# SUPERVISED IMAGE SEGMENTATION FOR HIGH DYNAMIC RANGE IMAGING

*Ali Reza Omrani*[1,2], *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

## ABSTRACT

Regular cameras and cell phones are able to capture limited luminosity. In terms of quality, most of the produced images by such devices are not similar to the real world. Various methods, which fall under the name of High Dynamic Range (HDR) Imaging, can be utilised to cope with this problem and produce an image with more details. However, most methods for generating an HDR image from Multi-Exposure images only focus on how to combine different exposures and do not consider the choice the best details of each image. By convers, in this research it is strived to detect the most visible areas of each image with the help of image segmentation. Two methods of producing the Ground Truth are considered, as manual and Otsu thresholding, and two similar neural networks are used to train segment these areas. Finally, it is shown that the neural network is able to segment the visible parts of pictures acceptably.

***Index Terms—*** Image Segmentation, Otsu Threshold, Multi-Exposure, High Dynamic Range, Deep Learning.

## 1. INTRODUCTION

Natural scenes have a vast luminosity; however, regular cameras are capable of capturing a limited dynamic range of that luminance. Therefore, the generated image has regions with High- (overly bright) and Low-Exposure (too dark), and the detail is not well visible. These types of pictures are called Low Dynamic Range (LDR) images.

The first solution to this problem is to utilise cameras with special sensors, which can obtain more luminance than regular cameras and produce images with more details and more similar to the real-world [1-7]. However, due to the high cost of such equipment, it is not affordable and usable for regular users.

Another solution for this issue is using software development methods known as High Dynamic Range (HDR) imaging. Various algorithms have been proposed recently, and the existing techniques can be divided into HDR imaging with Single-Exposure and Multi-Exposure methods. In the Single-Exposure, various techniques can produce an HDR image starting from a single LDR image. However, these methods are not satisfying since the detail cannot be restored goodly. In [8], the authors proposed an algorithm to generate an HDR image from an LDR image. Still, their method was affected by two problems: the inability to reconstruct details of dark and overly saturated areas. More precisely, this algorithm was not able to retrieve the details in the excessively saturated regions. Therefore, [9] proposed to first merge input images with different exposures and afterwards feed the wavelet coefficient of the merged image to the network to produce more details in a shorter time. Fortunately, unlike the Single-Exposure methods, Multi-Exposure ones are more effective and can reconstruct more detail. Several LDR images are combined in such techniques and produce an HDR image. Although Multi-Exposure methods perform almost perfectly on static scenes, they can encounter problems such as ghosting in dynamic scenes due to moving objects. However, several algorithms have been proposed to solve this issue [10-15].

Additionally, deep learning has been a great help in computer vision in recent years. For instance, [8] used a deep neural network to produce an HDR image in the logarithmic domain. Also, [16] used deep learning to reconstruct the detail of an image with different row-wise exposure in the irradiance domain. The works [17,18], unlike other methods, used neural networks to produce several LDR images with different exposures from a single LDR image. Additionally, [11] first aligned images with the optical flow and eventually used deep learning to fuse the aligned images to produce an image with more details. In [10], two deep learning methods were used to align images and generate an HDR image. Neural networks with different scales of images were used in [19] to learn the relative relation between input images and their Ground Truth.

Image Segmentation is one of the tasks in computer vision whose objective is to simplify image analysis. This task is typically used to detect objects or better understand images, such as medical ones. Image segmentation can be utilised to extract the regions of images with more details. In [20], the authors analysed images in HSV colour space to segment pixels based on the value of Intensity or Hue. Additionally, other works proposed two methods for image segmentation based on luminance: histogram division [21] and clustering based on the Gaussian Mixture Models (GMM) of the histogram [22]. Furthermore,[23] proposed a method to find the optimal valley point based on the slope between the histogram value of each pixel and other neighbouring points and used that valley point to segment regions.

The main contributions of this paper are as follows: (i) we propose two methods to extract the best areas of images with more details; (ii) we compare the proposed methods to specify the best one.

## 2. PROPOSED METHOD

### 2.1 Producing ground truth

Most proposed algorithms in HDR imaging are concentrated on how to produce them, while less attention has been paid to extracting suitable features. In this research, the proposed method focuses on extracting the most suitable regions for HDR imaging. Indeed by finding the areas with more details, the HDR algorithm can produce an image free of overly saturated or dark parts. More specifically, an Image segmentation method is proposed to segment areas with the most detail. A neural network can then be utilised to extract the desired regions of input images, which will be discussed in future work. Additionally, two different methods, i.e. manual thresholding and Otsu segmentation, were used to produce the Ground Truth, which will be compared with each other.

In the manual technique, several experts investigated the best possible range of intensity in YCbCr colour space for extracting the areas with the most detail empirically. Eventually, an average of the scopes was calculated for each image. The selected ranges for image intensity with Low and High-Exposure are [120,255] and [0,200], respectively. Generally, the objective is to acquire areas with less darkness and saturation. Therefore, because most of the regions in Low-Exposure images are dark, we would like to extract the areas with the highest pixel values, which indicate the most visible ones. Conversely, because most pixels in High-Exposure images are saturated, the objective is to extract pixels with the lowest values. Certainly, by choosing pixel values in the luminance channel, some of the visible pixels with the lowest values cannot be selected. For example, although the grey area of the mountain in Fig. 1 is visible, it was not selected in the segmentation process.

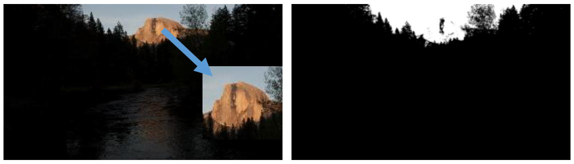

**Figure 1. The image on the left is the input image, and on the right is its Ground Truth produced by the manual method. The picture is taken from [24].**

The second method is called the Otsu technique, which calculates a threshold based on the intensities of images and segment pixels. More precisely, the pixels greater than the threshold are considered foreground (white), and those with lower values as background (black). The difference between these two methods is that the Otsu technique threshold is computed based on the histogram of each image. Whereas in the manual, all the pictures of each exposure have the same range. Moreover, in Otsu, all the pixels of Low-Exposure images greater than the threshold are considered the desired pixels, while the pixels lower than the threshold in High-Exposure pictures are desirable.

### 2.2 Neural network structure

Unfortunately, each image has various intensities, and it would be challenging to use non-machine learning methods to predict them. Moreover, it is a time-consuming task to extract a range for each image separately. Therefore, a neural network has been proposed in this research to learn how to extract the best area of each image based on the proposed ranges in the training stage.

Two similar U-Net-shaped networks were used for segmentation in this research, and each network is trying to learn how to map from each exposure to its Ground Truth. As can be seen in Fig. 2, the U-Net consists of 2 parts. In the first part, the subnetwork strives to extract features, and the second subnetwork tries to produce an output similar to the Ground Truth. The encoder section includes five blocks, and each block has two convolutional layers with ReLU function, DropOut, and MaxPool layers, respectively. Additionally, kernels of convolutional layers in each block are 16,32,64,128,256, respectively. Moreover, the decoder has four blocks, and each block consists of one transpose convolutional, concatenate, convolutional layer with ReLU activation, Drop Out, and another convolutional layer with ReLU, respectively. Furthermore, all convolutional and transpose convolutional layers used a kernel size of 3x3, and the last layer used a kernel size of 1x1.

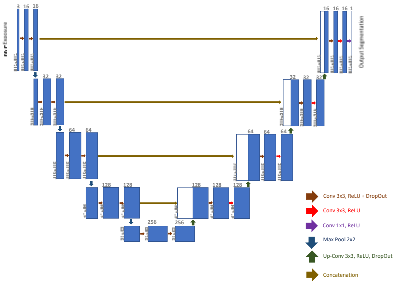

**Figure 2. Total scheme U-net Architecture, which was used in this experiment. The blue boxes denote feature maps; their number is on the top of each box, and their size is indicated on the lower left side of each box.**

### 2.3 Loss functions

The loss function is one of the essential components in deep learning. Thus, three loss functions will be used and compared to select the best loss function for segmenting the regions with the most detail. The used loss functions are as follows:

1. Binary Cross Entropy (BCE): One of the most common functions, which is used in most image segmentation research is the BCE loss function, and it can be represented as follows:

$$L_{BCE} = -\sum\left(y log\hat{y} + (1-y)log(1-\hat{y})\right) \qquad (1)$$

Where y and $\hat{y}$ represent Ground Truth and the network's output, respectively, and the sum is over all the pixels.

2. Focal Loss: This loss function is used for imbalance data and focuses on hard data:

$$L_{focal} = -\sum(\alpha.y.(1-\hat{y})^{\gamma}log(\hat{y}) + (1-\alpha)(1-y)(\hat{y})^{\gamma}\log(1-\hat{y})) \quad (2)$$

Where α and γ are hyper-parameters and, as a default, they are equal to 0.25 and 2.0, respectively.

3. Combo Loss (Dice Cross-Entropy): This loss function is also used for imbalanced data and is produced by a combination of Cross-Entropy and Dice loss functions. Eq (3) represents Dice loss, and Eq (4) is for Combo loss:

$$DL(y,\hat{y}) = 1 - \frac{2y\hat{y}+1}{y+\hat{y}+1} \quad (3)$$

The number one added to the numerator and the denominator avoids undefined errors, such as y=$\hat{y}$=0.

$$L_{DiceCE} = L_{Dice} + L_{BCE} \quad (4)$$

## 3. EXPERIMENT RESULTS

### 3.1 Dataset

Recently, a new dataset was collected for High Dynamic Range (HDR) Imaging Challenge called NTIRE 2021 [25]. 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 [26] 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. Moreover, all the images were already aligned and gamma corrected.

### 3.2 Evaluation metrics

Several evaluation parameters have been used in this research to evaluate the results and are discussed as follows:

1. Dice Index: This metric is region based and evaluates the similarity and the overlaps of two samples.

$$Dice\ (A,B) = 2\frac{|A \cap B|}{|A|+|B|} \quad (5)$$

2. Jaccard Index: This metric works similarly to Dice and calculates the similarity of two samples.

$$Jaccard\ (A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

3. Two other metrics are Sensitivity and Specificity, which calculate True Positive and True Negative pixels.

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

4. Area under Curve (AUC): this metric is commonly used in image segmentation algorithms.

$$AUC = 1 - \frac{1}{2}\left(\frac{FP}{FP+TN} + \frac{FN}{FN+TP}\right) \quad (9)$$

### 3.3 Ground truth generation

As the used dataset is not consisting of Ground Truths for segmentation, the first objective of this research is to produce Ground Truths that cover the most area of scenes. Thus, after frequent and visual studying of produced Ground Truths by both manual and Otsu techniques, it became evident that the manual method has more coverage than the latter one. For instance, as can be seen in Fig. 3, both approaches worked almost the same on images with Low-Exposure. However, the manual method succeeded in covering more areas in images with High-Exposure. Additionally, as can be seen in the last row, the total covered area by the manual method is larger than in the Otsu technique. Therefore, the produced Ground Truth from the manual method will be used for the rest of the research.

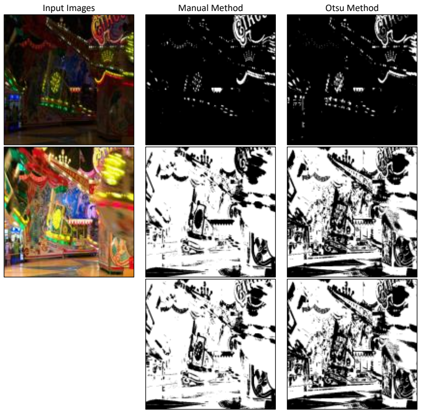

**Figure 3. Produced Ground Truth of both Manual and Otsu Methods. The first row is generated from the Low-Exposure image, the second one is obtained from the High-Exposure image, and the third row is a merged output of both rows.**

### 3.4 Other details

Additionally, the training process for each loss function was 50 epochs, which took around 200 minutes on NVIDIA DGX A100 and less than 2 minutes for testing, and the model was trained as parallel on 4 GPUs. Moreover, the number of images for the training set and the validation set was about 1300 and 200 images with a resolution of 512x512 and a batch size of 32, respectively. Furthermore, Adam optimiser with a learning rate of 0.001 was used. Finally, the neural network was implemented in Tensorflow (Keras) framework. During experiments, 3 input images with different exposures were used for image segmentation, in which, after obtaining the suitable areas of Low- and High-Exposure images, the remaining regions were extracted from the Medium-Exposure images. However, the acquired areas of the Medium-Exposure were not sensible because most of them were only a few pixels with no shapes. Thus, it was difficult for the network to segment them. Fig. 4 demonstrates an example of the extracted regions in the Medium-Exposure image.

### 3.5 Results

The predicted segmentation outputs by 3 different loss functions were compared quantitatively with their produced Ground Truth by manual technique. As can be seen in Table 1, which demonstrates the evaluation results of Low-Exposure Image Segmentation, different loss functions outperformed the others in different evaluation metrics. For instance, the Focal Loss function performed better than the others in Jaccard and Sensitivity evaluation metrics. Although they have equal values in the AUC evaluation metric, the Focal loss was better than Dice-BCE and BCE averagely. Additionally, Table 2 indicates that the Dice-BCE loss function worked better than the other two losses in Jaccard and Sensitivity evaluation metrics, but as a result, BCE was better on average. Therefore, it can be concluded that Focal loss function can segment better illumination in Low-Exposures and BCE in High-Exposures. Figs. 5 and 6 demonstrate produced outcomes by different loss functions for both images with Low- and High- Exposure. As can be seen, although all the outputs are almost identical visually and are difficult to distinguish differences between them, the quantitative results demonstrated that the output of Dice-BCE is not as well as the output of the other two. Moreover, Fig. 7 indicates more examples of losses.

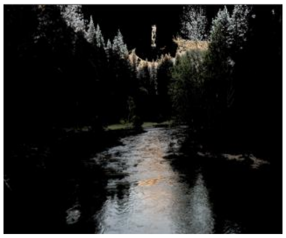

**Figure 4. An example of extracted areas from a Medium-Exposure image. The picture is taken from [24].**

**Table 1. Quantitative evaluation results of Low-Exposure Image Segmentation. M row determines metrics, which are specified as M1: Dice, M2: Jaccard, M3: Sensitivity, M4: Specificity, M5: AUC, and AVG is the average of the metrics.**

| **Loss functions** | M1 | M2 | M3 | M4 | M5 | AVG |
|---|---|---|---|---|---|---|
| BCE | 0.951 | 0.905 | 0.912 | **0.999** | 0.498 | 0.853 |
| Focal | 0.916 | **0.936** | **0.997** | 0.997 | 0.498 | **0.869** |
| Dice - BCE | **0.965** | 0.933 | 0.912 | **0.999** | 0.498 | 0.861 |

**Table 2. Quantitative evaluation results of High-Exposure Image Segmentation. M row is specified in Table 1.**

| **Loss functions** | M1 | M2 | M3 | M4 | M5 | AVG |
|---|---|---|---|---|---|---|
| BCE | **0.994** | 0.909 | 0.765 | 0.754 | **0.68** | **0.82** |
| Focal | 0.989 | 0.89 | 0.753 | **0.763** | 0.675 | 0.814 |
| Dice - BCE | 0.991 | **0.912** | **0.77** | 0.73 | 0.67 | 0.815 |

## 4. CONCLUSION AND FUTURE WORKS

As discussed in the proposed method section, Otsu and manual methods were used in this research, and in the manual technique, a range was computed empirically. Although experiments demonstrated that the empirical approach had better outcomes than Otsu, it has two cons. Firstly, failure to recognise dark visible areas, such as the mountain peak illustrated in Fig. 1. Secondly, if the segmentation process is performed with the manual technique, the calculated range needs to be applied to all images, and it is possible that the computed span is not suitable for some photos, and calculating a specific span for each picture is also a time-consuming task. Therefore, it is better to work on a new automatic technique to estimate these ranges for each image. In addition to working on a novel method for calculating an automatic range for each image in future work, it is feasible to use extracted regions from segmentation techniques in HDR imaging to produce an HDR image with more details. Additionally, this work can help reduce the complexity of networks for generating an HDR image.

In summary, two methods for segmenting visible regions were used in this research, and a manual technique that is an empirical approach was chosen after comparing them to produce the Ground Truth. Moreover, deep neural networks were used to learn to extract the regions with the help of produced Ground Truths in each exposure. Additionally, three different loss functions were utilised in this article, and the quantitative metrics demonstrated that the focal and BCE loss functions outperformed in Low-Exposure and High-Exposure images, respectively.

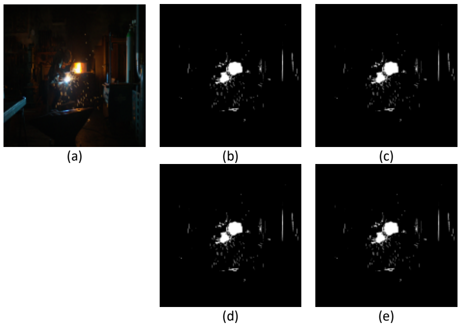

**Figure 5. Output results of other losses. (a) Low-Exposure input image, (b) Dice-BCE Output, (c) BCE Output, (d) Focal Output, (e) Ground Truth.**

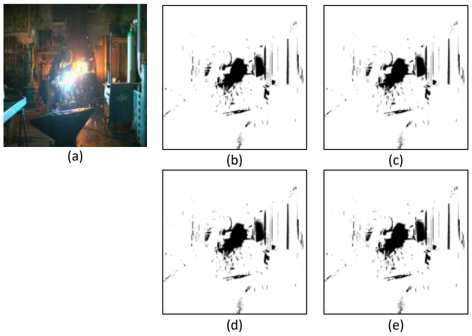

**Figure 6. Output results of different losses. (a) High-Exposure input image, (b) Dice-BCE Output, (c) BCE Output, (d) Focal Output, (e) Ground Truth.**

## 6. 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*, 2000.

[2] J. Tumblin and et al., "Why I want a gradient camera," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.

[3] M. McGuire and et al., "Optical Splitting Trees for High-Precision Monocular Imaging," *IEEE Computer Graphics and Applications*, vol. 27, pp. 23-42, 2007.

[4] M. D. Tocci and et al., "A Versatile HDR Video Production System," *ACM Trans. Graph*, vol. 30, 2011.

[5] S. Hajisharif and et al., "Adaptive dualISO HDR reconstruction," *EURASIP Journal on Image and Video Processing*, 2015.

[6] H. Zhao and et al., "Unbounded High Dynamic Range Photography Using a Modulo Camera," 2015.

[7] A. Serrano and et al., "Convolutional Sparse Coding for High Dynamic Range Imaging," 2016.

[8] G. Eilertsen and et al., "HDR Image Reconstruction from a Single Exposure Using Deep CNNs," *ACM Trans. Graph.*, vol. 36, p. 15, 2017.

[9] A. Omrani and et al., "High dynamic range image reconstruction using multi-exposure Wavelet HDRCNN," in *2020 International Conference on Machine Vision and Image Processing (MVIP)*, 2020.

[10] K. Green Rosh and et al., "Deep Multi-Stage Learning for HDR With Large Object Motions," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019.

[11] N. K. Kalantari and R. Ramamoorthi, "Deep High Dynamic Range Imaging of Dynamic Scenes," *ACM Trans. Graph.*, vol. 36, p. 12, 2017.

[12] S. Wu and et al., "Deep High Dynamic Range Imaging with Large Foreground Motions," in *Computer Vision – ECCV*, 2018.

[13] Q. Yan and et al., "Attention-Guided Network for Ghost-Free High Dynamic Range Imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] K. R. Prabhakar and et al., "A Fast, Scalable, and Reliable Deghosting Method for Extreme Exposure Fusion," in *2019 IEEE International Conference on Computational Photography (ICCP)*, 2019.

[15] K. R. Prabhakar and et al., "Towards Practical and Efficient High-Resolution HDR Deghosting with CNN," in *Computer Vision – ECCV 2020*, 2020.

[16] 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)*, 2017.

[17] S. Lee and et al., "Deep Chain HDRI: Reconstructing a High Dynamic Range Image from a Single Low Dynamic Range Image," *IEEE Access*, vol. 6, pp. 49913-49924, 2018.

[18] Y. Endo and et al., "Deep Reverse Tone Mapping," *ACM Trans. Graph.*, vol. 36, p. 10, 2017.

[19] Q. Yan and et al., "Multi-Scale Dense Networks for Deep High Dynamic Range Imaging," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[20] A. Vadivel and et al., "Segmentation Using Saturation Thresholding and Its Application in Content-Based Retrieval of Images," in *Campilho, A., Kamel, M. (eds) Image Analysis and Recognition. ICIAR 2004.*, 2004.

[21] Y. Kinoshita and H. Kiya, "Scene Segmentation-Based Luminance Adjustment for Multi-Exposure Image Fusion," *IEEE Transactions on Image Processing*, pp. 4101-4116, 2019.

[22] 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*, p. 22, 2018.

[23] B. D. Lee and M. H. Sunwoo, "HDR Image Reconstruction Using Segmented Image Learning," *IEEE Access*, vol. 9, pp. 142729-142742, 2021.

[24] M. D. Fairchild, The HDR photographic survey, 2007.

[25] E. Perez-Pellitero and et al., "NTIRE 2021 Challenge on High Dynamic Range Imaging: Dataset, Methods and Results," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 691-700, 2021.

[26] J. Froehlich, and et al., "Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays," in *In Proc. of SPIE Electronic Imaging*, 2014.