32**, 1–8 (2013).

[42] Y. Gryaditskaya, T. Pouli, E. Reinhard, K. Myszkowski, and H.-P. Seidel, "Motion aware exposure bracketing for HDR video," *Computer Graphics Forum* (Wiley, Hoboken, NJ, 2015), Vol. 34, pp. 119–130.

[43] Y. Li, C. Lee, and V. Monga, "A maximum *a posteriori* estimation framework for robust high dynamic range video synthesis," IEEE Trans. Image Process. **26**, 1143–1157 (2016).

[44] G. Eilertsen, R. K. Mantiuk, and J. Unger, "Single-frame regularization for temporally stable cnns," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 11176–11185.

[45] S. Y. Kim, J. Oh, and M. Kim, "Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video," *Proc. AAAI Conf. on Artificial Intelligence* (AAAI, Washington, DC, 2020), Vol. 34, pp. 11287–11295.

[46] H. Zhang, L. Song, W. Gan, and R. Xie, "Multi-scale-based joint super-resolution and inverse tone-mapping with data synthesis for UHD HDR video," Displays **79**, 102492 (2023).

[47] X. Chen, Z. Zhang, J. S. Ren, L. Tian, Y. Qiao, and C. Dong, "A new journey from SDRTV to HDRTV," *Proc. IEEE/CVF Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2021), pp. 4500–4509.

[48] M. Anand, N. Harilal, C. Kumar, and S. Raman, "HDRVideo-GAN: deep generative HDR video reconstruction," *Proc. Twelfth Indian Conf. on Computer Vision, Graphics and Image Processing* (ACM, New York, NY, 2021), pp. 1–9.

[49] Y. Yang, J. Han, J. Liang, I. Sato, and B. Shi, "Learning event guided high dynamic range video reconstruction," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2023), pp. 13924–13934.

[50] Q. Yang, Y. Liu, and J. Yang, Efficient HDR reconstruction from real-world raw images, *arXiv preprint* arXiv:2306.10311 (2023), 10 pages.

[51] U. Cogalan, M. Bemana, K. Myszkowski, H.-P. Seidel, and T. Ritschel, "Learning HDR video reconstruction for dual-exposure sensors with temporally-alternating exposures," Comput. Graph. **105**, 57–72 (2022).

[52] Z. Liu, Z. Li, W. Chen, X. Wu, and Z. Liu, "Unsupervised optical flow estimation for differently exposed images in ldr domain," IEEE Trans. Circuits Syst. Video Technol. 1–1 (2023).

[53] O. Martorell and A. Buades, "Variational temporal optical flow for multi-exposure video," *VISIGRAPP (4: VISAPP)* (SciTePress, Setubal, 2022), pp. 666–673.
[54] Y. Jiang, I. Choi, J. Jiang, and J. Gu, HDR video reconstruction with tri-exposure quad-bayer sensors, *arXiv preprint* arXiv:2103.10982 (2021), 10 pages.
[55] U. Cogalan, M. Bemana, K. Myszkowski, H.-P. Seidel, and T. Ritschel, "Learning HDR video reconstruction for dual-exposure sensors with temporally-alternating exposures," Comput. Graph. **105**, 57–72 (2022) https://www.sciencedirect.com/science/article/pii/S0097849322000607.
[56] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," *Proc. IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2017), pp. 764–773.
[57] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *Proc. European Conf. on Computer Vision (ECCV)* (Springer, Cham, 2018), pp. 3–19.
[58] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2019), pp. 510–519.
[59] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint* arXiv:1704.04861 (2017), 9 pages.
[60] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2020), pp. 3360–3369.
[61] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2018), pp. 2472–2481.
[62] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," *SIGGRAPH '97: Proc. 24th Annual Conf. on Computer Graphics and Interactive Techniques* (ACM, New York, NY, 1997), pp. 369–378.
[63] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proc. Thirteenth Int'l. Conf. on Artificial Intelligence and Statistics, JMLR Workshop and Conf. Proc.* (JMLR, Cambridge, MA, 2010), pp. 249–256.
[64] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," ACM Trans. Graphics **27**, 1–10 (2008).
[65] M. Själander, M. Jahre, G. Tufte, and N. Reissmann, EPIC: an energy-efficient, high-performance GPGPU computing research infrastructure, arXiv:1912.05848 (2019), 6 pages.
[66] M. Narwaria, M. P. Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," Signal Process., Image Commun. **35**, 46–60 (2015).
[67] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," Color Res. Appl. **26**, 340–350 (2001).
[68] J. Kronander, S. Gustavson, G. Bonnet, A. Ynnerman, and J. Unger, "A unified framework for multi-sensor HDR video reconstruction," Signal Process., Image Commun. **29**, 203–215 (2014).
[69] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," Int. J. Comput. Vis. **127**, 1106–1125 (2019).