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, Pisa, Italy
[2]Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy
[*]{ali.omrani, davide.moroni}@isti.cnr.it

July 28, 2023

1 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 methods in this topic perform well, they mainly concentrate on combining different exposures and have 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 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 State-Of-The-Art algorithms.

Keywords: Deep Neural Network, High Dynamic Range imaging, Image Segmentation, Multi-exposure Image, Visual Attention Module

2 Introduction

In the scope of photography, the real world consists of an unlimited range of luminance. However, most devices are capable of capturing merely limited 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 [1–7] 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 mostly 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 [8–11] 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 not capable 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 [15–20].

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 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. Thus, to reach this point, Visual Attention Modules (VAMs) will be proposed to obtain such regions. Moreover, in this research, Spatial and Attention modules have been used from a State-Of-The-Art method, 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 State-Of-The-Art methods, the results still illustrated a slight amount of noise

that was extracted from the input images.

In section 3, the State-Of-The-Art in HDR imaging and related image segmentation is presented. In section 4, the proposed method is discussed in detail. Section 5 demonstrates the experimental results and comparison with the State-Of-The-Art methods. Moreover, section 6 concludes this article with ideas for further works. Finally, the code will be available at the **github page**.

3 Related work

In this section, we will discuss the State-Of-The-Art methods in the scope of HDR imaging in the Multi-Exposure category (Section 3.1) and survey unsupervised Image Segmentation methods for extracting regions (Section 3.2).

3.1 Multi-Exposure Methods

[24] proposed a two-stage algorithm, in which 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 at 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 are 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 State-Of-The-Art 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.

[31] demonstrated that the ghosting problem is mainly in short-frequency signals, and therefore, they proposed a wavelet-based model to merge images in the frequency domain and avoid any ghosting problems. [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 esti-

mated 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. [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. [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.

3.2 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].

4 Proposed Method

4.1 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.

4.2 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.

4.2.1 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 the neural network and Otsu method were used for the image segmentation stage; however, the neural network resulted in overfitting. Thus, the Otsu method has been selected to segment the visible areas of the pictures. Therefore, the images are converted into the YUV color space to calculate a threshold based on the histograms of Short- and Long-Exposure images.

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

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.

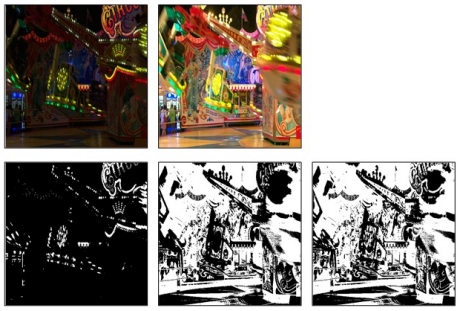

Figure 1: Produced masks of Short- and Long-Exposure images.

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, 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.

4.3 Proposed Method Structure

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

4.3.1 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:

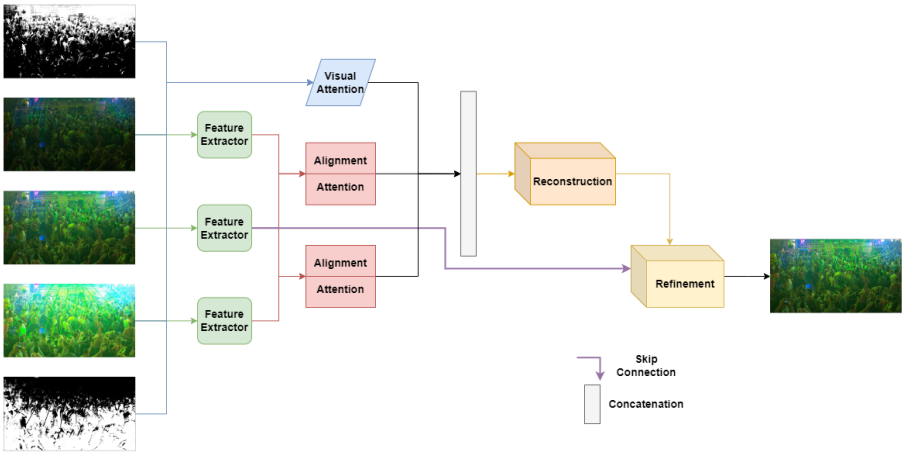

Figure 2: The total pipeline of the proposed.

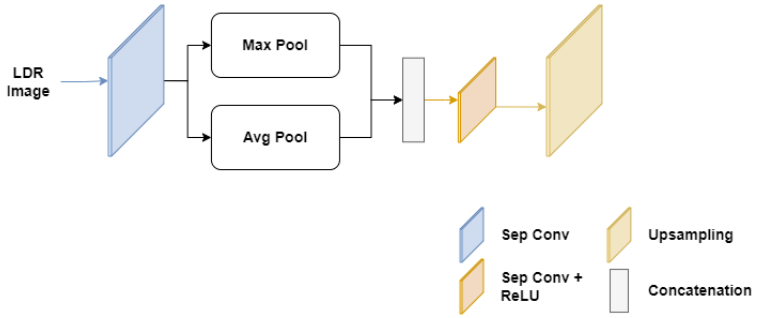

Figure 3: The structure of the Feature Extraction Block.

$$C_i = \text{concat}\left(M(\text{SepConv}(I_i)), A(\text{SepConv}(I_i))\right) \tag{5}$$

$$F_i = \text{Upsample}(\text{ReLU}(\text{SepConv}(C_i))) \tag{6}$$

for $i = 1, 2, 3$, where $A()$ and $M()$ 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.

4.3.2 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

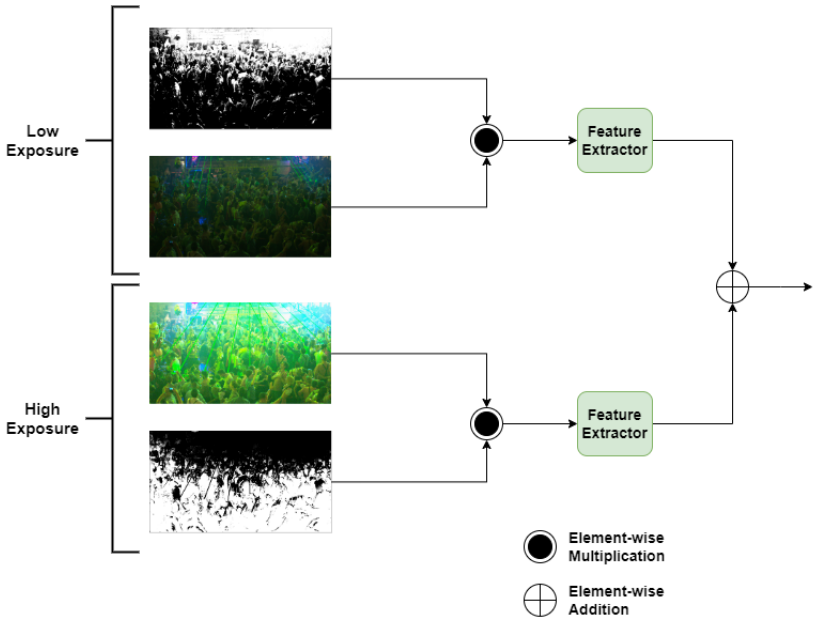

Figure 4: The structure of the Visual Attention Module (VAM).

Features. Finally, they are added together element-wisely. The VAM can be formally defined as follows:

$$\text{features}_L = F(\text{multiply}(\text{mask}_L, I_L)) \tag{7}$$

$$\text{features}_H = F(\text{multiply}(\text{mask}_H, I_H)) \tag{8}$$

$$V = \text{add}(\text{features}_L, \text{features}_H) \tag{9}$$

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

4.3.3 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-wisely 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-wisely with M_i. Formally, the operation in the module can be written as follows:

$$\text{Ref}_1 = \text{ReLU}(\text{Conv}(\text{ref features})) \tag{10}$$

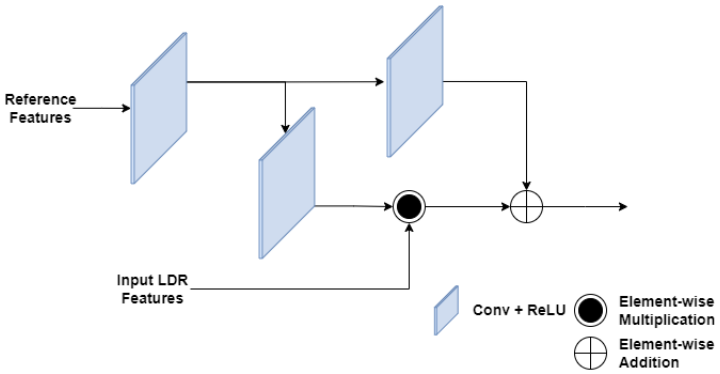

Figure 5: The structure of the Spatial Alignment Module.

$$M_i = \text{multiply}(\text{ReLU}(\text{Conv}(\text{Ref}_1)),\ \text{inp features}_i) \tag{11}$$

$$\text{out}_i = \text{add}(\text{ReLU}(\text{Conv}(\text{Ref}_1)),\ M_i) \tag{12}$$

4.3.4 Attention Module

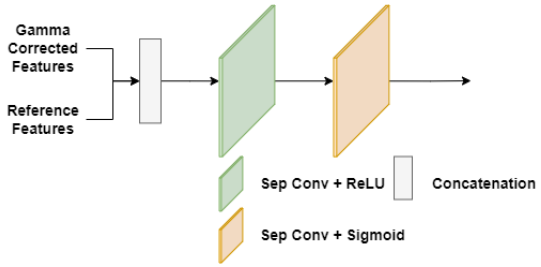

Figure 6: The structure of the Attention Module.

The Attention Module is almost similar to [30], 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:

$$R_i = \text{ReLU}(\text{SepConv}(\text{concat}(f_i, f_r)) \quad \text{for } i = 1, 3 \tag{13}$$

$$S_i = \text{Sigmoid}(\text{SepConv}(R_i)) \tag{14}$$

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

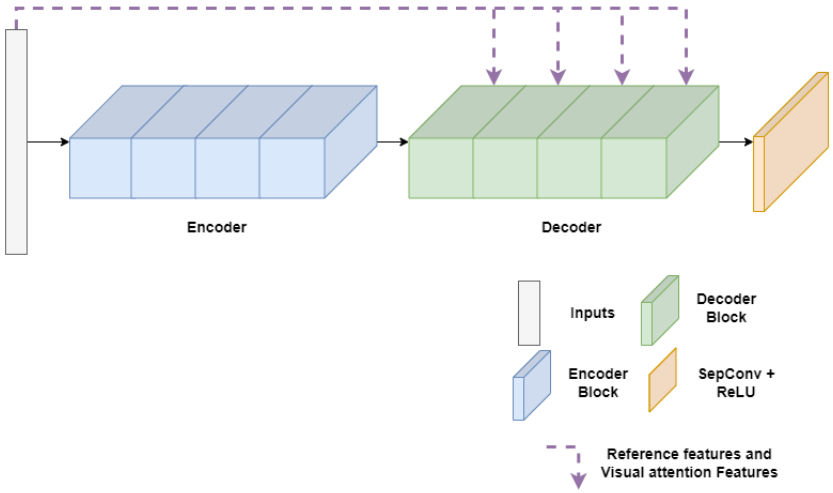

Figure 7: The total Scheme of the Reconstruction stage.

4.3.5 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.

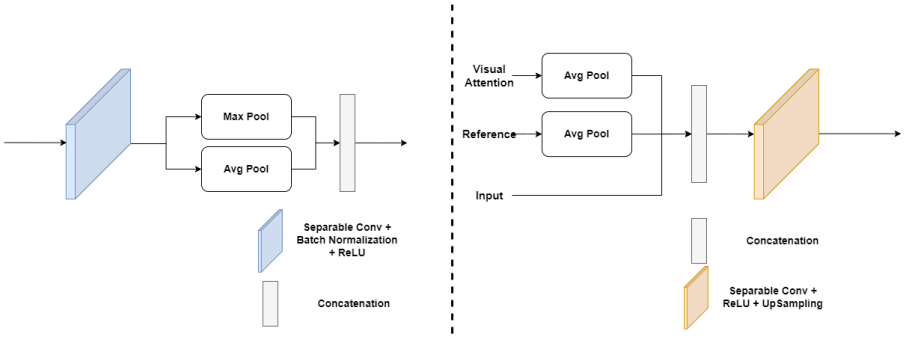

Figure 8: The structure of the blocks in the encoder (**left**) and the decoder (**right**).

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.

4.3.6 Refinement

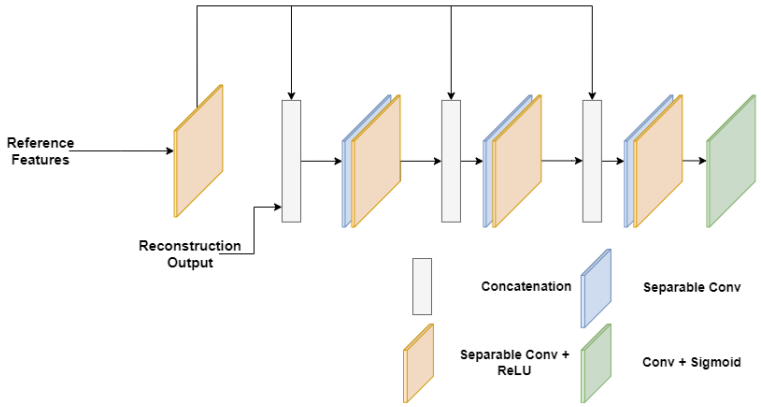

Figure 9: The structure of the Refinement Stage.

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

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. 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; therefore, by mapping them to sigmoid space, the proposed model outperforms the model in the HDR space.

Algorithm 1 Pseudo-code of Refinement Stage.

$\hat{f}_r = \text{ReLU}(\text{SepConv}(f_r))$
$i \leftarrow 0$
while $i < 3$ **do**
  if $i == 0$ **then**
    $c \leftarrow concat(Reconstruction_o, \hat{f}_r)$
    $x \leftarrow \text{ReLU}(ConvSep(ConvSep(c)))$
  else
    $c \leftarrow concat(x, \hat{f}_r)$
    $x \leftarrow \text{ReLU}(ConvSep(ConvSep(c)))$
  end if
  $i \leftarrow i + 1$
end while
$out \leftarrow Sigmoid(Conv(x))$

5 Experiments and Results

5.1 Dataset

A new dataset was collected for HDR Imaging Challenge [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 1900x1060. 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.

5.2 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.

5.2.1 Loss function

The Mean Absolute Error (MAE) loss function is used to train the model. 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.

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	

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

$$GT_n = sigmoid(GT) \tag{15}$$

$$L(\hat{y}, GT_n) = |GT_n - \hat{y}| \tag{16}$$

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:

$$HDR = \log(\frac{(\hat{y})}{1 - \hat{y}}) \tag{17}$$

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

5.2.2 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 256x256. 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 Tensorflow (Keras) framework and is trained on a DGX-A100 GPU.

5.2.3 Validation

The images are first padded from 1900x1060 to 1920x1080 and then fed to the model without any augmentation methods during validation.

Methods	PSNR	μ-PSNR	GMACs	Param. $\times 10^3$
GSANet	36.88	35.57	<u>199.38</u>	**80**
DRHDR	38.5	**36.91**	1701.932	1190
Vien et al.	<u>39.44</u>	35.39	**198.819**	1301
ours	**43.25**	<u>35.86</u>	234.107	<u>570</u>

Table 2: Comparison with the State-Of-The-Art methods. The bold numbers are the best values, and the underline ones are the second best.

5.3 Evaluation Metrics and Comparison

5.3.1 Quantitative Comparison

The results in this paper are compared with the State-Of-The-Art methods by $PSNR$ in HDR and Tone-mapped domains. The $PSNR-\mu$ is the tone-mapped version, where the images were tone-mapped in $\mu - law$. Moreover, the results are compared with the State-Of-The-Art methods in $GMACs$ and the number of parameters.

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 $\mu - PSNR$. 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.

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

Table 3: Comparison between the proposed method in HDR and Sigmoid Spaces.

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.

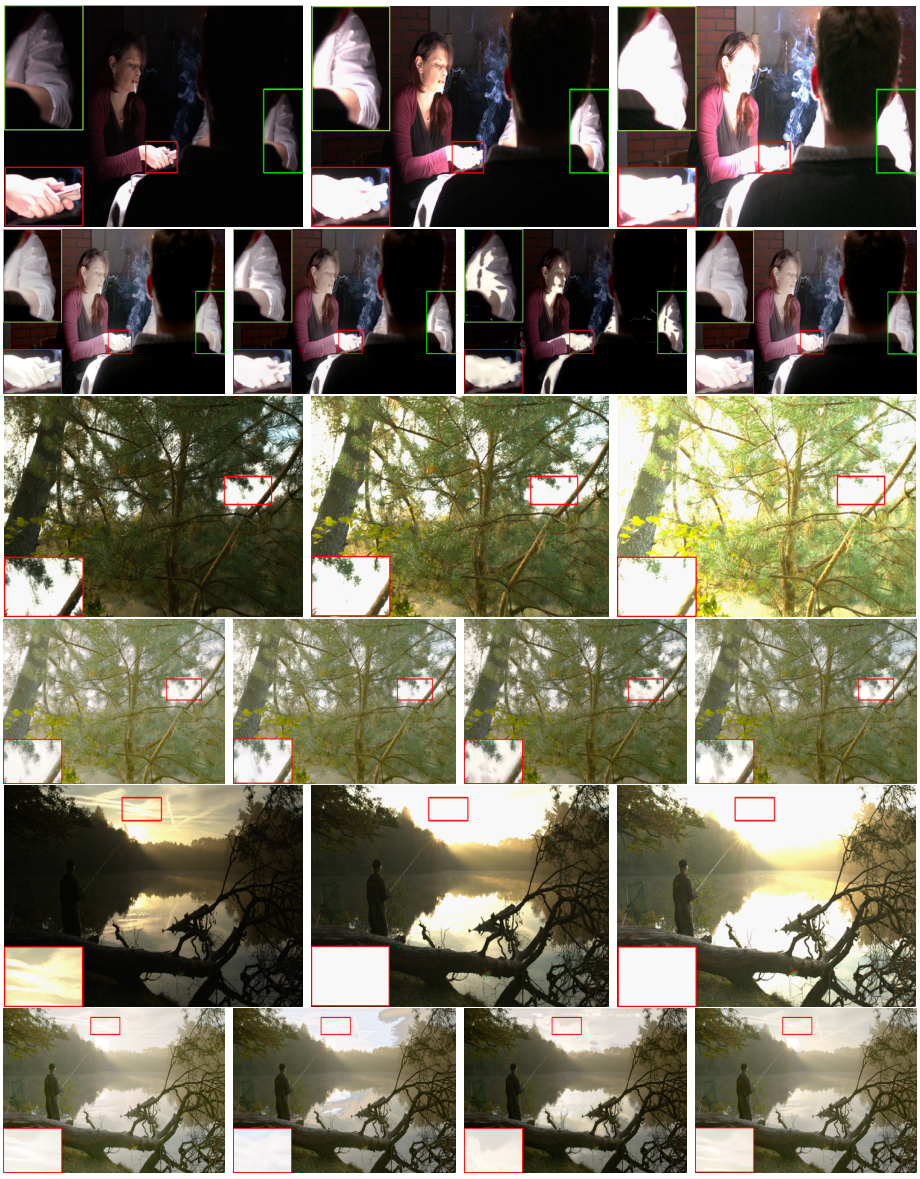

Figure 10: Qualitative Comparison with the State-Of-The-Art. The first row of each scene contains short, medium, and long exposure images, respectively. The second row includes ours, DRHDR, An et al., and GSANet outcomes, respectively.

5.3.2 Qualitative Comparison

As can be seen in Fig. 10, the produced images by ours, worked better in terms of image reconstruction compared to DRHDR and An et al. methods. More specifically, Fig. 10 demonstrates the results of ours, DRHDR [26], An et al. [33], and GSANet [24]. As can be seen, the output of An 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, 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 An 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.

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, if the input images have a severe amount of movement, the output might also encounter a ghosting issue. 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.

6 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 State-Of-The-Art 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. Therefore, for future research, we would like to focus on investigating these two problems.

References

[1] S. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, vol. 1, pp. 472–479 vol.1, 2000.

[2] J. Tumblin, A. Agrawal, and R. Raskar, "Why i want a gradient camera," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 103–110 vol. 1, 2005.

[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 Computer Graphics and Applications*, vol. 27, no. 2, pp. 32–42, 2007.

[4] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile hdr video production system," *ACM Trans. Graph.*, vol. 30, jul 2011.

[5] S. Hajisharif, J. Kronander, and J. Unger, "Adaptive dualiso hdr reconstruction," *EURASIP Journal on Image and Video Processing*, vol. 2015, 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 *2015 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, 2015.

[7] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masia, "Convolutional sparse coding for high dynamic range imaging," in *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics*, EG '16, (Goslar, DEU), p. 153–163, Eurographics Association, 2016.

[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, 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," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105947, 2023.

[10] P.-H. Le, Q. Le, R. Nguyen, and B.-S. Hua, "Single-image hdr reconstruction by multi-exposure generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023.

[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, feb 2023.

[12] A. Omrani, M. R. Soheili, and M. Kelarestaghi, "High dynamic range image reconstruction using multi-exposure wavelet hdrcnn," in *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pp. 1–4, 2020.

[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, nov 2017.

[15] G. R. K.S., A. Biswas, M. S. Patel, and B. H. P. Prasad, "Deep multi-stage learning for hdr with large object motions," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4714–4718, 2019.

[16] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, jul 2017.

[17] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[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 *2019 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–8, 2019.

[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), pp. 497–513, Springer International Publishing, 2020.

[21] V. G. An and C. Lee, "Single-shot high dynamic range imaging via deep convolutional neural network," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1768–1772, 2017.

[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 *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 41–50, 2019.

[23] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1032–1040, June 2022.

[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 and Applications*, vol. 81, pp. 12925–12944, Apr 2022.

[26] J. Mar'in-Vega, M. Sloth, P. Schneider-Kamp, and R. Rottger, "Drhdr: A dual branch residual network for multi-bracket high dynamic range imaging," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 843–851, 2022.

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

[28] J. Xiao, Q. Ye, T. Liu, C. Zhang, and K.-M. Lam, "Multi-scale sampling and aggregation network for high dynamic range imaging," 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 823–831, 2022.

[30] G. Yu, J. Zhang, Z. Ma, and H. Wang, "Efficient progressive high dynamic range image restoration via attention and alignment network," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1123–1130, 2022.

[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," 2022.

[32] K. R. Prabhakar, S. Agrawal, and R. V. Babu, "Segmentation guided deep hdr deghosting," 2022.

[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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1182–1189, 2022.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 547–557, June 2022.

[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), pp. 344–360, Springer Nature Switzerland, 2022.

[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, Heidelberg), pp. 33–40, Springer Berlin Heidelberg, 2004.

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

[38] Y. Kinoshita and H. Kiya, "Automatic exposure compensation using an image segmentation method for single-image-based multi-exposure fusion," *APSIPA Transactions on Signal and Information Processing*, vol. 7, 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. Springer Science & Business Media, 2010.

[41] F. Yi and I. Moon, "Image segmentation: A survey of graph-cut methods," in *2012 international conference on systems and informatics (ICSAI2012)*, pp. 1936–1941, IEEE, 2012.

[42] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[43] L. Wang and K.-J. Yoon, "Deep learning for hdr imaging: State-of-the-art and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8874–8895, 2022.

[44] E. Pérez-Pellitero, S. Catley-Chandar, A. Leonardis, R. Timofte, X. Wang, Y. Li, T. Wang, F. Song, Z. Liu, W. Lin, X. Li, Q. Rao, T. Jiang, M. Han, H. Fan, J. Sun, S. Liu, , *et al.*, "Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 691–700, 2021.

[45] E. Pérez-Pellitero, S. Catley-Chandar, R. Shaw, A. Leonardis, R. Timofte, Z. Zhang, C. Liu, Y. Peng, Y. Lin, G. Yu, *et al.*, "Ntire 2022 challenge on high dynamic range imaging: Methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1009–1023, 2022.

[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," in *Digital Photography X* (N. Sampat, R. Tezaur, S. Battiato, and B. A. Fowler, eds.), vol. 9023, p. 90230X, International Society for Optics and Photonics, SPIE, 2014.