IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 February 2024, accepted 27 March 2024, date of publication 8 April 2024, date of current version 15 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3386096

RESEARCH ARTICLE

High Dynamic Range Imaging via Visual Attention Modules

ALI REZA OMRANI[1,2] **AND DAVIDE MORONI**[1]

[1]Institute of Information Science and Technologies (ISTI), National Research Council of Italy, 56124 Pisa, Italy
[2]Department of Engineering, Università Campus Bio-Medico di Roma, 00128 Rome, Italy

Corresponding author: Ali Reza Omrani (ali.omrani@isti.cnr.it)

ABSTRACT Thanks to High Dynamic Range (HDR) imaging methods, the scope of photography has seen profound changes recently. To be more specific, such methods try to reconstruct the lost luminosity of the real world caused by the limitation of regular cameras from the Low Dynamic Range (LDR) images. Additionally, although the State-Of-The-Art (SOTA) methods in this topic perform well, they mainly concentrate on combining different exposures and pay less attention to extracting the informative parts of the images. Thus, this paper aims to introduce a new model capable of incorporating information from the most visible areas of each image extracted by a Visual Attention Module (VAM) which is a result of a segmentation strategy. In particular, the model, based on a deep learning architecture, utilizes the extracted areas to produce the final HDR image. The results demonstrate that our method outperformed most of the SOTA algorithms.

INDEX TERMS Deep neural network, high dynamic range imaging, image segmentation, multi-exposure image, visual attention module.

I. INTRODUCTION

In the scope of photography, the real world consists of an unlimited range of luminance. However, most devices are capable of capturing only a limited amount of that light. Therefore, the taken images are not desirable and consist of saturated regions, in which some parts of the images are too dark (underexposed) or overly bright (overexposed). These types of pictures are called LDR images.

Thus, in order to cope with this problem, highly advanced cameras [1], [2], [3], [4], [5], [6], [7] can be used, which have special sensors to capture more light. However, such devices are mainly too expensive and overly heavy, which are not suitable for daily life, and instead, are primarily used in industries.

A possible resolution for this drawback is developing software algorithms called HDR imaging techniques. Moreover, HDR images can be implemented by a single image [8], [9], [10], [11] or fusing a stack of images with different exposures, which are called single- and multi-exposure methods, respectively. In algorithms with a single LDR image, an HDR image can be produced from one image. However, the generated picture might not be as informative as the HDR image produced by several LDR images because the amount of detail in one single picture is limited compared to several images with different exposures. More precisely, [8] implemented an algorithm that only reconstructs the detail of bright saturated areas. However, the model is not only incapable of restoring the detail of dark regions but also does not perform well if the amount of bright saturation is too much. Thus, [12] first combined several LDR images and then fed the low-frequency response of the wavelet transform to the network to produce more detail in a shorter time.

Luckily, multi-exposure methods are more effective and informative compared to single-exposure techniques. Moreover, these methods perform well when the images are static [13], [14], while when there are movements in the sequence of pictures, the ghosting problem emerges, which is almost solved in [15], [16], [17], [18], [19], and [20].

Deep learning has been a significant means of producing an HDR image for the past decade. For instance, [8] produced an HDR picture in the logarithmic domain with the help of a deep neural network. Additionally, [21], used a neural network to reconstruct the detail of an image with different exposure in each row in the irradiance domain. Moreover, unlike other multi-exposure methods, [13], [14] used a neural

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

network to produce synthetic LDR images with different exposures from a single image. Furthermore, [16] proposed to first align images with the help of the optical flow method, and then use a deep neural network to combine them. Therewith, [15], instead of using optical flow for alignment, proposed to use two different neural networks first to align them and then combine the aligned images with the second neural network. Finally, [22] used a neural network to learn the relative relation between the inputs and the Ground Truth using input images in different scales.

In this article, we would like to exploit image segmentation with the help of the Otsu method [23] in HDR imaging to extract the most visible areas of the images and help the model produce pictures with more detail. By doing so, we are able to reduce the complexity of the neural network model and obtain similar or better results. Indeed, deep learning methods have demonstrated outstanding capabilities in identifying the most relevant features in the images, and for the present task, they are, in principle, able to identify the most informative areas in the images with different exposures. However, this may require a higher network complexity and an intractable number of parameters. Conversely, this paper investigates the possible role of segmentation in driving the network architecture to a superior HDR reconstruction. To reach this point, VAMs will be proposed to obtain such regions. Moreover, in this research, Spatial and Attention modules have been used from SOTA methods, and a new architecture for the Reconstruction stage was designed and implemented, in which the visual attention and the reference image were used in the decoder part. Finally, although VAMs helped in producing pictures with more details and outperformed most of the SOTA methods, the results still illustrated a slight amount of noise that was extracted from the input images.

In section II, the SOTA in HDR imaging and related image segmentation are presented. In section III, the proposed method is discussed in detail. Section IV demonstrates the experimental results and comparison with the SOTA methods. Moreover, section V presents ablation studies to prove the value of the single proposed steps, concentrating on the use of VAM and of the Refinement stage. Finally, section VI concludes this article with ideas for further work. The code will be available at the **github page**.

II. RELATED WORK

In this section, we will discuss the SOTA methods in the scope of HDR imaging in the Multi-Exposure category (Section II-A) and survey unsupervised Image Segmentation methods for extracting regions (Section II-B).

A. MULTI-EXPOSURE METHODS

Reference [24] proposed a two-stage algorithm, in which, in the first phase they extracted features from the input images and merged them to produce the HDR image in the latter one. Additionally, to cope with the appeared noise from the gamma correction operation on input images, i.e. the gamma-corrected Short-Exposure image becoming similar to Medium-Exposure, they used a U-net to extract noiseless features from it. Moreover, [25] implemented a model in which images with lower scales were used to reduce the consuming sources. Additionally, a novel loss function was defined to focus more on the motion. Furthermore, [26] forwarded features with different scales to deformable and spatial attention blocks to align images in the feature space and also extract the features of the specific areas of the input images. Moreover, [27] proposed a model that first estimated the optical flow from the two input images in different scales and then fused them to produce the final output. In [28], the features are extracted from different scales and then processed by sampling and aggregation modules to align the pixels of the non-reference features.

The work [29] implemented a baseline that had lower computational resources and acceptable results compared to the other SOTA models. They used a dual attention module, which includes both spatial and channel attention modules, to cope with misalignment and to better learn the details of the produced areas. In [30], the authors proposed a model that first extracts features from input images by multi-scale encoding modules and then produces an HDR image by progressively dilated U-shape blocks.

Reference [31] demonstrated that the ghosting problem is mainly in short-frequency signals. Therefore, they proposed a wavelet-based model to merge images in the frequency domain and avoid any ghosting problems. Reference [32] implemented an algorithm that extracted dynamic areas of the images with the help of image segmentation and applied two neural networks separately on the static and dynamic scenes. Finally, they merged the information to produce an HDR image without ghosting. In [33], a model based on bidirectional motion estimation was proposed, in which the amount of optical flow between LDR images was estimated by motion estimation with cyclic cost volume and spatial attention maps, and eventually, an HDR image was produced with the help of the extracted local and global features. Reference [34] implemented the first multi-bracket HDR pipeline using event cameras, in which they merged the extracted features of images and the events to produce an HDR image. Reference [35] proposed a transformer-based baseline, in which they used a context-aware vision transformer to extract local and global features to model the movement of objects and the diversity of intensity.

B. IMAGE SEGMENTATION

Image segmentation is a crucial task in computer vision, which tries to partition images into segments to analyze the pictures more easily. Additionally, image segmentation not only can be used for object recognition, detection, and medical purposes but also can be applied for extracting regions of pictures with more details. In [36] images were analyzed in HSV color space to segment pixels based on Intensity or Hue value. Moreover, two image segmentation methods were proposed based on luminance:

histogram division [37] and clustering based on Gaussian Mixture Model (GMM) of histogram [38]. Furthermore, [39] calculated an optimal valley point based on the slope between the histogram value of each pixel and the neighboring points, and used the computed valley point to segment regions. The literature on the topic is endless, depending on applications and methodologies, from level set methods [40] to graph cut [41] to recent deep learning-based frameworks [42].

III. PROPOSED METHOD

A. OVERVIEW

As cited in [43], it might be beneficial to first segment images based on exposure information to extract the best and more detailed regions from the Over- and Under-Exposure regions and exploit this knowledge in reconstructing an HDR image. Following this idea, in this paper, a model is proposed in which, with the help of image segmentation, regions with more detail are segmented first in the preprocessing stage. Finally, they are fed to the model along with the input images to produce an HDR image with the help of VAMs.

Generally, the model can be divided into several sections. Firstly, the input images are fed into the feature extraction module, and afterward, the extracted features enter the attention and spatial alignment modules to cope with any possible misalignment. Moreover, the input images with their corresponding masks go to the VAM simultaneously to extract the visible areas of the LDR images. Next, the outputs of the three modules are fed to the Reconstruction stage to produce the initial HDR image. Finally, the generated outcomes with the features of the reference image enter the refinement section to construct the final HDR image.

B. PREPROCESS

In this article, the inputs are three LDR images with different exposures, and the image with Medium-Exposure is considered the reference image. Moreover, before feeding the input images to the model, they are first mapped to the HDR domain with the help of gamma correction. Finally, they are concatenated channel-wise with their corresponding LDR images.

$$\hat{I}_i = \frac{(I_i)^{\gamma}}{t_i} \qquad \text{for } i = 1, 2, 3 \tag{1}$$

where t_i is the exposure time of I_i. γ is the gamma correction parameter, which was 2.24, and $\hat{I}_i$ is the gamma-corrected image.

1) SEGMENTATION

Most of the present algorithms in HDR imaging focus more on the approach of image production, but not many pay attention to how to extract the most helpful features. Thus, in this research, the regions of the pictures with more details are segmented and extracted as a preprocess and finally are fed to the proposed model along with the LDR images as the inputs.

Different methods, such as neural network-based approaches and the Otsu method, were used for the image segmentation stage; however, in our experiments, the neural network-based approaches resulted in overfitting. Thus, the widely adopted Otsu method has been selected, also given its simplicity, to segment the visible areas of the pictures. To this end, the images are first converted into the YUV color space and then the luminance channel Y is taken into consideration by computing a threshold based on the histograms of Short- and Long-Exposure images. More exactly, in each sample, different thresholds are calculated based on the histogram of each image in each exposure. Thus, the threshold parameter for each image is a variable threshold based on each sample.

$$\text{thresh}_i = G(Y_i) \qquad \text{for } i = 1, 3 \tag{2}$$

In which Y_i is the luminance channel of the LDR image, $G()$ is the Otsu function, and thresh_i is the threshold value of image i.

In the Short-Exposure image, because most of the pixels are dark, and the objective is to extract the regions with visible pixels, the values equal to or more than the threshold are considered one, and the rest are zero for the Short-Exposure mask.

$$\begin{cases} 1 & p \geq \text{thresh}_1 \\ 0 & p < \text{thresh}_1 \end{cases} \tag{3}$$

where thresh_1 is the threshold value of the Short-Exposure image, and p is the pixel.

On the other hand, because most of the pixels in the Long-Exposure image are saturated, and the visible pixels have the lowest values, the values that are less than the threshold were considered one, and the rest as zero in the Long-Exposure mask.

$$\begin{cases} 0 & p \geq \text{thresh}_3 \\ 1 & p < \text{thresh}_3 \end{cases} \tag{4}$$

By doing so, the masks of the areas with more detail are extracted and can help to produce an HDR image.

Generally, most of the pixels in Short- and Long-Exposure images are too dark or bright, respectively. Therefore, the location of the areas with surplus information is extracted and fed to the model. Doing so reduces the amount of calculation and helps in producing an HDR image with more detail. Fig. 1 demonstrates the segmented and visible regions of both Short- and Long-Exposure pictures (first and second masks from the left, respectively), and the third mask is a sum of both generated masks.

Moreover, during experiments, three input images with different exposures were used for image segmentation, in which, after obtaining the suitable areas of Short- and Long-Exposure images, the remaining regions were extracted from the Medium-Exposure image. However, the acquired areas of the Medium-Exposure were not sensible, as most of them were only a few pixels. Thus, two reasons exist for not using Medium-Exposure in the segmentation stage. First,

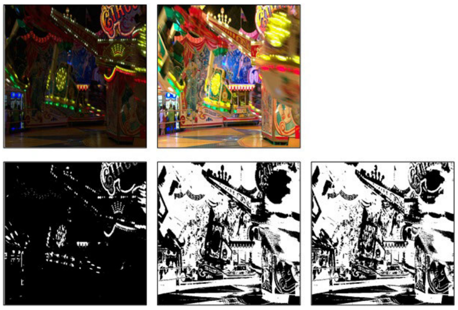

FIGURE 1. Produced masks of Short- and Long-Exposure images.

it would be challenging to calculate a range for the visibility of the pixels. Second, Medium-Exposure is the reference image, and the picture will be used in the neural network. Therefore, it is not necessary to use segmentation for it.

C. PROPOSED METHOD STRUCTURE

As shown in Fig. 2, the proposed algorithm consists of six stages, which will be discussed separately and in detail.

1) FEATURE EXTRACTION

Fig. 3 illustrates the Feature Extraction block, in which a SepConv is applied to the image to extract 32 feature maps. Afterward, a Max Pool and an Average Pool are used to not only smooth the features and focus on the details but also pay more attention to the edges. Next, the outputs of Poolings are concatenated, and another SepConv + ReLU is used to reduce the number of channels to 32. Finally, the extracted features are Upsampled to make them the same size as the input image. The feature extraction can be written as follows:

$$\text{features}_i = \text{SepConv}(I_i) \tag{5}$$

$$C_i = \text{concat}\,(M(\mathit{features}_i), A(\mathit{features}_i)) \tag{6}$$

$$F_i = \text{Upsample}(\text{ReLU}(\text{SepConv}(C_i))) \tag{7}$$

for $i = 1, 2, 3$, where $M()$ and $A()$ functions are Max Pooling and Average Pooling, respectively, and C_i is the output of Concatenation. Finally, F_i is the output of the Feature Extraction Block.

2) VISUAL ATTENTION MODULE

As it was mentioned, in this article, Image Segmentation is used to help the model to produce a better image. Therefore, as shown in Fig. 4, the input images are multiplied element-wise by their corresponding masks first. By doing so, the regions with more details are kept, and those that are overly dark or too bright will be removed. Next, they are fed to the Feature Extractor to extract Features. Finally, they are added together element-wise. The VAM can be formally defined as follows:

$$\text{features}_L = F(\text{multiply}(\text{mask}_L, I_L)) \tag{8}$$

$$\text{features}_H = F(\text{multiply}(\text{mask}_H, I_H)) \tag{9}$$

$$V = \text{add}(\text{features}_L, \text{features}_H) \tag{10}$$

where F is a feature extractor function, and V is the output feature of the VAM.

3) SPATIAL ALIGNMENT MODULE

Because the input LDR images are not aligned, the extracted features from the LDR images without the gamma correction images are fed to an *ad hoc* module for aligning them. To this end, we used the same Feaure-alignment Module used in [30]. As can be seen in Fig. 5, first a Conv + ReLU is applied to the Reference Features, which can be called as Ref_1. Next, a Conv + ReLU is applied to Ref_1 and is multiplied element-wise by the input LDR features, which can be called M_i (for $i = 1, 3$). Finally, another Conv + ReLU is applied to the Ref_1 and is added element-wise with M_i. Formally, the operation in the module can be written as follows:

$$\text{Ref}_1 = \text{ReLU}(\text{Conv}(\text{ref features})) \tag{11}$$

$$M_i = \text{multiply}(\text{ReLU}(\text{Conv}(\text{Ref}_1)), \text{ inp features}_i) \tag{12}$$

$$\text{out}_i = \text{add}(\text{ReLU}(\text{Conv}(\text{Ref}_1)), M_i) \tag{13}$$

4) ATTENTION MODULE

The Attention Module is similar to [30] in terms of the structure, but it differs in details, in which, as shown in Fig 6, feature maps are produced for Short- and Long-Exposure images to merge them with the reference image as guidance. After feeding the features of gamma-corrected images with the reference image, they are concatenated. Afterward, SepConv + ReLU and SepConv + Simgoid operations are applied to them. The module can be considered as follows:

$$R_i = \text{ReLU}(\text{SepConv}(\text{concat}(f_i, f_r)) \quad \text{for } i = 1, 3 \tag{14}$$

$$S_i = \text{Sigmoid}(\text{SepConv}(R_i)) \tag{15}$$

where f_i and f_r are the features of gamma-corrected and reference images, respectively.

5) RECONSTRUCTION

All the extracted features from the modules are concatenated and fed to the reconstruction stage. As shown in Fig. 7, with the help of four encoder blocks, the input is merged, and new features are produced. Next, each decoder block receives features from the encoder along with features of the reference image and VAM. Finally, a SepConv + ReLU is used to produce the output of the stage.

Each encoder block (Fig. 8, left) initially applies SepConv, Batch Normalization, and ReLU layers to the inputs. Afterward, similar to Feature Extraction Module, Max and AVG Poolings are used. Finally, they are concatenated and sent to the next block.

Moreover, each decoder block (Fig. 8, right) consists of three inputs, which are features of the VAM, features of the reference image, and the output of the previous block. First, AVG pooling is applied to the first two inputs to make them the same size as the output of the previous block, and then they are concatenated with each other. Finally, SepConv + ReLU and Upsampling are used, respectively.

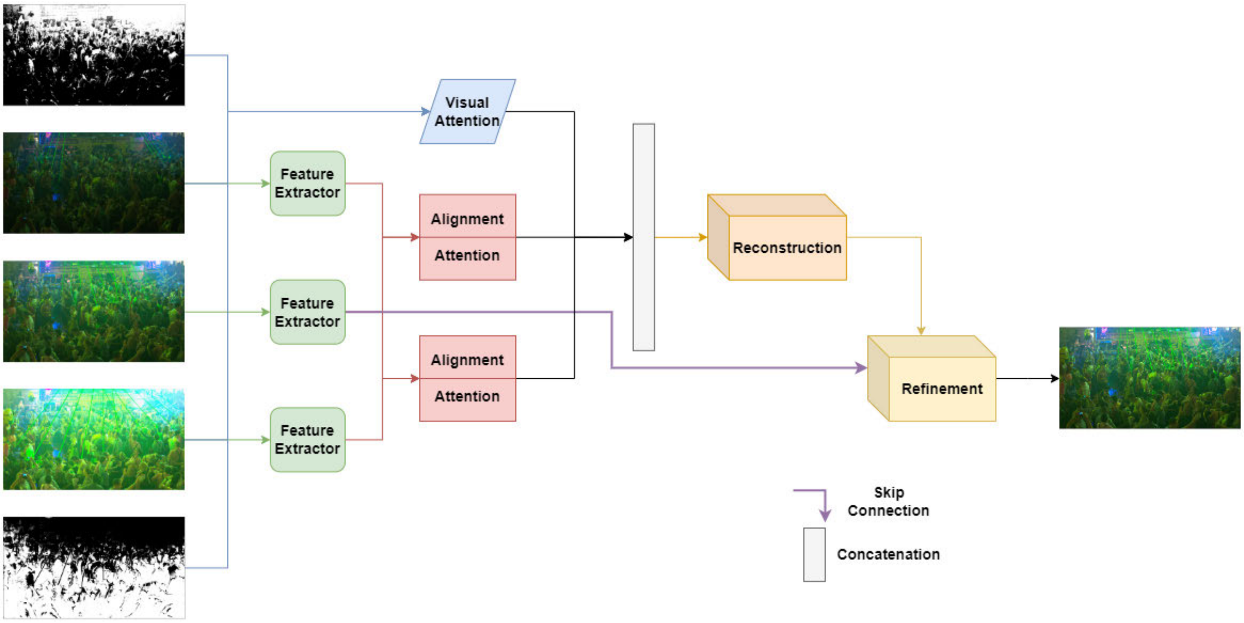

FIGURE 2. The total pipeline of the proposed.

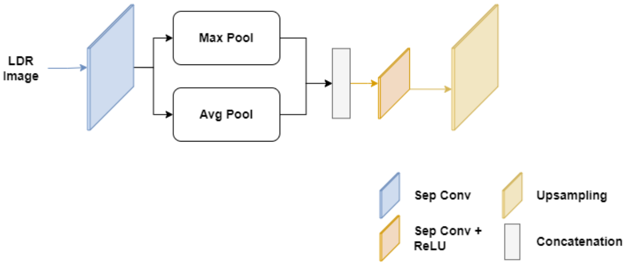

FIGURE 3. The structure of the feature extraction block.

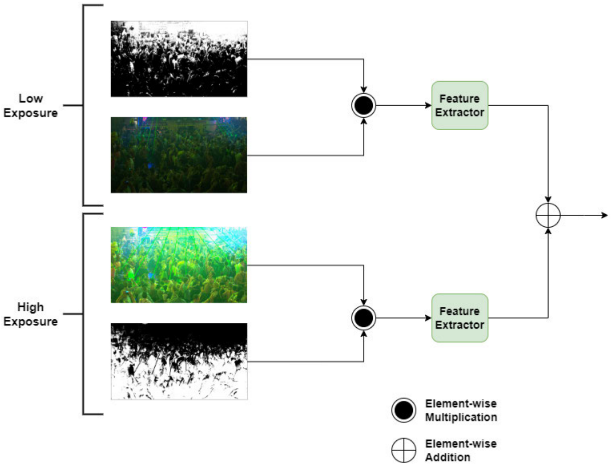

FIGURE 4. The structure of the visual attention module (VAM).

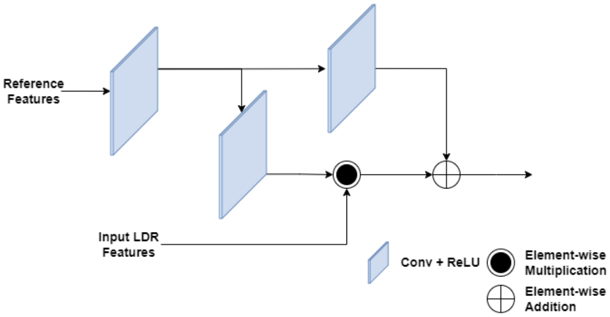

FIGURE 5. The structure of the spatial alignment module.

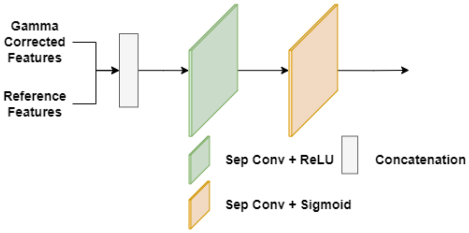

FIGURE 6. The structure of the attention module.

6) REFINEMENT

Unfortunately, the output of the reconstruction stage may have blurry, saturated, or dark areas; therefore, to cope with such possible issues with the help of features of the reference image, a refinement section also has been added.

As Fig. 9 illustrates, SepConv + ReLU is applied to the features of the reference image to reduce the number of feature maps. Furthermore, after concatenating the inputs, SepConv and SepConv + ReLU are used, respectively. The process is repeated two more times, and eventually, *Conv* + *Sigmoid* is applied to produce the final image in Sigmoid

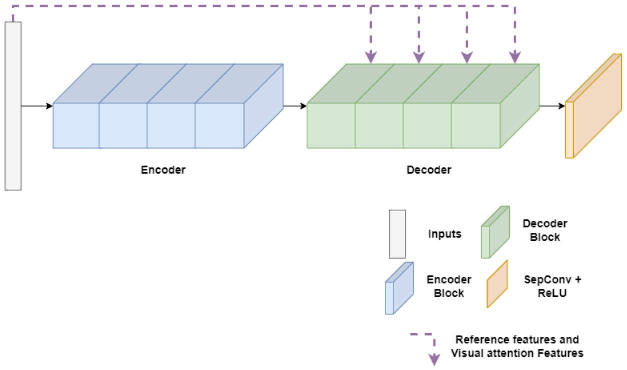

FIGURE 7. The total Scheme of the reconstruction stage.

space. The process in Refinement can be represented in pseudo-code as shown in Algorithm 1.

Algorithm 1 Pseudo-Code of Refinement Stage

```
1:  Inputs: The output of the Reconstruction stage (as $Reconstruction_o$) and extracted features of the referenced image ($f_r$).
2:  Output: The final image in the Sigmoid Space.
3:  $\hat{f}_r = \text{ReLU}(\text{SepConv}(f_r))$
4:  $i \leftarrow 0$
5:  while $i < 3$ do
6:      if $i == 0$ then
7:          $c \leftarrow concat(Reconstruction_o, \hat{f}_r)$
8:          $x \leftarrow \text{ReLU}(SepConv(SepConv(c)))$
9:      else
10:         $c \leftarrow concat(x, \hat{f}_r)$
11:         $x \leftarrow \text{ReLU}(SepConv(SepConv(c)))$
12:     end if
13:     $i \leftarrow i + 1$
14: end while
15: $out \leftarrow Sigmoid(Conv(x))$
```

As visible in Algorithm 1, the first two lines show the inputs and the output of the refinement stage. Moreover, the *concat*, *SepConv*, and *SepConv* + *ReLU* steps can be considered a block of the stage, respectively, which are applied three times. The first block receives the *reconstruction*$_o$ and $\hat{f}_r$ as inputs (lines 6-8). However, the output of the previous block and $\hat{f}_r$ are fed to the next blocks.

Notice that, in this research, the Ground Truth images are mapped from HDR Space into sigmoid space. Indeed, based on our experiments, transforming the values into sigmoid helps the network converge more conveniently (see Fig. 10) for a comparison of train and validation loss in Sigmoid and HDR space). The reason for changing the space is that the values in HDR space are too large, and a model with a low number of parameters is not able to learn to produce an HDR image correctly; conversely, by mapping them to the sigmoid space the values are converted between 0 and 1, which helps the proposed model to learn the data more efficiently.

TABLE 1. Brief highlights regarding the training and validation settings for the proposed method.

Dataset	NTIRE Challenge	
Optimizer	Adam Optimizer	
Initial LR	0.001 with LR decay	
	Train	Validation
Batch Size	16	2
Input Size	256x256	1920x1088
Augmentation	True	False
Epoch	100	
Loss	MAE	

IV. EXPERIMENTS AND RESULTS

A. DATASETS

For the experimental test and validation of the proposed method, standard benchmark datasets were used. The main dataset is the NTIRE dataset which was collected for the HDR Imaging Challenge (NTIRE) [44], [45]. In this dataset, two types of pictures (Single-Exposure and Multi-Exposure images) were provided; however, Multi-Exposure images only were used in this research. More specifically, this dataset includes images from [46] that were generated as follows. First, HDR images were produced natively by two Alexa Arri cameras with a mirror rig; then, their corresponding LDR images were generated synthetically with noise sources. There are approximately 1500 pairs of HDR/LDR images in this dataset for the training set, 40 for the validation set, and 200 pictures for the test set with a resolution of 1900 × 1060. However, in this research, we randomly selected 200 images of the training set as a test set and trained the model with around 1300 pairs.

In addition to the NTIRE dataset, we also tested our method on two other datasets: Kalantari et al. [16] and Hu et al. [47]. Both datasets contain dynamic scenes with large motions between the medium, low- and high-exposure images. Kalantari et al. dataset was created by capturing static scenes and introducing motion either by having a human actor move or by shifting the camera position between the acquisitions of the different LDR images. Hu et al. dataset consists of a set of sensor-realistic synthetic images generated using Unreal Engine and then calibrated to match the color gamut of a real sensor.

B. IMPLEMENTATION DETAILS

The highlights of the model are demonstrated in Table 1 briefly. Additionally, the weights of the model were initialized randomly and no pre-trained weights were used. Finally, the information regarding the proposed method will be discussed in the following subsections.

1) LOSS FUNCTION

Out of the various potential loss functions, the Mean Absolute Error (MAE) loss function has been chosen for training the model. This decision stems from the experimental findings outlined in [48], specifically in the closely related task of image denoising. In this study, the authors demonstrate

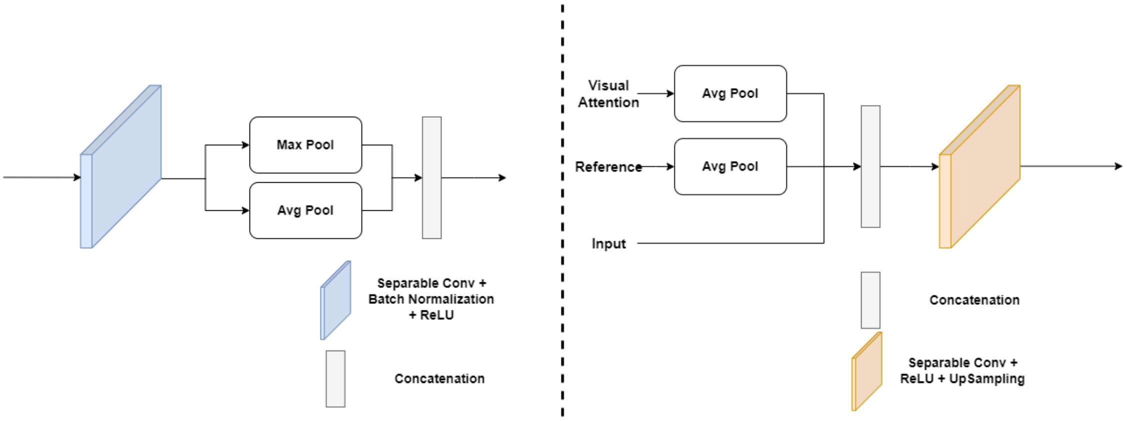

FIGURE 8. The structure of the blocks in the encoder (left) and the decoder (right).

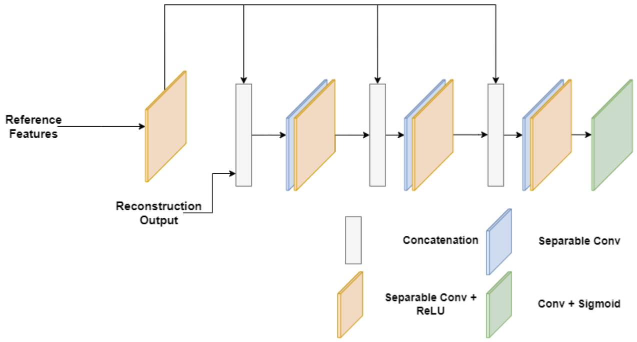

FIGURE 9. The structure of the refinement stage.

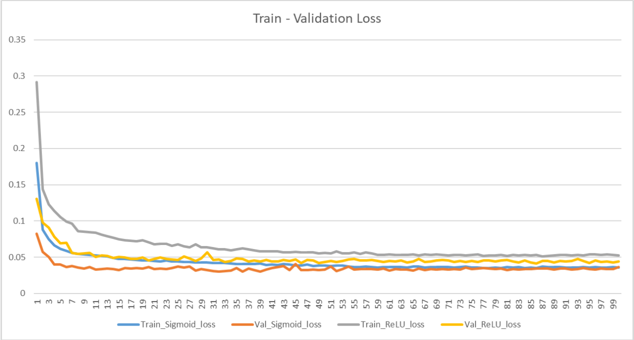

FIGURE 10. Train and validation loss in sigmoid and HDR spaces.

that the three loss functions MS-SSIM+MAE, MAE, and MS-SSIM consistently stand out as the best options. In this paper, MAE is favoured for its simplicity in terms of computation, making it a practical and effective choice for our model. Operatively, the difference is that the Ground Truth is first mapped to Sigmoid Domain, and eventually, MAE is calculated in Sigmoid space between the Ground Truth and the output of the model.

$$GT_n = sigmoid(GT) \tag{16}$$

$$L(\hat{y}, GT_n) = |GT_n - \hat{y}| \tag{17}$$

where GT_n is the Ground Truth image in the new domain, and L is the loss between Ground Truth and the output.

Furthermore, after training the model in sigmoid space, inverse sigmoid is used to re-map the output to HDR space. The inverse sigmoid can be written as follows:

$$HDR = \log(\frac{(\hat{y})}{1-\hat{y}}) \tag{18}$$

where HDR is the output in HDR space and $\hat{y}$ is the image in the sigmoid domain.

2) TRAINING

Flipping the images vertically or horizontally is also used as an augmentation method during training. Moreover, before feeding the images to the model, they are resized into 256 × 256. The reason for doing so instead of producing patches is that some generated patches from the masks may be totally black or completely white, which causes the model to pay less attention to the images with Short-Exposure.

Moreover, batch size and the number of epochs are set to 16 and 100, respectively. In this article, Adam Optimizer with an initial learning of 0.001 is used, and it will be reduced by a factor of 0.1 if the validation accuracy does not improve. Finally, the whole model is implemented in the Tensorflow (Keras) framework and is trained on a DGX-A100 GPU.

3) VALIDATION

The images are first padded from 1900 × 1060 to 1920 × 1080 and then fed to the model without any augmentation methods during validation.

C. EVALUATION METRICS AND COMPARISON

1) QUANTITATIVE COMPARISON

As Table 2 demonstrates, the results in this paper are compared with the SOTA methods by *PSNR* and *SSIM* in HDR and Tone-mapped domains. The $\mu - PSNR$ and $\mu - SSIM$ are the tone-mapped versions, where the images were tone-mapped in $\mu - law$. Moreover, in addition to

TABLE 2. Comparison with the SOTA methods, and ours also considering it without the refinement stage and the segmentation stage as described in Section V. The bold numbers are the best values, and the underlined ones are the second best.

Methods	PSNR	μ-PSNR	SSIM	μ-SSIM	LPIPS	delta-E	GMACs	Param. × 10^3
GSANet [24]	36.88	35.57	0.996	0.873	**0.02**	**0.40**	199.38	**80**
DRHDR [26]	38.5	**36.91**	0.996	0.86	0.21	**0.40**	1701.932	1190
Vien et al. [33]	39.44	35.39	0.994	0.837	0.34	0.45	**198.819**	1301
ours	**43.25**	35.86	**0.997**	**0.90**	0.03	0.57	234.107	570
ours-w-r	41.71	35.30	0.993	0.857	0.04	0.51	227.59	567
ours-w-s	40.27	34.99	0.993	0.842	0.05	0.66	223.96	545

TABLE 3. Comparison between the proposed method in HDR and sigmoid spaces.

Methods	PSNR	Mu-PSNR
Ours (HDR Space)	42.4	35.28
Ours (Sigmoid Space)	43.25	35.86

PSNR and SSIM, the results are compared with the SOTA methods by LPIPS [49], delta-E, *GMACs*, and the number of parameters. Learned Perceptual Image Patch Similarity (LPIPS) is a metric that computes the perceptual similarity of two images using a neural network. Delta-E is a metric that calculates the color difference of two images.

As mentioned in [45], the challenge focused on two tracks, which were Fidelity and low complexity. In the first one, the methods were required to obtain the highest $\mu - PSNR$ while the *GMACs* value is less than 200. In the latter track, it was asked to reduce the *GMACs* value to less than the baseline method while the *PSNR* and $\mu - PSNR$ values are almost the same as the baseline method. The proposed method has been compared with GSANet [24], DRHDR [26], and Vein et al. [33] methods. As can be seen, Table 2 shows the proposed method has the highest value in terms of *PSNR*, while having the second highest value in $\mu - PSNR$.

Additionally, all the methods were close in SSIM, however, our method was able to outperform the SOTA in both SSIM and $\mu - SSIM$. Furthermore, although our result with the value of 0.03 is the second best in LPIPS, it performed worst in delta-E. On the other hand, Vien et al. [33] had the lowest GMACs value, and GSANet is ranked second lowest. Moreover, it is visible that in terms of the number of parameters, GSANet has the lowest and the proposed method is in the second place among the algorithms. As Table 2 shows, ours-w-r and ours-w-s methods refer to ours without refinement and segmentation, and although the number of parameters and the value of GMACs in those methods are lower than the total model, still in terms of metrics the full model has a better result.

Furthermore, for more study, the proposed method was trained and tested in HDR and Sigmoid Spaces to check which space is superior for training the model. Thus, as Table 3 demonstrates, the proposed method in Sigmoid Space outperformed the algorithm in the HDR domain. Moreover, during training, the model in Sigmoid space converged quicker than the model in the HDR domain.

2) QUALITATIVE COMPARISON

In terms of qualitative comparison, we used the images of the NTIRE [44], [45], Kalantari [16], and Hu [47] datasets. As can be seen in Fig. 11, the produced images by ours, worked better in terms of image reconstruction compared to DRHDR and Vien et al. methods. More specifically, Fig. 11 demonstrates the results of ours, DRHDR [26], Vien et al. [33], and GSANet [24]. As can be seen, the output of Vien et al. in the first scene has distortion in the bright areas, and it is visible that the algorithm cannot restore the details from these areas correctly. Furthermore, there is some degradation in the dark regions too. Moreover, although DRHDR worked great and reconstructed both areas, this method was not able to acquire the details in over-saturated areas. For instance, looking at the two red and green boxes, the model did not reconstruct the details of the hands and the shirt, while the proposed method produced more detail in these two regions. Moreover, the produced image from the GSANet method shows significant details and is almost similar to ours. More precisely, although both methods could reconstruct the shirt nicely, the details of the hand in the GSANet are more than ours.

Additionally, in the second scene, the DRHDR and Vien et al. methods were not able to reconstruct the branches that were only visible in the short exposure image and restored only a part of them. In contrast, the proposed method and the GSANet worked almost well in this regard. Finally, looking at the last scene, it is visible that the proposed method outperformed the first two algorithms and reconstructed more details in both dark and bright areas, and the details of the sky show this point.

As further research, we tested the model and the SOTA on two other datasets with much more movement [16], [47]. Unfortunately, because all the models were trained on a dataset with low movement, they did not perform as well as they worked on the NTIRE dataset. Therefore, because the quantitative results were not as acceptable as the NTIRE dataset, we only used qualitative results for comparison. As can be seen in Fig. 12 (first scene), three methods including ours almost worked nicely to cope with the motion problems. However, GSANet encountered a ghosting issue. Additionally, similar to the NTIRE dataset, our model was able to reconstruct more local details compared to the other methods. On the other hand, as the second scene in Fig. 12 demonstrates, by having more motion, none of the methods

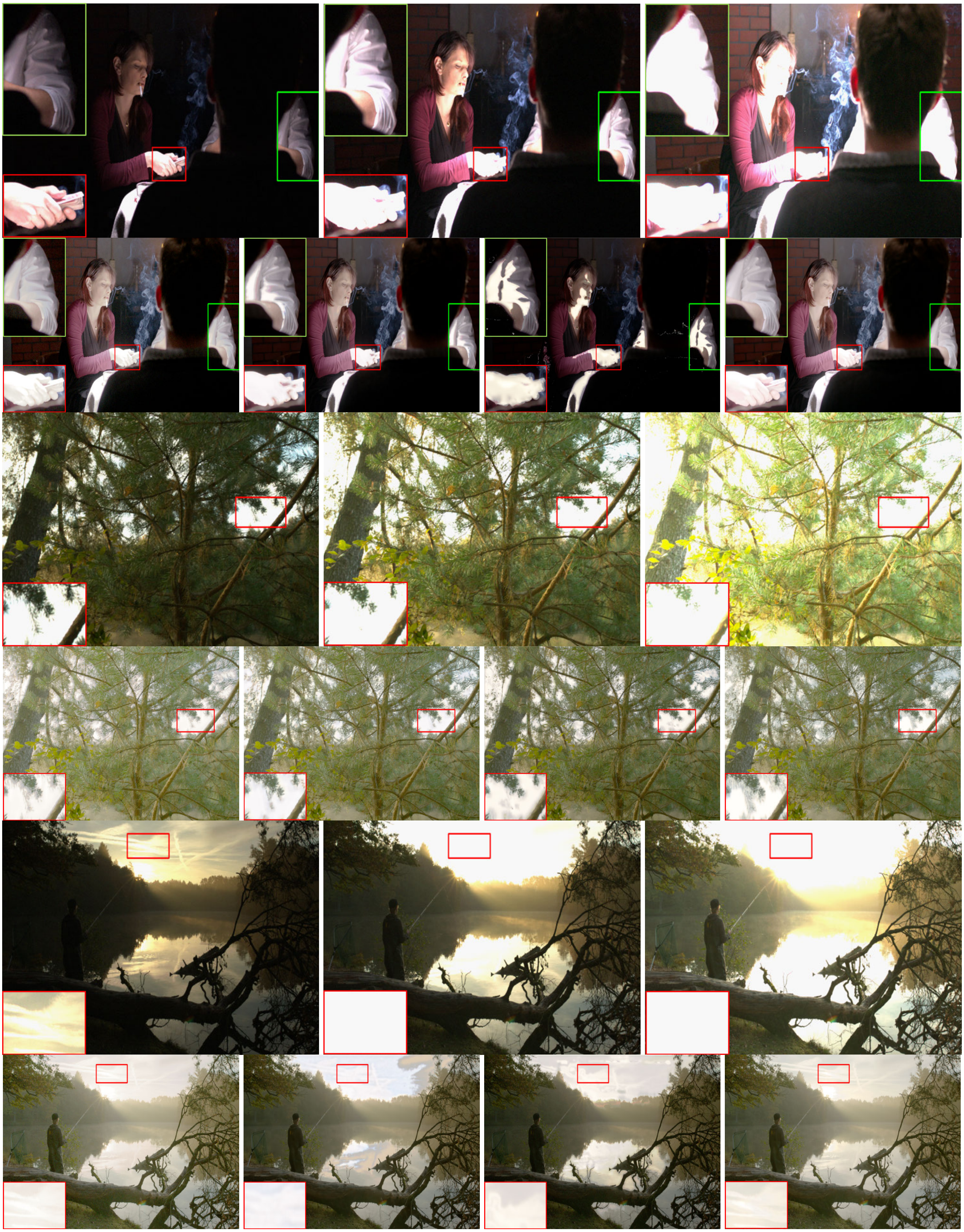

FIGURE 11. Qualitative comparison with the SOTA. The first row of each scene contains short, medium, and long exposure images, respectively. The second row includes ours, DRHDR, Vien et al., and GSANet outcomes, respectively.

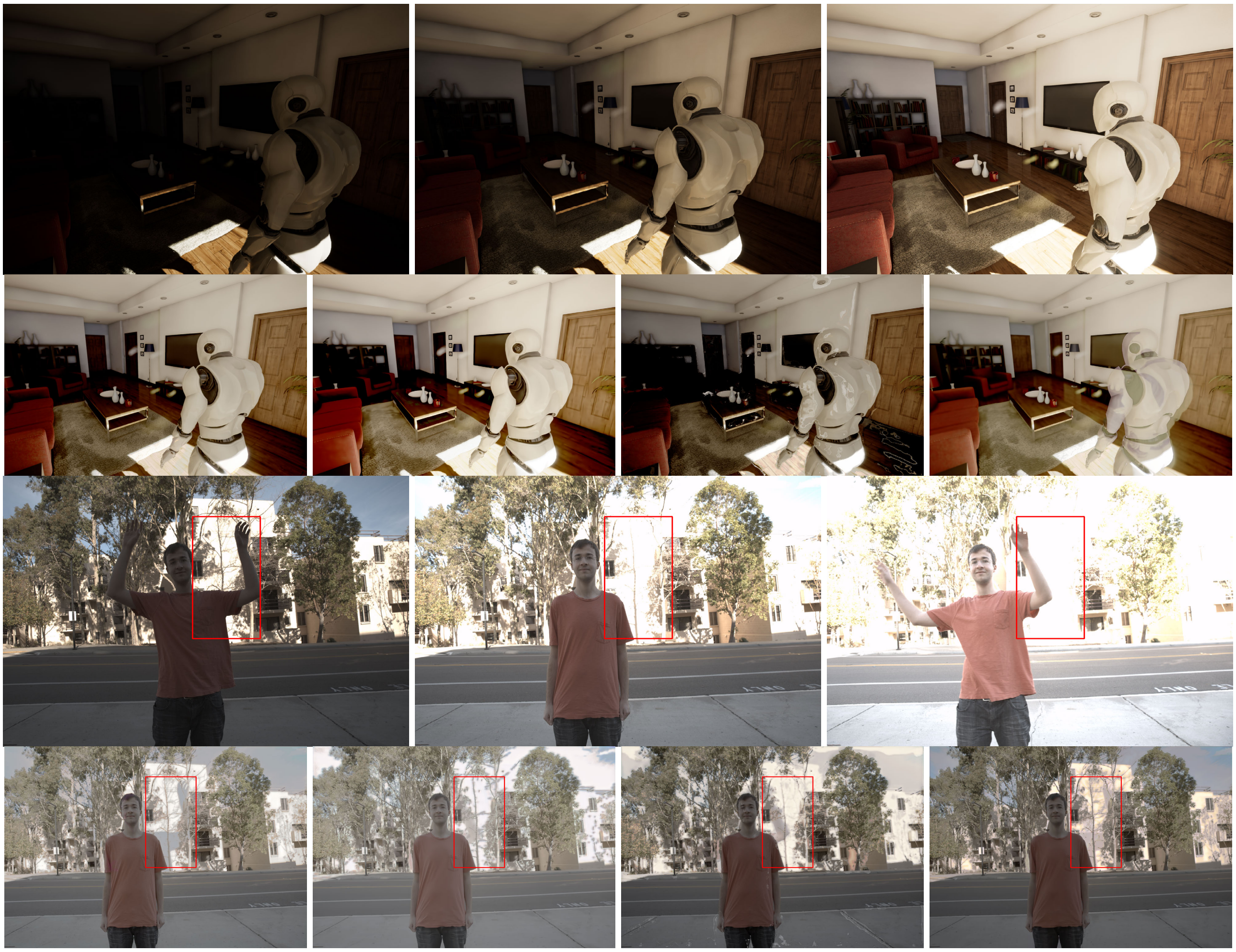

FIGURE 12. Qualitative comparison with the SOTA. The first row of each scene contains short, medium, and long exposure images, respectively. The second row includes ours, DRHDR, Vien et al., and GSANet outcomes, respectively. The First scene, and the second scene are acquired from [47] and [16], respectively.

can produce images without any ghosting problems. The red box in the images illustrates the common area where the methods had a ghosting issue.

Furthermore, although the segmentation helped the model to produce better results, the method might encounter two possible issues. Firstly, due to plausible noise in input images, using segmentation for extracting visible areas may also acquire the noise, and the produced image might become noisy. Lastly, although spatial alignment and attention modules are used to avoid any possible ghosting problems, the output might also encounter a ghosting issue if the input images have a severe amount of movement. Because the segmentation is applied to the Short- and Long-Exposure images and extracts their visible areas. Therefore, some parts of the images might not be aligned. Moreover, for future research, we would like to investigate possible methods to use segmentation and avoid any likely noise or misalignment.

V. ABLATION STUDY

In this paper, we proposed a model that included several stages, which are Attention, Reconstruction, and Refinement, and each one of the steps has an important role in this method. Therefore, to demonstrate the importance of each one, we removed the VAM and the refinement stages each time and retrained the model to compare them with the total method. Additionally, we tested our model with two other datasets [16], [47].

A. WITHOUT VISUAL ATTENTION MODULE

As mentioned in the III-C2, VAM is a helpful module that can help the model reconstruct a better image. Therefore, to demonstrate this statement, we kept the refinement stage, retrained the model without the VAM module, and compared the results with the main model.

Given the results in Fig 13, the segmentation has both benefits and drawbacks. The zoomed part of the images demonstrates that the segmentation helps the model to reconstruct the details. As can be seen, the model was able to reconstruct the wall and the cracks in the ground better than the model without the segmentation stage. Moreover, although the model was able to cope with the motion in

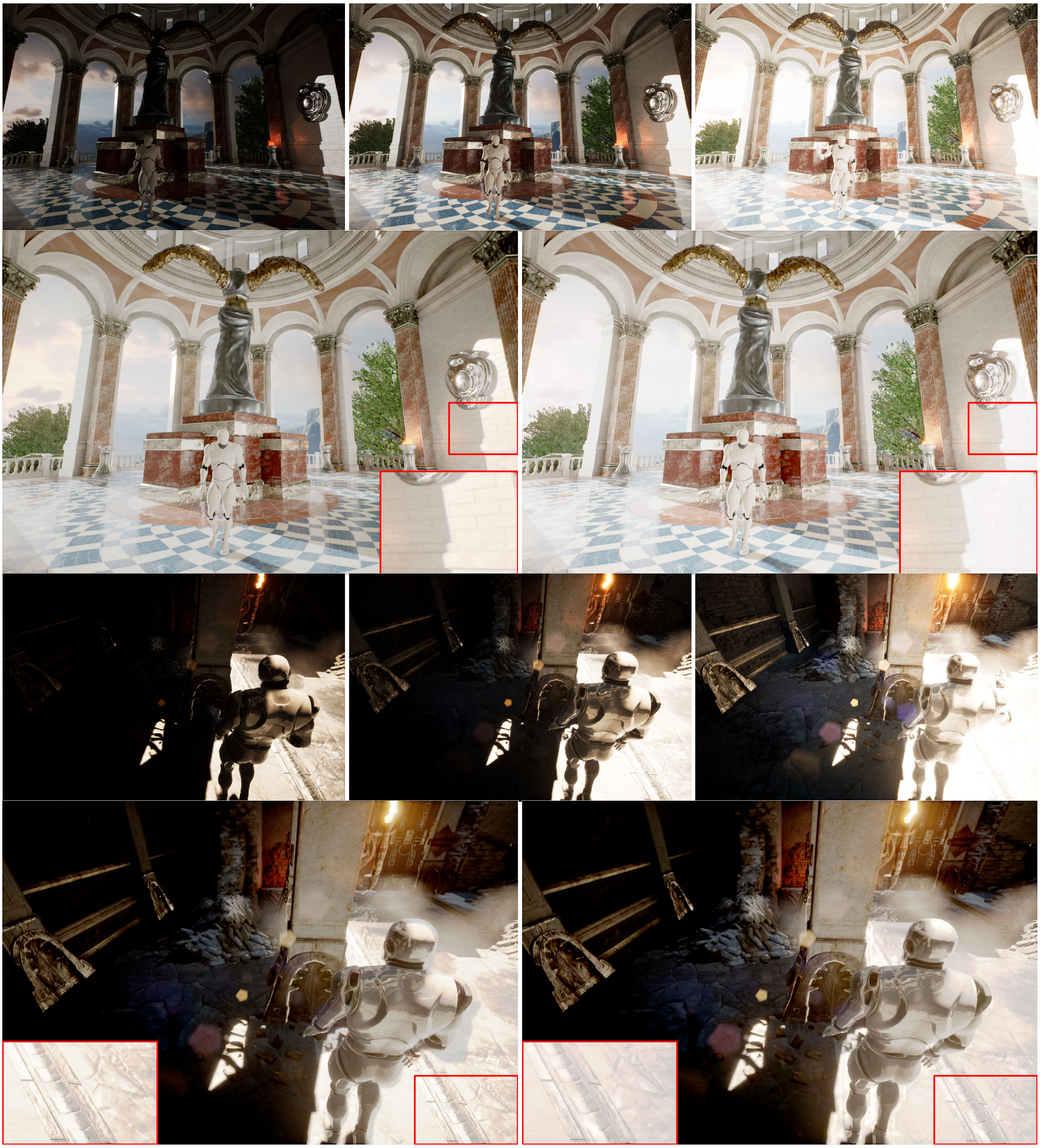

FIGURE 13. Qualitative comparison between the proposed method (on the left) and proposed method without VAM module (on the right). The image is acquired from [47].

the first scene of Fig 13, the model could not resolve the motion problem in the second scene due to the high volume of movement. Unfortunately, due to keeping the information from each exposure, the VAM module causes the ghosting problem.

B. WITHOUT REFINEMENT

Additionally, it was mentioned in III-C6, that the Refinement stage was used to cope with possible distortions. Therefore, we retained the Segmentation part and retrained the model without the Refinement.

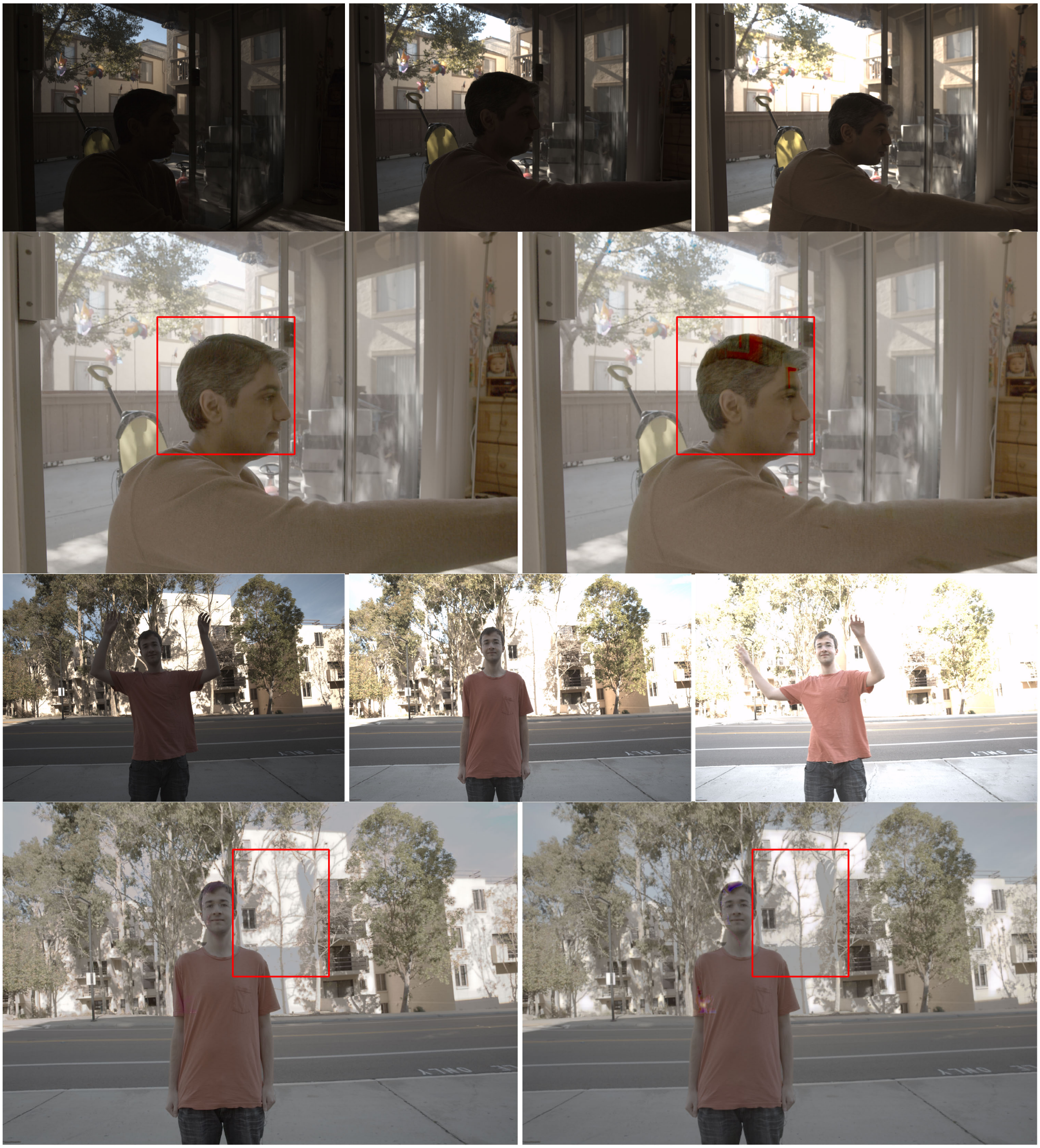

FIGURE 14. Qualitative comparison between the proposed method (on the left) and the proposed method without the refinement stage (on the right). The image is acquired from [16].

Given the outcomes in Fig 14, the first row of each scene contains input images, and the second row includes the outputs. The outputs on the right illustrate that the model without the refinement step causes distortion in both under-exposed and overexposed areas. More precisely, the specified box in the first scene indicates that the hair of the person is not reconstructed well and is noisy, while the complete model was able to reconstruct it well. Additionally, as can be seen in the second scene, the outputs of both models contain the ghosting problem. Moreover, the produced results of the model without the refinement stage suffer from the lack of local details.

VI. CONCLUSION

In this article, we proposed a new method for HDR imaging with the help of image segmentation. More specifically, we first applied the Otsu method on Short- and Long-Exposure images to acquire the areas with more details. Afterward, the input images along with the segmentation outputs were fed to the model to produce the HDR image. The results show that the proposed method outperformed the SOTA and generated more details. However, the proposed model is not free of issues, and in case of possible noise or misalignment in input images, the output might have a slight amount of noise or misalignment due to extracting areas of input images. More exactly, the experiments show that the model is incapable of producing a ghosting-free image when the level of motion is high because of the Segmentation stage. Therefore, for future research, we would like to focus on investigating these two problems.

REFERENCES

[1] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2000, pp. 472–479.

[2] J. Tumblin, A. Agrawal, and R. Raskar, "Why I want a gradient camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2005, pp. 103–110.

[3] M. McGuire, W. Matusik, H. Pfister, B. Chen, J. F. Hughes, and S. K. Nayar, "Optical splitting trees for high-precision monocular imaging," *IEEE Comput. Graph. Appl.*, vol. 27, no. 2, pp. 32–42, Mar. 2007.

[4] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011.

[5] S. Hajisharif, J. Kronander, and J. Unger, "Adaptive dualISO HDR reconstruction," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, p. 41, Dec. 2015.

[6] H. Zhao, B. Shi, C. Fernandez-Cull, S.-K. Yeung, and R. Raskar, "Unbounded high dynamic range photography using a modulo camera," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, Apr. 2015, pp. 1–10.

[7] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masia, "Convolutional sparse coding for high dynamic range imaging," in *Proc. 37th Annu. Conf. Eur. Assoc. Comput. Graph.*, 5555, pp. 153–163.

[8] 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.*, vol. 36, no. 6, pp. 1–15, Nov. 2017.

[9] L. She, M. Ye, S. Li, Y. Zhao, C. Zhu, and H. Wang, "Single-image HDR reconstruction by dual learning the camera imaging process," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105947.

[10] P.-H. Le, Q. Le, R. Nguyen, and B.-S. Hua, "Single-image HDR reconstruction by multi-exposure generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4052–4061.

[11] G. Cao, F. Zhou, K. Liu, A. Wang, and L. Fan, "A decoupled kernel prediction network guided by soft mask for single image HDR reconstruction," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2s, pp. 1–23, Feb. 2023.

[12] A. Omrani, M. R. Soheili, and M. Kelarestaghi, "High dynamic range image reconstruction using multi-exposure wavelet HDRCNN," in *Proc. Int. Conf. Mach. Vis. Image Process. (MVIP)*, Feb. 2020, pp. 1–4.

[13] S. Lee, G. H. An, and S.-J. Kang, "Deep chain HDRI: Reconstructing a high dynamic range image from a single low dynamic range image," *IEEE Access*, vol. 6, pp. 49913–49924, 2018.

[14] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–10, Nov. 2017.

[15] A. Biswas, M. S. Patel, and B. H. P. Prasad, "Deep multi-stage learning for HDR with large object motions," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4714–4718.

[16] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017.

[17] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 1–12.

[18] Q. Yan, D. Gong, Q. Shi, A. van den Hengel, C. Shen, I. Reid, and Y. Zhang, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1751–1760.

[19] K. R. Prabhakar, R. Arora, A. Swaminathan, K. P. Singh, and R. V. Babu, "A fast, scalable, and reliable deghosting method for extreme exposure fusion," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2019, pp. 1–8.

[20] K. R. Prabhakar, S. Agrawal, D. K. Singh, B. Ashwath, and R. V. Babu, "Towards practical and efficient high-resolution HDR deghosting with CNN," in *Computer Vision—ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 497–513.

[21] V. G. An and C. Lee, "Single-shot high dynamic range imaging via deep convolutional neural network," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1768–1772.

[22] Q. Yan, D. Gong, P. Zhang, Q. Shi, J. Sun, I. Reid, and Y. Zhang, "Multi-scale dense networks for deep high dynamic range imaging," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 41–50.

[23] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[24] F. Li, R. Gang, C. Li, J. Li, S. Ma, C. Liu, and Y. Cao, "Gamma-enhanced spatial attention network for efficient high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1031–1039.

[25] Y. Deng, Q. Liu, and T. Ikenaga, "Attention-guided network with inverse tone-mapping guided up-sampling for HDR imaging of dynamic scenes," *Multimedia Tools Appl.*, vol. 81, no. 9, pp. 12925–12944, Apr. 2022.

[26] J. Marín-Vega, M. Sloth, P. Schneider-Kamp, and R. Röttger, "DRHDR: A dual branch residual network for multi-bracket high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 843–851.

[27] Q. Ye, M. Suganuma, J. Xiao, and T. Okatani, "Learning regularized multi-scale feature flow for high dynamic range imaging," 2022, *arXiv:2207.02539*.

[28] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Multi-scale sampling and aggregation network for high dynamic range imaging," *ArXiv*, vol. abs/2208.02448, 2022.

[29] Q. Yan, S. Zhang, W. Chen, Y. Liu, Z. Zhang, Y. Zhang, J. Q. Shi, and D. Gong, "A lightweight network for high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 823–831.

[30] G. Yu, J. Zhang, Z. Ma, and H. Wang, "Efficient progressive high dynamic range image restoration via attention and alignment network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1123–1130.

[31] T. Dai, W. Li, X. Cao, J. Liu, X. Jia, A. Leonardis, Y. Yan, and S. Yuan, "Wavelet-based network for high dynamic range imaging," *Comput. Vis. Image Understand.*, vol. 238, Jan. 2024, Art. no. 103881.

[32] K. Ram Prabhakar, S. Agrawal, and R. Venkatesh Babu, "Segmentation guided deep HDR deghosting," 2022, *arXiv:2207.01229*.

[33] A. G. Vien, S. Park, T. T. N. Mai, G. Kim, and C. Lee, "Bidirectional motion estimation with cyclic cost volume for high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1182–1189.

[34] N. Messikommer, S. Georgoulis, D. Gehrig, S. Tulyakov, J. Erbach, A. Bochicchio, Y. Li, and D. Scaramuzza, "Multi-bracket high dynamic range imaging with event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 546–556.

[35] Z. Liu, Y. Wang, B. Zeng, and S. Liu, "Ghost-free high dynamic range imaging with context-aware transformer," in *Computer Vision—ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham, Switzerland: Springer, 2022, pp. 344–360.

[36] A. Vadivel, M. Mohan, S. Sural, and A. K. Majumdar, "Segmentation using saturation thresholding and its application in content-based retrieval of images," in *Image Analysis and Recognition*, A. Campilho and M. Kamel, Eds. Berlin, Germany: Springer, 2004, pp. 33–40.

[37] Y. Kinoshita and H. Kiya, "Scene segmentation-based luminance adjustment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4101–4116, Aug. 2019.

[38] Y. Kinoshita and H. Kiya, "Automatic exposure compensation using an image segmentation method for single-image-based multi-exposure fusion," *APSIPA Trans. Signal Inf. Process.*, vol. 7, no. 1, p. e22, 2018.

[39] B. D. Lee and M. H. Sunwoo, "HDR image reconstruction using segmented image learning," *IEEE Access*, vol. 9, pp. 142729–142742, 2021.

[40] A. Mitiche and I. B. Ayed, *Variational and Level Set Methods in Image Segmentation*, vol. 5. Cham, Switzerland: Springer, 2010.

[41] F. Yi and I. Moon, "Image segmentation: A survey of graph-cut methods," in *Proc. Int. Conf. Syst. Informat. (ICSAI)*, May 2012, pp. 1936–1941.

[42] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[43] L. Wang and K.-J. Yoon, "Deep learning for HDR imaging: State-of-the-art and future trends," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8874–8895, Dec. 2022.

[44] E. Pérez-Pellitero, S. Catley-Chandar, A. Leonardis, and R. Timofte, "NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 691–700.

[45] E. Pérez-Pellitero et al., "NTIRE 2022 challenge on high dynamic range imaging: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1008–1022.

[46] 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*, vol. 9023, Jul. 1117, Art. no. 90230X.

[47] J. Hu, G. Choe, Z. Nadir, O. Nabil, S.-J. Lee, H. Sheikh, Y. Yoo, and M. Polley, "Sensor-realistic synthetic data engine for multi-frame high dynamic range photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2180–2189.

[48] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

ALI REZA OMRANI was born in Deir, Bushehr, Iran, in 1995. He received the bachelor's degree in software engineering from the Shahid Bahonar Technical College, Shiraz, in 2016, and the master's degree in artificial intelligence from Kharazmi University, in 2020. He is currently pursuing the Ph.D. degree with the Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy, and the Institute of Information Science and Technologies (ISTI), National Research Council of Italy, Pisa, Italy.

His current research interests include image reconstruction, image enhancement, high dynamic range imaging, and behavior analysis.

DAVIDE MORONI received the M.Sc. degree (Hons.) in mathematics from the University of Pisa, in 2001, the Diploma degree from Scuola Normale Superiore di Pisa, in 2002, and the Ph.D. degree in mathematics from the University of Rome La Sapienza, in 2006. He is a Researcher with the Institute of Information Science and Technologies (ISTI), National Research Council, Italy, Pisa, where he is currently the Head of the Signals and Images Laboratory. His main research interests include geometric modeling, computational topology, image processing, computer vision, and medical imaging. He is the Chair of the MUSCLE Working Group, European Consortium for Informatics and Mathematics. Since 2018, he has been the Chair of the Technical Committee 16 on Algebraic and Discrete Mathematical Techniques in Pattern Recognition and Image Analysis of the International Association for Pattern Recognition (IAPR). He is an Associate Editor of *IET Image Processing*.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE-ITALY'
within the CRUI CARE Agreement