A Fast, Scalable, and Reliable Deghosting Method for Extreme Exposure Fusion

Video Analytics Lab, Department of Computational and Data Sciences,
Indian Institute of Science, Bengaluru, Karnataka-560012, INDIA

HDR fusion of extreme exposure images with complex camera and object motion is a challenging task. Existing patch-based optimization techniques generate noisy and/or blurry results with undesirable artifacts for difficult scenarios. Additionally, they are computationally intensive and have high execution times. Recently proposed CNN-based methods offer fast alternatives, but still fail to generate artifact-free results for extreme exposure images. Furthermore, they do not scale to an arbitrary number of input images. To address these issues, we propose a simple, yet effective CNN-based multi-exposure image fusion method that produces artifact-free HDR images. Our method is fast, and scales to an arbitrary number of input images. Additionally, we prepare a large dataset of 582 varying exposure images with corresponding deghosted HDR images to train our model. We test the efficacy of our algorithm on publicly available datasets, and achieve significant improvements over existing state-of-the-art methods. Through experimental results, we demonstrate that our method produces artifact-free results, and offers a speed-up of around 54x over existing state-of-the-art HDR fusion methods.

Index Terms—Exposure Fusion, High Dynamic Range Imaging, Deghosting.

I. INTRODUCTION

A DEQUATE lighting is important for commercial cameras to capture visually appealing photographs. However, it is unlikely that such a condition is encountered in common scenes. Often, the dynamic range of the scene far exceeds the hardware limit of standard digital camera sensors. In such scenarios, the resulting photos will consist of saturated regions, which are either too dark or too bright to visually comprehend. Keeping in mind the constraints posed by camera hardware, an algorithmic solution is to merge multiple Low Dynamic Range (LDR) images with varying exposure, also known as an exposure stack, into a single High Dynamic Range (HDR) image. This helps bring out details in both dark and bright saturated regions.

Multi-Exposure Fusion (MEF) methods merge well-exposed regions from numerous LDR input images to produce a single visually appealing LDR result that appears to possess higher dynamic range. It is a simple problem in case of perfectly static sequences (\(1\)–\(3\)). However, in most practical situations, a certain amount of camera and object motions are inevitable, leading to ghost-like artifacts in the final fused result.

Various techniques have been proposed in literature to address the ghosting problem in complex HDR fusion\(^1\). While camera motion can be addressed using global alignment methods (\(9\), \(10\)), deghosting of moving objects is much harder to tackle. Rejection based deghosting methods (\(11\)–\(13\)) fail to reconstruct HDR image content for moving objects with complex motion. Using non-rigid registration techniques (\(14\)–\(16\)) introduce visible artifacts in case of complex deformable motion. A paradigm shift in deghosting methods was offered by patch based synthesis methods such as Hu et al. (\(17\)) and Sen et al. (\(18\)). They offer impressive high quality results even for moving and occluded pixels. However, these methods are computationally heavy and still produce noisy or blurry results with artifacts for scenes with very complex and extreme dynamic range. Gallo et al. (\(19\)) offer a locally non-rigid registration method that strikes balance between execution time and accuracy.

Learning-based methods: Kalantari and Ramamoorthi (\(1\)) proposed the first deep learning based HDR deghosting algorithm. In their method, the input images are aligned using a traditional optical flow technique. The distortions introduced due to optical flow warping are corrected using a simple CNN. However, their method still generates artifacts for extreme dynamic range scenes (see Fig. 1). Recently, Wu et al. (\(2\)) proposed treating HDR fusion as an image translation problem, and solving it using a CNN-based architecture. While they do not explicitly perform foreground alignment (like (\(1\), \(17\), \(18\)), they showed that their network generates accurate HDR images irrespective of large foreground motion. However, this method is highly dependent on the structure of the reference image. Thus, if certain regions are saturated in the reference image, it fails to accurately reconstruct them in the final result (see Figure 5). A major bottleneck with learning-based approaches is scalability to an arbitrary number of images. Extending both methods to an arbitrary number of images requires training a new model, which can be computationally expensive.

Limitations: The deghosting algorithms mentioned above face a number of major challenges: (1) Handling extreme exposure images, (2) Complex scene motion, (3) Scalability to an arbitrary number of images, and (4) Speed. In order to address these challenges, we propose a simple, yet effective CNN-based MEF algorithm. Our proposed network, diagrammatically represented by Fig. 2 consists of multiple components. We start by registering the input images using a powerful optical flow network and a refinement network (Section II-B). The registered images are then passed to a feature encoder module to extract exposure-invariant features.

\* Equal contribution.
\(^1\)For more detailed review on literature methods, please refer to [7], [8].
We achieve scalability by training a single fully convolutional CNN feature encoder to extract features, irrespective of exposure bias. The common feature extractor ensures that our model remains independent of the number of input images. Finally, the extracted features are merged using a decoder CNN to generate artifact-free fused images (Section II-C).

In summary, our contributions are as follows:

- We present a simple, fast and scalable CNN model for MEF deghosting that can handle even extreme exposure differences in the input exposure stack.
- We provide a large varying exposure image dataset of 582 sequences with moving objects and corresponding ground-truth fused HDR images.
- We conduct extensive quantitative and qualitative evaluation of our model against state-of-the-art deghosting algorithms with publicly available datasets.

II. PROPOSED METHOD

A. Overall approach

In this section, we describe our-CNN based MEF deghosting method. The block diagram of our proposed approach is shown in Fig. 2. Our model consists of three major modules: an image alignment module, a feature extraction module, and an image reconstruction module. Given an exposure stack with N varying exposure images, one of them is selected as the reference image. In our implementation, we choose the middle image in the stack as the reference image during training (under exposed image, in case of image pairs). During testing however, any image in the stack can be used as the reference image to generate results. We select the image with the least amount of saturated pixels as the reference. The remaining N-1 images in the stack are aligned to the reference image using a CNN-based optical flow network. We use PWC-Net [20], a lightweight pyramidal optical flow estimation network for this purpose. The aligned images may contain artifacts in regions with non-rigid motion and occluded pixels. To correct this, we refine the warped output using our proposed refinement network (see Fig. 3). From the N aligned and refined images, we extract N convolutional features from using a feature encoder. These features are merged using a feature aggregation operation to generate the features corresponding to the fused result. Finally, the fused image is reconstructed from the fused feature map using a decoder.

Since all the stages in our model are independent of the number of input images, we explain the specifics of our method in further sections using two images: the reference image (I1) and the source image (I2), with the aim of aligning the source image to the reference image and fusing them. Extending our model to any arbitrary N images, requires N-1 optical flow estimation operations, N feature extraction operations, and one image reconstruction operation. Unlike [1] and [2], our method does not require retraining for a different number of inputs. We now examine each of our modules in detail.

B. Image alignment

Optical flow network N1: The objective of the alignment phase is to warp the source image I2 to the reference image I1. The input images I1 and I2 are assumed to be in linear RAW format. If not, we linearize them with the help of the Camera Response Function [21]. We use PWC-Net [20] for computing optical flow because of its high accuracy and low execution time. However, any optical flow estimation method can be utilized in this step. Since optical flow computation requires brightness constancy for better performance, we raise the exposure of the darker image to match that of the brighter image using the formula:

\[
I_1' = \left( \frac{I_1^\gamma \times t_2}{t_1} \right)^\frac{1}{\gamma}
\]  

where, t1 and t2 are the exposure times of I1 and I2, and \( \gamma = 2.2 \). The flow (\( \mathcal{F} \)) obtained from PWC-net (\( N_1 \)) is used to warp \( I_2 \) to generate \( I_2^w \):

\[
[\mathcal{F}, I_2^w] = N_1(I_1', I_2)
\]  

Refinement network N2: \( I_2^w \) is prone to artifacts due to errors in optical flow estimation and occlusion. To correct these, we utilize a U-Net [22] style network to refine \( I_2^w \) and estimate \( I_2' \), which represents an approximation of \( I_1' \) at the exposure value of \( I_2 \). \( N_2 \) takes three inputs: \( I_2^w, I_1' \) (obtained from Eqn. 1), and flow \( \mathcal{F} \). We stack the inputs in...
**Fig. 2: Overview of our approach.** $N$ varying exposure input images are registered to reference image (shown in middle) using optical flow and refinement network. The registered images are passed to a feature extraction network. We aggregate the $N$ features to obtain fused image feature. Finally, a feature decoder model is used to reconstruct final pseudo-HDR image.

**Fig. 3: We illustrate the working of our method with a varying exposure image pair ($I_1, I_2$) as input.** First, we compute the flow between the exposure modified reference image, $I_1'$ and $I_2$. The warped image $I_w^2$ may contain artifacts that are corrected by the refinement network ($N_2$) with the help of the optical flow ($F$) and $I_1'$. Finally, the artifact-free warped image $\hat{I}_2$ is fused with $I_1$ by Fusion net ($N_3$) to generate the pseudo-HDR result. The overall approach consists of two stages of training. In the first stage, the fusion network ($N_3$) is trained separately with static input images. In the second stage, $N_3$ is frozen and the loss between the predicted fused image and the ground truth image ($\text{Loss}_2$) is used to train the refinement network along with $\text{Loss}_1$. Note that the weights of the optical flow network ($N_1$) are frozen throughout the training process.

The depth dimension to obtain eight channel input data. $N_2$ is then trained to predict a single channel weight map:

$$ W = N_2(I_w^2, I_1', F) $$

Finally, $\hat{I}_2$ is obtained using the formula:

$$ \hat{I}_2 = (1 - W) \times I_w^2 + W \times I_1' $$

(3)

Then, $N_2$ is trained with loss computed between predicted image $\hat{I}_2$ and ground truth $I_w^g$ with $L_2$ loss:

$$ \text{Loss}_1 = \|\hat{I}_2 - I_w^g\|_2 $$

(4)

The weight map primarily needs to emphasize details from $I_2$ that may be lost during the estimation of $I_1'$. Training the refinement network with just the above mentioned $L_2$ loss causes the network to get stuck in a local minimum, outputting uniform values of one in $W$, giving $I_1'$ as output. This occurs because for most pixels, $I_1'$ is very close to the ground truth. To counter this, we regularize the loss function using the $L_2$ norm of the weight map.

$$ \text{Loss}_1 = \|\hat{I}_2 - I_w^g\|_2 + \alpha_1 \times \|W\|_2 $$

(5)

The use of such a regularizer can give rise to ghosting artifacts. To minimize this, the weight of the regularizer is decreased after 15 epochs. A sample result from the refinement stage can be seen in Fig. 4. It should be noted that $\text{Loss}_1$ is used to train only $N_2$, while the weights of $N_1$ are frozen.

The refinement network is trained with the Adam optimizer for 45 epochs, starting with a learning rate of 1e-4 and $\alpha_1$ of 1e-3 for the first 15 epochs. After that, $\alpha_1$ is set to 2.5e-5, and is decreased by 0.25e-5 every 5 epochs till the 35th epoch, after which these parameters are kept unchanged for the final 10 epochs. The learning rate is decreased by 0.25e-4 every 5 epochs till the 25th epoch, and is then kept unchanged.
three unique methods of feature aggregation: 1) Mean - Mean activation function everywhere with \( \alpha \) spatial resolution of the feature maps. We use the LeakyReLU consists of four transpose-convolutional layers to increase the downsample the feature maps by a factor of 2. The decoder consists of 4 convolutional layers with 8, 16, 32 and 64 filters respectively, followed by a strided convolution that increase the resolution of input images, our method can be used to merge any number of input images. Within the fusion module (Fig. 2), we utilize an encoder \( E \) to extract individual image features, merge these features based on certain heuristics, and finally reconstruct the fused image from the merged features. To our knowledge, our method is the first CNN-based fusion approach that can fuse an arbitrary number of images.

The input to our fusion module is the set of \( N \) aligned images. Within the fusion module (Fig. 2), we utilize an encoder \( E \) to obtain exposure-invariant features from all \( N \) images, merge these \( N \) features using a feature aggregator operation, and finally use a decoder \( D \) to reconstruct the final fused image. Since the feature extractor \( E \) is unchanged for all input images, our method can be used to merge any number of images.

We follow a standard encoder-decoder architecture for image fusion, with the exception that encoder features from all the images are aggregated before being passed to decoder. The encoder consists of 4 convolutional layers with 8, 16, 32 and 64 filters respectively, followed by a strided convolution that downsample the feature maps by a factor of 2. The decoder consists of four transpose-convolutional layers to increase the spatial resolution of the feature maps. We use the LeakyReLU activation function everywhere with \( \alpha = 0.2 \). We investigate three unique methods of feature aggregation: 1) Mean - Mean of \( N \) feature maps, 2) Max - Max of \( N \) feature maps, and 3) Mean + Max - both mean and max of \( N \) features maps are concatenated. From the results reported in Table II, we observe that the Mean + Max fusion strategy scales well for multiple images. Thus, we choose Mean + Max as the preferred feature aggregation operator. The aggregated features from the corresponding encoding block are also fed to the decoder through skip connections. The fusion network is trained using an \( L_2 \) loss, an MS-SSIM loss [23], and a gradient magnitude loss between the ground truth (\( H^{gt} \)) and the predicted pseudo HDR image (\( \hat{H} \)):

\[
\text{Loss}_2 = \beta_1 \| \hat{H} - H^{gt} \|_2 + \beta_2 [1 - \text{MSSSIM}(\hat{H}, H^{gt})] + \beta_3 \| \nabla_m \hat{H} - \nabla_m H^{gt} \|_2
\]

where, \( \nabla_m \hat{H} \) denotes the gradient magnitude of image \( \hat{H} \). Empirically, we found that, setting \([\beta_1, \beta_2, \beta_3] \) with [0.3, 1.6] gave better results. We train our fusion network with 3 images of varying exposure but it is extensible to arbitrary number of images. The fusion network is trained with Adam optimizer over static images for 45 epochs. Learning rate is set to 1e-4 for first 15 epochs, after which, learning rate is halved after every 10 epochs.

III. EVALUATION AND RESULTS

A. Dataset

Considering the scarcity of large datasets for MEF in literature, we create a new dataset that could prove to be beneficial for the research community. We follow the capture process used by Kalantari and Ramamoorthi [1] to capture two sets of images for the same scene: reference and ghosted sets. In the reference set, the subject was asked to remain still during the capture process. In the ghosted set, the subject was asked to perform some action. Each set consists of three to seven images with exposure values between \(-3EV \) to \(+3EV \)(inclusive). All the images were captured in the camera’s native RAW format, and later converted to TIFF using dcraw software. The captured images have different resolutions, ranging from 1 to 4 Megapixels. We captured around 700 sequences with real-life object motion in various scenarios. We used three different capture devices to collect these sequences: Canon EOS 600D, Canon EOS 500D and an LG V30 smartphone. All the images were captured by mounting the device on a static tripod. In order to avoid camera motion due to manual shutter release, we used a wireless remote trigger. Out of the 700 captured sequences, we removed 118 sequences with background motion or object motion.

Finally, we split the 582 sequences into random, distinct sets:
TABLE I: Quantitative comparison with state of the art methods. Here, we compare proposed method with Hu et al. [17] and Ma et al. [5] with full reference IQA metrics PSNR and SSIM. We report the results for 2 and 3 images per sequence fusion.

<table>
<thead>
<tr>
<th>Number of images (→)</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>Methods (→)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu et al. [17]</td>
<td>27.04</td>
<td>0.925</td>
</tr>
<tr>
<td>Ma et al. [5]</td>
<td>25.67</td>
<td>0.913</td>
</tr>
<tr>
<td>Proposed (with flow)</td>
<td>29.61</td>
<td>0.952</td>
</tr>
<tr>
<td>Proposed (without flow)</td>
<td>28.12</td>
<td>0.938</td>
</tr>
</tbody>
</table>

The ground truth pseudo-HDR images were generated with the static reference set using Mertens’ et al. [4] algorithm. Additionally, we also test our model on Sen [18], Karaduzovic [3] and Tursun [8] datasets.

B. Quantitative evaluation

1) Comparison with SoTA methods

We perform quantitative comparisons of our method with two other, well-known state of the art methods: 1) Hu et al. [17] 2) Ma et al. [5]. We compare against these methods using two full-reference image quality assessment metrics: PSNR and SSIM [23] in Table I. The results for these methods are generated with the default parameters specified by the authors. We compare the results in two categories: image pair fusion and triple-image fusion. From Table I, we observe that Hu et al. [17] performs better in terms of PSNR than our proposed method for triple-image fusion. However, our method achieves a higher SSIM score in the same category. For extreme image pair fusion, our method outperforms both Hu et al. [17] and Ma et al. [5] in terms of PSNR and SSIM. We did not compare with Sen et al. [18], Kalantari and Ramamoorthi [1], and Wu et al. [2], as they generate HDR images directly. Since the ground truth fusion strategy for HDR imaging is different than that of MEF (triangular weighting for HDR and Mertens et al. [4] for MEF), it might be inaccurate to compare MEF and HDR methods quantitatively.

In Table III, we benchmark our method for HDR image deghosting against Sen et al. [18], Kalantari and Ramamoorthi [1], and Wu et al. [2] on the test set of UCSD dataset. The PSNR score for our model is slightly lower as a consequence of performing mean and max. Thus, in order to prove the effectiveness of our refinement net (N₂) model, we modified our fusion net (N₃) model to reflect the same setting as [2]. That is, weights for the fusion model (N₃) is not shared among input images. We train this model for a fixed number of inputs (three images in UCSD dataset) by using a separate encoder for each of the inputs and concatenating the feature maps instead of aggregating them using max/mean and obtained better PSNR values than [2] with a much shallower network. This modified model is referred to as untied weights in Table III, while the original model is referred to as tied weights.

TABLE II: Fusion scalability analysis: In this table, we report the no-reference IQA metric MEF SSIM scores for three fusion strategies, trained to fuse three input images only and tested to fuse any arbitrary number of input images without retraining for them. See IV-2 for more details.

<table>
<thead>
<tr>
<th>Methods/Metrics</th>
<th>PSNR(db)</th>
<th>HDR-VDP-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods/Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu et al. [17]</td>
<td>33.61</td>
<td>60.12</td>
</tr>
<tr>
<td>Sen et al. [18]</td>
<td>40.43</td>
<td>62.11</td>
</tr>
<tr>
<td>Kalantari and Ramamoorthi [1]</td>
<td>42.70</td>
<td>66.64</td>
</tr>
<tr>
<td>Wu et al. [2]</td>
<td>41.65</td>
<td>67.96</td>
</tr>
<tr>
<td>Proposed method (Tied weights)</td>
<td>40.47</td>
<td>66.80</td>
</tr>
<tr>
<td>Proposed method (Untied weights)</td>
<td>41.80</td>
<td>66.52</td>
</tr>
</tbody>
</table>

TABLE III: Quantitative comparison with state-of-art HDR deghosting algorithms on UCSD dataset.

<table>
<thead>
<tr>
<th>Methods/Metrics</th>
<th>PSNR(db)</th>
<th>HDR-VDP-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods/Metrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu et al. [17]</td>
<td>0.9768</td>
<td>0.9809</td>
</tr>
<tr>
<td>Sen et al. [18]</td>
<td>0.9787</td>
<td>0.9810</td>
</tr>
<tr>
<td>Kalantari and Ramamoorthi [1]</td>
<td>0.9787</td>
<td>0.9810</td>
</tr>
<tr>
<td>Wu et al. [2]</td>
<td>0.9774</td>
<td>0.9823</td>
</tr>
<tr>
<td>Mertens et al. [4]</td>
<td>0.9826</td>
<td>0.9844</td>
</tr>
</tbody>
</table>

C. Qualitative evaluation

We perform qualitative comparisons with Kalantari and Ramamoorthi [1] and Wu et al. [2] in Fig. 1 and Fig. 5. As seen in the highlighted area, the result generated by Kalantari and Ramamoorthi [1] still has artifacts due to optical flow misalignment. Additionally, their result tends to consist of small speckles of artifacts with erroneous colors as shown in Fig. 5. One reason for such artifacts could be the fact that Kalantari and Ramamoorthi [1] use a single CNN network to perform both deghosting and exposure fusion. This may cause the network to be confused between moving objects and highly saturated regions. In contrast, we perform both operations separately with different CNNs. Thus, our network is able to distinguish between moving objects and saturated pixels. In Fig. 1 and 5, we can observe ghosting artifact in results generated by Wu et al. [2] due to object motion.

In Fig. 6, we compare our results with Hu et al. [17], Sen et al. [18] and Ma et al. [5] for extreme image pair fusion with images from Tursun’s dataset [8]. The results are generated with the under-exposed image as the reference, and are displayed after tonemapping using Adobe Photoshop. It can be observed that Hu et al. [17], Sen et al. [18] and Ma et al. [5] hallucinate details in regions of heavy saturation. The artifacts are generated due to incorrect pixel correspondence during the alignment phase. Comparatively, though the aligned images consist of artifacts, our refinement network successfully identifies them as incorrect matches, thus avoiding such artifacts in the final result.

D. Running time

In Table IV, we compare the execution times of five state of the art methods for triple-image fusion with 1260×800 resolution. Compared to image synthesis approaches such as Hu et al. [17] and Sen et al. [18], our method achieves a speed up of up 54×. It should be noted that for Kalantari and Ramamoorthi [1], their non-deep optical flow alignment
method takes about 58 seconds to align three images. We plot PSNR against execution time in Fig. 9 for triple-image fusion.

IV. DISCUSSION

1) With and Without Optical Flow:

As shown in Table I, we performed experiments to deduce the efficacy of appending optical flow as input to the refinement network. The results indicate that the network is able to perform better when flow is present as auxiliary input. We hypothesize that the flow information helps the network identify pixels with occlusion and non-smooth flow regions.

2) Scalability Analysis:

In Table II, we report the no-reference image quality metric MEF-SSIM performance of three feature-agglomeration techniques, trained to fuse three input images only and tested to fuse any arbitrary number of input images without retraining. We trained three models using the Mean, Max and Mean + Max feature merging techniques with three input images (see section II C for details). We tested these models for sequences with 2, 5, and 7 input images. The number of streams in the feature extractor are replicated N times for N input images. For the purpose of comparison, we have included the scores for the ground truth images as well. Mean + Max fusion strategy scales best for a higher number of input images. It should be
TABLE IV: Execution time comparison between different methods for an image resolution of 1260 × 800.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (in seconds)</td>
<td>330.65</td>
<td>318.57</td>
<td>12.37</td>
<td>61.53</td>
<td>0.29</td>
<td>0.32</td>
<td>6.66</td>
</tr>
<tr>
<td>Speed up factor</td>
<td>56 ×</td>
<td>54 ×</td>
<td>2 ×</td>
<td>192 ×</td>
<td>0.9 ×</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 7: Comparison with Ma et al. [5] for an input sequence from Sen’s dataset [18]. Ma et al. [5] introduces structural artifacts in the facial region. Best viewed in color when zoomed in on an electronic display.

Fig. 8: Choosing a reference image with saturated pixels leads to artifacts, as highlighted in the red box of (c). To overcome this problem, one can choose the image with the least number of saturated pixels as the reference image, as shown in (d).

Fig. 9: PSNR vs Execution Time scatter plot for triple-image fusion.

noted that since the network was trained with Mertens et al. [4] fused images, it can not outperform it.

3) Choice of Reference Image:

An important parameter for the alignment stage is the choice of the reference image. Choosing an over-exposed reference image can lead to incorrect reconstruction if the overexposed pixels in source image are occluded, since the network has no way of guessing the correct value for clipped pixels. On the other hand, choosing the most underexposed image can lead to noise when it is mapped to the higher exposure image. Thus, it is desirable that the reference image is not saturated. We find that choosing image with least number of saturated pixels in the HSV color space provides a balanced middle ground, and more complex techniques such as color quantization and histogram analysis can potentially produce better results.

V. CONCLUSION

We have proposed a CNN-based fast, scalable image deghosting method that generates visually pleasing results even for extreme exposure inputs. In our method, we compensate for background and foreground motion using optical flow. The artifacts introduced by optical flow are corrected using a refinement network. Finally, the aligned images are fused using a novel scalable exposure fusion CNN module. The proposed method can fuse an arbitrary number of images without the need for re-training. The proposed method shows state-of-the-art performance with low computation time, achieving a speed up of around 54 × compared to existing methods.

The current model is trained with ground truth generated by Mertens et al. [4]. An interesting avenue of research would be to train the model with an NR-IQA metric such as MEF-SSIM. Another potentially useful idea would be to extend the current model to generate HDR videos. To do so, one has to take temporal structural and color coherency into account. Models such as 3D-CNN can be helpful in such a scenario.

ACKNOWLEDGMENT

The authors would like to thank Nima Kalantari for his helpful suggestions and clarifications.

REFERENCES


