Marcel Santana Santos

Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2025

Marcel Santana Santos

Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração
Inteligência computacional

Orientadores
Tsang Ing Ren
Nima Khademi Kalantari

Recife
2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Santana Santos, Marcel.
Single Image HDR Reconstruction Using a CNN ith Masked Features and Perceptual Loss / Marcel Santana Santos. - Recife, 2020.
59 p. : il., tab.

Orientador(a): Tsang Ing Ren
Cooorientador(a): Nima Khademi Kalantari
Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2020.
Inclui referências, apêndices, anexos.

1. Alto alcance dinâmico. 2. Redes Neurais Convolucionais. 3. Função de Perda Perceptual. I. Ing Ren, Tsang. (Orientação). II. Khademi Kalantari, Nima. (Coorientação). IV. Título.

000 CDD (22.ed.)

Marcel Santana Santos

“Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 31/08/2020.

BANCA EXAMINADORA

____

Prof. Dr. Sílvio de Barros Melo
Centro de Informática / UFPE

____

Prof. Dr. Manuel Menezes de Oliveira Neto
Instituto de Informática / UFRGS

____

Prof. Dr. Tsang Ing Ren
Centro de Informática / UFPE
(Orientador)

ACKNOWLEDGEMENTS

I thank my mother, father, and grandmother for doing their best to assist me with anything necessary for my education. I thanks especially my advisor, professor Tsang, for giving me the necessary support for developing this work. Besides that, I thank my co-advisor professor Nima Kalantari for the valuable discussions and for giving me the opportunity to work with him at Texas A&M University. Professor Kalantari has been a source of inspiration to me since my last year of graduation and I am very thankful for be working with him. Also, I would like to thank Thais Aquino, my fiancé, for all positive influence and the patience in the time I have been absent doing research. I am very thankful to my partners from Vektore, in special Rogerio, Sahadia, and R. Figueiredo for all the support they gave me for my research. Finally, I thanks my collegues from Aggieland Graphics Lab for the valuable discussions. My research was funded by the Brazilian agency CNPQ grant 161268/2018-8.

What I cannot create I do not understand (Richard Feynman).

RESUMO

Câmeras digitais convencionais não são capazes de capturar completamente o alcance de iluminação das cenas (expressa por uma grandeza conhecida por luminância). Consequentemente, as imagens produzidas por estes dispositivos geralmente apresentam regiões com saturação e, portanto, informações da cena são perdidas. Métodos tradicionais para reconstrução desse intervalo perdido pela captura não são capazes de reconstruir as texturas e detalhes das cenas, produzindo resultados com artefatos nas regiões saturadas. No presente trabalho, foram investigados métodos baseados em redes neurais convolucionais para reconstrução de imagens com alto alcance dinâmico (HDR) a partir de apenas uma imagem capturada com câmeras convencionais (LDR). Essas imagens HDR são capazes de expressar com fidelidade os detalhes das cenas e se aproximam do que o sistema visual humano é capaz de capturar. O método proposto é capaz de reconstruir as regiões saturadas das imagens de entrada com um alto grau realismo. Para alcançarmos este resultado, diversas contribuições foram realizadas. Primeiramente, os métodos baseados em redes convolucionais em geral aplicam o mesmo conjunto de filtros convolucionais nas regiões saturadas e não saturadas das imagens. No entanto, as regiões saturadas não contém informação válida, o que causa ambiguidade durante o treinamento causando diversos artefatos no resultado final. Para resolver este problema, foi proposto um mecanismo (apelidado feature masking) para reduzir a contribuição das regiões saturadas no cálculo das convoluções. Além disso as funções de erro perceptual (comumente utilizadas em problemas de síntese de imagens) para o treinamento da rede foram revisitadas e adaptadas para o problema de reconstrução de imagens HDR. Como resultado, o método proposto é capaz de produzir texturas realísticas e com um alto grau de fidelidade a cena original. Além disso, como as bases de dados de treinamento para o presente problema ainda são limitadas, foi proposto realizar o treinamento do método em duas etapas. Especificamente, o método é inicialmente treinado em um número grande de imagens em uma tarefa auxiliar (image inpainting, neste caso) e então refinado para a tarefa de reconstrução de imagens HDR. Por fim, como a maioria das imagens de treinamento contém regiões simples de serem reconstruídas, foi proposto uma estratégia para selecionar regiões difíceis para serem utilizadas durante a etapa de refinamento da rede neural. Essa estratégia simples é capaz de aumentar a robustez e reduzir o tempo de treinamento do método. Diversos experimentos foram conduzidos em uma grande variedade de cenários para demonstrar visualmente e numericamente que o método proposto é capaz de produzir imagens HDR com alto grau de realismo e melhor que os métodos estado-da-arte. Um artigo decorrente do presente trabalho foi aceito na conferência ACM SIGGRAPH 2020.

Palavras-chaves: Alto alcance dinâmico, Redes Neurais Convolucionais, Função de Perda Perceptual.

ABSTRACT

Digital cameras can only capture a limited range of real-world scenes' luminance, producing images with saturated pixels. Existing single image high dynamic range (HDR) reconstruction methods attempt to expand the range of luminance, but are not able to hallucinate plausible textures, producing results with artifacts in the saturated areas. In this thesis, we present a novel learning-based approach to reconstruct an HDR image by recovering the saturated pixels of an input LDR image in a visually pleasing way. Previous deep learning-based methods apply the same convolutional filters on well-exposed and saturated pixels, creating ambiguity during training and leading to checkerboard and halo artifacts. To overcome this problem, we propose a feature masking mechanism that reduces the contribution of the features from the saturated areas. Moreover, we adapt the VGG-based perceptual loss function to our application to be able to synthesize visually pleasing textures. Since the number of HDR images for training is limited, we propose to train our system in two stages. Specifically, we first train our system on a large number of images for image inpainting task and then fine-tune it on HDR reconstruction. Since most of the HDR examples contain smooth regions that are simple to reconstruct, we propose a sampling strategy to select challenging training patches during the HDR fine-tuning stage. We demonstrate through experimental results that our approach can reconstruct visually pleasing HDR results, better than the current state of the art on a wide range of scenes.

Key-words: High dynamic range imaging, convolutional neural network, feature masking, perceptual loss

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\hat{H}$	Final HDR image
H	Linear ground truth image
T	Input LDR image
$\hat{Y}$	Network output in the logarithmic domain
M	Saturation (soft) mask
ϕ_l	Activation function in layer l
$\mathcal{T}(H)$	μ-law range compressor function
$G_l(X)$	Gram matrix in layer l

CONTENTS

Chapter 1

INTRODUCTION

The illumination of real-world scenes is high dynamic range, however standard digital cameras sensors can only capture a limited range of luminance. Therefore, these cameras typically produce images with under/over-exposed areas. As an example, when attempting to capture an object in a dark indoor environment in front a bright windows, one has to choose between properly expose the bright background or the object in the foreground. Properly expose the background causes the information to be lost in the dark (underexposed) areas of the foreground (see Figure 1 left), while choosing to properly expose the object in the foreground will cause the loss of information of the saturated (overexposed) background (see Figure 1 right). On the other hand, it is usually not a problem for the human eye to simultaneously register both background and foreground due to the wider dynamic range of the human visual system (HVS) when compared to the conventional cameras, as shown in Figure 2. This difference in the dynamic range of the standard cameras compared to the HVS motivates several techniques that can produce high dynamic range (HDR) images.

A large number of approaches propose to generate a HDR image by combining a set of low dynamic range images (LDR) of the scene at different exposures (DEBEVEC; MALIK, 1997). For long exposures images, the details in dark areas are captured while information in bright areas vanishes due to sensor saturation. For short exposures images,

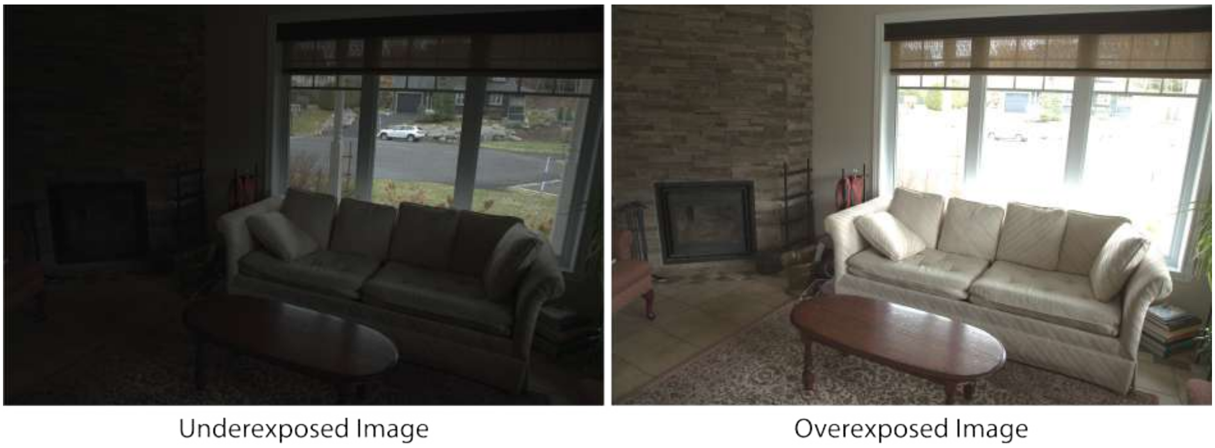

Figure 1 – Standard cameras often cannot capture the high dynamic range of the scene. The user has to choose between properly expose the background, which causes the information to be lost in the foreground areas (left), or the foreground that leads the information to be lost in the background (right).

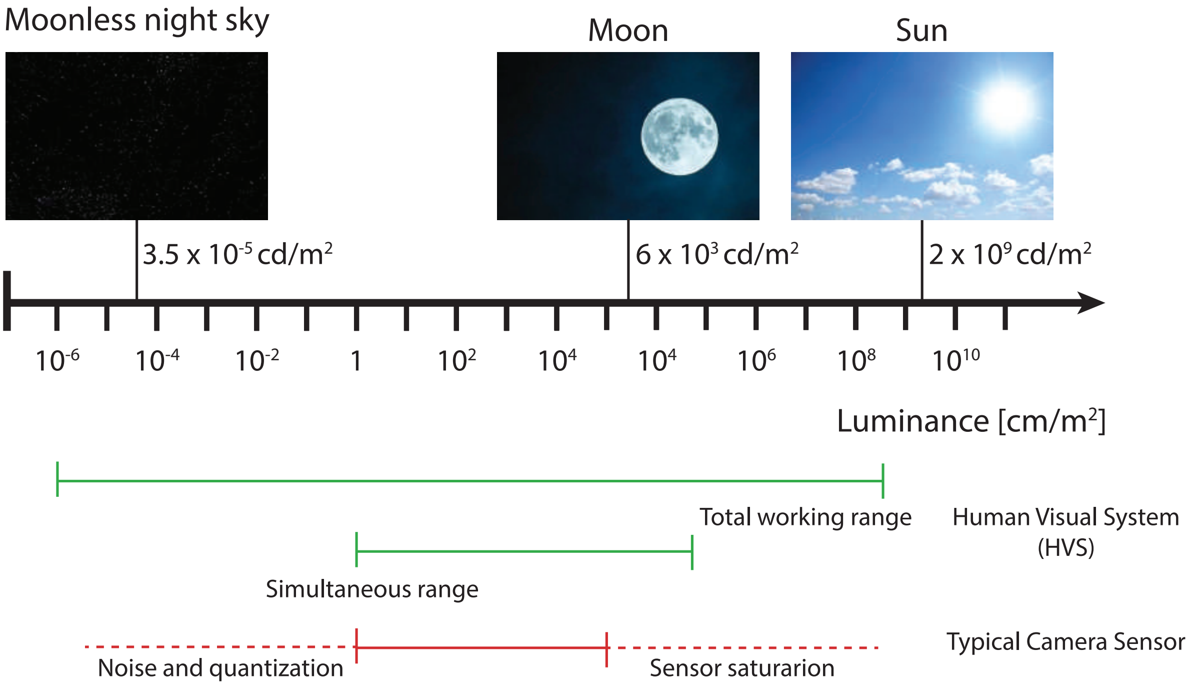

Figure 2 – Comparison of the dynamic ranges of human visual system and typical cameras sensors. The simultaneous dynamic range can be defined as the range which the visual system can detect objects while being in a state of full adaptation. This sensitivity is much lower than the total working range to which the visual system can adapt for several reasons.

the bright areas are properly registered while darker parts are lost in quantization and noise. Combining these different exposures means that both dark and bright image regions, which are outside the range of a conventional sensor, will be represented and therefore more information will be captured. However, these methods either have to handle the scene motion (KALANTARI; RAMAMOORTHI, 2017; WU et al., 2018; KANG et al., 2003; SEN et al., 2012; HU et al., 2013; OH et al., 2014) or require specialized bulky and expensive optical systems (MCGUIRE et al., 2007; TOCCI et al., 2011). Single image dynamic range expansion approaches avoid these limitations by reconstructing an HDR image using one image. These approaches can work with images captured with any standard camera or even recover the full dynamic range of legacy LDR content. As a result, they have attracted considerable attention in recent years.

Several existing methods for single image dynamic range expansion extrapolate the light intensity using hand-crafted features and rules (BANTERLE et al., 2006; REMPEL et al., 2007; BIST et al., 2017), but are not able to properly recover the brightness of saturated areas as they do not utilize context. On the other hand, recent deep learning approaches (ENDO; KANAMORI; MITANI, 2017; LEE; AN; KANG, 2018a; EILERTSEN et al., 2017) systematically utilize contextual information using convolutional neural networks (CNNs) with large receptive fields. However, these methods usually produce results with blurriness, checkerboard, and halo artifacts in saturated areas.

In this dissertation, we present a novel learning-based technique to reconstruct an HDR image by recovering the missing information in the saturated areas of an LDR image. We design our approach based on two main observations. First, applying the same convolutional filters on well-exposed and saturated pixels, as done in previous approaches, results in ambiguity during training and leads to checkerboard and halo artifacts. Second, using simple pixel-wise loss functions, utilized by most existing approaches, the network is unable to hallucinate details in the saturated areas, producing blurry results. To address these limitations, we propose a feature masking mechanism that reduces the contribution of features generated from the saturated content by multiplying them to a soft mask. With this simple strategy, we are able to avoid checkerboard and halo artifacts as the network only relies on the valid information of the input image to produce the HDR image. Moreover, inspired by image inpainting approaches, we leverage the VGG-based perceptual loss function, introduced by Gatys, Ecker e Bethge (2016), and adapt it to the HDR reconstruction task. By minimizing our proposed perceptual loss function during training, the network can synthesize visually realistic textures in the saturated areas.

Since a large number of HDR images, required for training a deep neural network, are currently not available, we perform the training in two stages. In the first stage, we train our system on a large set of images for the inpainting task. During this process, the network leverages a large number of training samples to learn an internal representation that is suitable for synthesizing visually realistic texture in the incomplete regions. In the next step, we fine-tune this network on the HDR reconstruction task using a set of simulated LDR and their corresponding ground truth HDR images. Since most of the HDR examples contain smooth and textureless regions that are simple to reconstruct, we propose a simple method to identify the textured patches and only use them for fine-tuning.

Our approach can reconstruct regions with high luminance and hallucinate textures in the saturated areas, as shown in Figure 3. We also demonstrate trough several examples that our approach can produce better results and is faster than the state-of-the-art methods both on simulated images and on images taken with real-world cameras.

1.1 OVERVIEW OF THIS DISSERTATION

1.1.1 Objective

The main goal of this dissertation is to develop a method for single-image HDR reconstruction using convolutional neural networks. Given a single overexposed LDR image, our method must be able to reconstruct an HDR image by synthesizing visually realistic textures and details in the saturated areas.

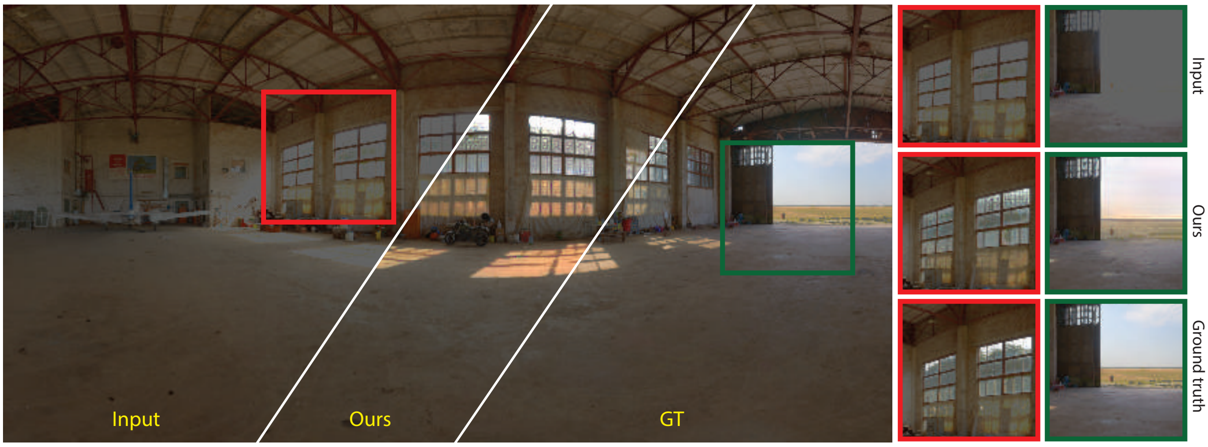

Figure 3 – We propose a novel deep learning system for single image HDR reconstruction by synthesizing visually pleasing details in the saturated areas. We introduce a new feature masking approach that reduces the contribution of the features computed on the saturated areas, to mitigate halo and checkerboard artifacts. To synthesize visually pleasing textures in the saturated regions, we adapt the VGG-based perceptual loss function to the HDR reconstruction application. Furthermore, to effectively train our network on limited HDR training data, we propose to pre-train the network on inpainting task. Our method can reconstruct regions with high luminance, such as the bright highlights of the windows (red inset), and generate visually pleasing textures (green insert). The images have been gamma corrected for display purposes.

The specific objectives of this research are:

- To purpose an approach to mitigate halo and checkerboard artifacts present in the current state-of-the-art methods.
- To purpose a loss function capable of synthesizing visually pleasing textures in the saturated regions.
- To be able to effectively train our network on limited HDR training data.
- To be able to generalize to a wide range of scenarios such as indoor, outdoor, day and night.
- To be able to generalize to several real-world cameras.

1.1.2 Contributions

To achieve the previously stated objectives, we made the following contributions:

1. We propose a feature masking mechanism to avoid relying on the invalid information in the saturated regions. This masking approach significantly reduces the artifacts and improves the quality of the final results (Figure 16).

2. We adapt the VGG-based perceptual loss function to the HDR reconstruction task. Compared to pixel-wise loss functions, our loss can better reconstruct sharp textures in the saturated regions (Figure 18).

3. We propose to pre-train the network on inpainting before fine-tuning it on HDR generation. We demonstrate that the pre-training stage is essential for synthesizing visually pleasing textures in the saturated areas (Figure 17).

4. We propose a simple strategy for identifying the textured HDR areas to improve the performance of training. This strategy improves the network ability to reconstruct sharp details (Figure 17).

The supplementary materials and source code are available at the project website: <https://marcelsan.github.io/SIGGRAPH2020/>.

1.1.3 Publications

The work presented in this dissertation in built on the following publication:

Marcel Santana Santos, Tsang Ing Ren and Nima Khademi Kalantari. Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss. *ACM Transactions on Graphics*, 39, 4, Article 80 (July 2020).

1.1.4 Organization of the document

The present document is organized as follows:

Chapter 2 discusses the basic concepts of High Dynamic Range imaging and methods for HDR image synthesis. It also discusses the key terminology that will be used throughout the dissertation.

Chapter 3 discusses the proposed approach for single image HDR Reconstruction, including the feature masking strategy, our loss function and the two-stage training.

Chapter 4 discuss some implementation details, shows the result of our method and compares them to other state-of-the-art methods. Also, we discuss several ablations we performed to our system and show some failure cases of our method.

Chapter 5 finally presents our conclusions and list some potential directions for future work that we think are promising.

Chapter 2

BACKGROUND AND RELATED WORK

High dynamic range imaging enables to capture a wide range of the illumination of a scene and therefore produces images that closely resemble what humans can see. Therefore, HDR imaging is useful for improving the viewing experience on HDR displays or by means of tone-mapping (OU et al., 2020). Furthermore, an HDR image represents a photometric measurement of the physical lighting incident on the camera sensor. As such, HDR panoramas can be used as a light source for synthesizing photo realistic images by using image-based lighting (IBL) techniques (DEBEVEC, 2008). These techniques can also be useful in the visual effects (VFX) movies industry to insert computer graphics generated content in a filmed shot (DEBEVEC et al., 2000). Finally, other fields such as medical imaging, simulations, virtual reality and surveillance, to name a few, can also benefit from the more accurate lighting measurement from HDR images. Given the aforementioned applications of HDR imaging, methods for capturing high dynamic range images have been subject of extensive research for decades. In particular, most recently, the success of deep learning on various image processing and synthesizes tasks (e.g. image inpainting, style transfer, image colorization) has prompted research in developing learning-based approaches for HDR image synthesis.

In this chapter, we introduce the relevant literature on high dynamic range image synthesis as well as the key terminology that will be used throughout this dissertation. We start by briefly discussing the methods that require multiple exposures (Section 2.1) and special hardware (Section 2.2) for producing high dynamic range images. We then focus on single-image methods in Section 2.3, which we categorize in non-learning and learning-based methods.

2.1 MULTI-EXPOSURE METHODS

The most common method for HDR image synthesis is merging multiple low dynamic range (LDR) images taken from the same scene at different exposures (DEBEVEC; MALIK, 1997; MADDEN, 1993). This process is generally composed of two distinct steps. First, the camera response function (CRF) needs to be estimated and inverted to obtain pixel values that linearly corresponds to the captured luminances (DEBEVEC; MALIK, 1997; GROSSBERG; NAYAR, 2003a). Notice that, in modern cameras systems, it is possible to easily access these linear measurements stored in an increased precision (usually, 12-14 bits) in a RAW format. Secondly, the set of images with different exposures in the linear domain needs to be merged to produce the final HDR image. Straightforward methods for HDR fusion include picking one single exposure for each pixel (MADDEN, 1993) or using a simple triangular filter (DEBEVEC; MALIK, 1997). Other more robust methods consider noise models (MITSUNAGA; NAYAR, 1999; TSIN; RAMESH; KANADE, 2001) or even the camera pipeline (HAJSHARIF; KRONANDER; UNGER, 2014; HEIDE et al., 2014; KRONANDER et al., 2013) when fusing the multi-exposure images. In general, these methods produce high quality results for tripod mounted cameras and static scenes, however dynamic scenes or hand-held camera poses a challenging to them as robust alignment is needed to avoid ghosting artifacts.

To account for small amounts of camera shake or motion, the LDR sources may be globally registered (WARD, 2003; TOMASZEWSKA; MANTIUK, 2007). The static pixels will have the same color across the multi-exposure images stack and then can be merged into an HDR image as usual. If a pixel is moving, these methods detect it and reject it. Existing approaches have different ways of rejecting the ghosting regions. While several methods propose specific formulations for detecting these regions (JACOBS; LOSCOS; WARD, 2008; JINNO; OKUDA, 2008; GALLO et al., 2009; MIN; PARK; CHANG, 2009; WU et al., 2010), other algorithms do not require the explicit identification of the ghosted pixels at all (KHAN; AKYUZ; REINHARD, 2006; EDEN; UYTTENDAELE; SZELISKI, 2006; HEO et al., 2010).

On the other hand, for dynamic scenes with large motions, the problem is more challenging demanding registration on the local level. These approaches try to align the different exposures before merging them into the final HDR image. Although the problem of image alignment has been subject of extensive study in the computer vision community for decades, its application for the HDR imaging is far from trivial. Here, the input images have different exposures and therefore violate the color constancy assumption. Even if we use the inverse of camera response function to map the pixels to the linear domain, these pixels will have regions that are too dark or bright and should therefore not be considered for alignment. The simpler approaches to align the LDR images solve for a simple transformation (such as translation or homography) that accounts for camera motion between exposures (LI et al., 2017; AKYÜZ, 2011; YAO, 2011). More sophisticated alignment methods are based on optical flow (HAFNER; DEMETZ; WEICKERT, 2014; ZIMMER; BRUHN; WEICKERT, 2011) or patch-based (SEN et al., 2012; HU et al., 2013) approaches.

It is also important to mention, a few recent methods have shown considerable improvements in dynamic scenes by using learning-based methods. These methods are in essence extensions of the optical flow algorithm which use learned-based approaches to improve their results. For instance, Kalantari and Ramamoorthi (2017) address the multi-exposure fusion by using a convolutional neural network. This CNN-based fusion method improves their results as it can correct the misalignment caused by the optical flow. Yan et al. (2019) replace the optical flow by an attention neural network that excludes misaligned regions of the LDR sources. These images are then merged by using a fusion neural network. Chaudhari et al. (2019) also purpose to use a CNN for fusing the different exposures. However, they take as input RAW measurements to avoid common error propagation effects such as demosaicing artifacts, color distortions, or image alignment errors that are commonly amplified by the HDR merging step. While these methods present high-quality HDR results, they are still highly dependent on the number of exposures.

For a complete discussion, survey and categorization on multi-exposure fusion methods, we refer to the state-of-the-art report by Tursun et al. (2015).

2.2 SPECIFIC CAMERA HARDWARE

Another class of techniques for HDR image synthesis leverage specific optical systems or sensors for capturing HDR images in a single shot. For example, a few methods propose cameras with multiple sensors (MCGUIRE et al., 2007; TOCCI et al., 2011). The beam-splitter is placed behind the lens in the camera body and splits the light onto each sensor. This enables capturing a wider dynamic range than conventional cameras as each sensor will capture different exposures by restricting the light in each of them using different ND filters. Several approaches propose to reconstruct HDR images from coded per-pixel exposure (SERRANO et al., 2016; HEIDE et al., 2014; HAJISHARIF; KRONANDER; UNGER, 2015) or modulus images (ZHAO et al., 2015). These approaches can produce high-quality results even on dynamic scenes as they capture the entire image in a single shot. Unfortunately, they demand cameras with specific hardware that are often bulky and expensive and, therefore, are not available to the general public.

2.3 SINGLE IMAGE METHODS

Single image methods aim to reconstruct HDR images without requiring any special equipment or capturing techniques, nor multiple exposures. Therefore, the methods in this category can be easily applied to images or videos obtained from various sources such as standard cameras or the Internet. This flexibility makes the single-image methods more compelling as they enable the LDR content to be used in HDR applications such as the ones we mentioned at the beginning of this chapter (image-based lighting, HDR displays, tone-mapping etc.), see Figure 4 for a reference.

The problem of single image HDR reconstruction, also known as inverse tone-mapping (BANTERLE et al., 2006), has been extensively studied in the last couple of decades. However, this problem remains a major challenge as it requires recovering the details from regions with missing content. In this section, we discuss the existing techniques of single image HDR reconstruction by classifying them into two categories of non-learning and learning methods.

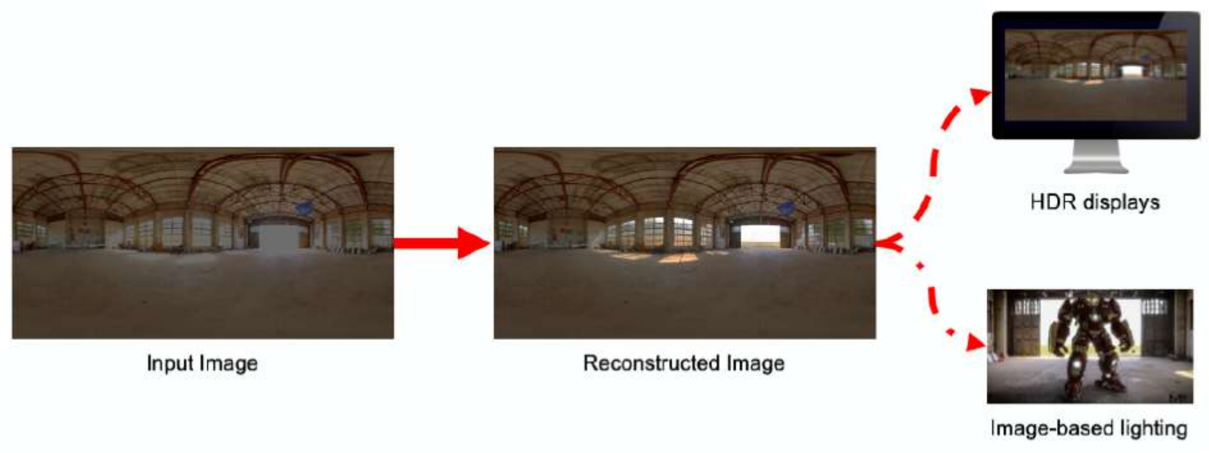

Figure 4 – Single image HDR reconstruction methods allow the LDR content to be enhanced to be used in HDR applications such as HDR displays and image-based lighting.

2.3.1 Non-learning Methods

Several approaches propose to perform inverse tone-mapping using global operators. Landis (2002) applies a linear or exponential function to the pixels of the LDR image above a certain threshold. Bist et al. (2017) approximates tone expansion by a gamma function. They use the characteristics of the human visual system to design the gamma curve. Luzardo et al. (2018) improve the brightness of the result by utilizing an operator based on the mid-level mapping.

A number of techniques propose to handle this application through local heuristics. Banterle et al. (2006) use median-cut (DEBEVEC, 2005) to find areas with high luminance. They then generate an expand-map to extend the range of luminance in these areas, using an inverse operator. Rempel et al. (2007) also utilize an expand-map but use a Gaussian filter followed by an edge-stopping function to enhance the brightness of saturated areas. Kovaleski and Oliveira (2014) extend the approach by Rempel et al. (2007) using a cross bilateral filter. These approaches simply extrapolate the light intensity by using heuristics and, thus, often fail to recover saturated highlights, introducing unnatural artifacts.

A few approaches propose to handle this application by incorporating user interactions in their system. For instance, Didyk et al. (2008) enhance bright luminous objects in video sequences by using a semi-automatic classifier to classify saturated regions as lights, reflections, or diffuse surfaces. Wang et al. (2007) recover the textures in the saturated areas by transferring details from the user-selected regions. Their approach demands user interactions that take several minutes, even for an expert user. In contrast to these methods, we propose a learning-based approach to systematically reconstruct HDR images from a wide range of different scenes, instead of relying on heuristics strategies and user inputs.

2.3.2 Learning-based Methods

In recent years, Convolutional Neural Networks (CNNs) (LECUN; BENGIO; HINTON, 2015) have been successful in several applications, achieving state of the art performance in a broad range of tasks including image classification (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), speech processing (OORD et al., 2016), text processing (DEVLIN et al., 2018; RADFORD et al., 2019) and many others. A CNN learns the weights of several filter kernels in each layer of a deep neural network. Each represent a different abstraction level. This means that the firsts layers learn local patterns while the next layers combine these local representations into local objects which are finally combined into high-level concepts (such as cat, dog, car etc.) by the last convolutional layers. By convolving the image with the learned filters, features can be extracted at different spatial locations using the same kernel. This makes neural networks for image processing and synthesis more tractable by making the connectivity of neurons between two adjacent layers sparse reducing therefore the number of parameters of the neural network.

Motivated by the recent development of CNNs in several image processing tasks such as image inpainting (LIU et al., 2018; YU et al., 2019), colorization (IIZUKA; SIMO-SERRA; ISHIKAWA, 2016; ZHANG; ISOLA; EFROS, 2016), super-resolution (TIAN et al., 2020; DONG et al., 2014) style transfer (JOHNSON; ALAHI; FEI-FEI, 2016) and image synthesis (CHEN; KOLTUN, 2017), several approaches have proposed to tackle the single image HDR reconstruction problem using deep convolutional neural networks. Zhang e Lalonde (2017) pioneered introducing CNNs to predict HDR panoramas from a LDR image with the purpose of using these panoramas to Image-based lighting (IBL). Endo, Kanamori e Mitani (2017) use an auto-encoder (HINTON; SALAKHUTDINOV, 2006) to generate a set of LDR images with different exposures, from a single input LDR image. These multi-exposure images are then combined to reconstruct the final HDR image. Lee, An e Kang (2018a) chain a set of CNNs to sequentially generate the bracketed LDR images. Later, they propose (LEE; AN; KANG, 2018b) to handle this application through a recursive conditional generative adversarial network (GAN) (GOODFELLOW et al., 2014) combined with a pixel-wise l_1 loss.

In contrast to these approaches, a few methods (EILERTSEN et al., 2017; YANG et al., 2018; MARNERIDES et al., 2018) directly reconstruct the HDR image without generating bracketed images. Eilertsen et al. (2017) use a network with U-Net architecture to predict the values of the saturated areas, whereas linear non-saturated areas are obtained from the input. Marnerides et al. (2018) present a novel dedicated architecture for end-to-end image expansion. Yang et al. (2018) reconstruct HDR image for image correction application. They train a network for HDR reconstruction to recover the missing details from the input LDR image, and then a second network transfers these details back to the LDR domain.

While these approaches produce state-of-the-art results, their synthesized images often contains halo and checkerboard artifacts and lacks textures in the saturated areas. This is mainly because of using standard convolutional layers and pixel-wise loss functions. Note that, several recent methods (LEE; AN; KANG, 2018b; XU et al., 2019; NING et al., 2018; KIM; OH; KIM, 2019) use adversarial loss instead of pixel-wise loss functions, but they still do not demonstrate results with high-quality hallucinated textures. This is potentially because the problem of HDR reconstruction is constrained in the sense that the synthesized content should properly fit the input image using a soft mask. Unfortunately, GANs are known to have difficulty handling these scenarios and manipulating existing images with GANs is challenging as the synthesized content does not usually fit the original image (BAU et al., 2019). In contrast, we propose a feature masking strategy and a more constrained VGG-based perceptual loss to effectively train our network and produce results with visually pleasing textures.

Chapter 3

PROPOSED APPROACH

3.1 INTRODUCTION

Our goal is to reconstruct an HDR image from a single LDR image by recovering the missing information in the saturated highlights. We achieve this using a convolutional neural network (CNN) that takes an LDR image as the input and estimates the missing HDR information in the saturated regions.

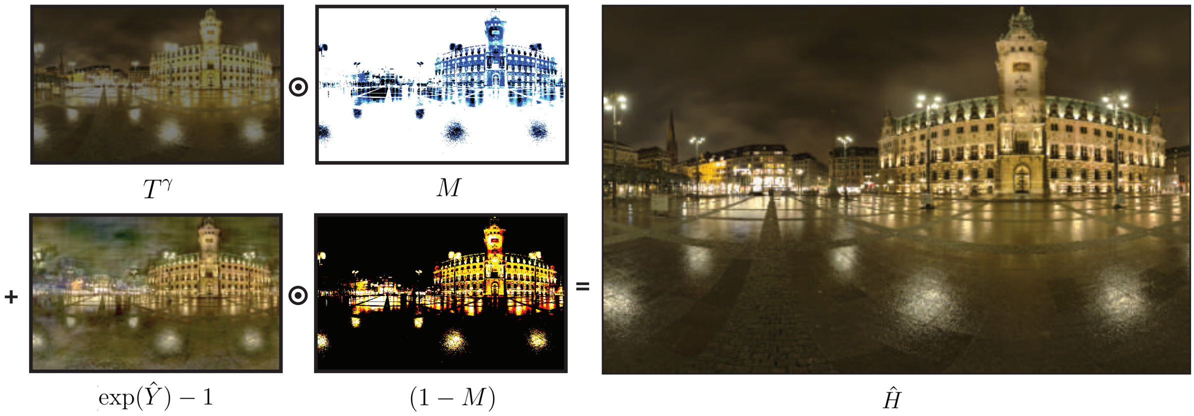

Figure 5 – Components of Equation 3.1 to compute the reconstructed HDR image $\hat{H}$. Here, T^γ is the input image in the linear domain, $\hat{Y}$ is the network output in the logarithmic domain and M is a soft mask that defines how well-exposed each pixel is.

To compute the final HDR image, we combine the well-exposed content of the input image and the output of the network in the saturated areas. Formally, we reconstruct the final HDR image $\hat{H}$, using the blending equation (see Figure 5) as follows:

$$\hat{H} = M \odot T^\gamma + (1-M) \odot [\exp(\hat{Y}) - 1], \tag{3.1}$$

where the $\gamma = 2.0$ is used to transform the input image to the linear domain, and $\odot$ denotes element-wise multiplication. Here, T is the input LDR image in the range $[0,1]$, $\hat{Y}$ is the network output in the logarithmic domain (Section 3.3), and M is a soft mask with values in the range $[0,1]$ that defines how well-exposed each pixel is. We obtain this mask by applying the function $\beta(\cdot)$ (see Figure 6) to the input image, i.e., $M = \beta(T)$.

In the following sections, we discuss our proposed feature masking approach, loss function, as well as the training process.

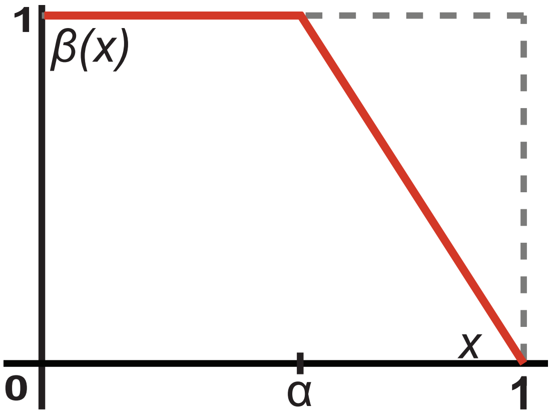

Figure 6 – We use this function to measure how well-exposed a pixel is. The value 1 indicates that the pixel is well-exposed, while 0 is assigned to the pixels that are fully saturated. In our implementation, we set the threshold $\alpha = 0.96$.

3.2 FEATURE MASKING

Standard convolutional layers apply the same filter to the entire image to extract a set of features. This is reasonable for a wide range of applications, such as image super-resolution (DONG et al., 2015), style transfer (GATYS; ECKER; BETHGE, 2016), and image colorization (ZHANG; ISOLA; EFROS, 2016), where the entire image contains valid information. However, in our problem, the input LDR image contains invalid information in the saturated areas. Since meaningful features cannot be extracted from the saturated contents, naïve application of standard convolution introduces ambiguity during training and leads to visible artifacts (Figure 16).

We address this problem by proposing a feature masking mechanism (Figure 7) that reduces the magnitude of the features generated from the invalid content (saturated areas). We do this by multiplying the feature maps in each layer by a soft mask, as follows:

$$Z_l = X_l \odot M_l, \tag{3.2}$$

where $X_l \in \mathbb{R}^{H \times W \times C}$ is the feature map of layer l with height H, width W, and C channels. $M_l \in [0,1]^{H \times W \times C}$ is the mask for layer l and has values in the range $[0,1]$. The value of one indicates that the features are computed from valid input pixels, while zero is assigned to the features that are computed from invalid pixels. Here, $l = 1$ refers to the input layer and, thus, $X_{l=1}$ is the input LDR image. Similarly, $M_{l=1}$ is the input mask $M = \beta(T)$. Note that, since our masks are soft, weak signals in the saturated areas are not discarded using this strategy. In fact, by suppressing the invalid pixels, these weak signals can propagate through the network more effectively.

Once the features of the current layer l are masked, the features in the next layer X_{l+1} are computed as usual:

$$X_{l+1} = \phi_l(W_l * Z_l + b_l), \tag{3.3}$$

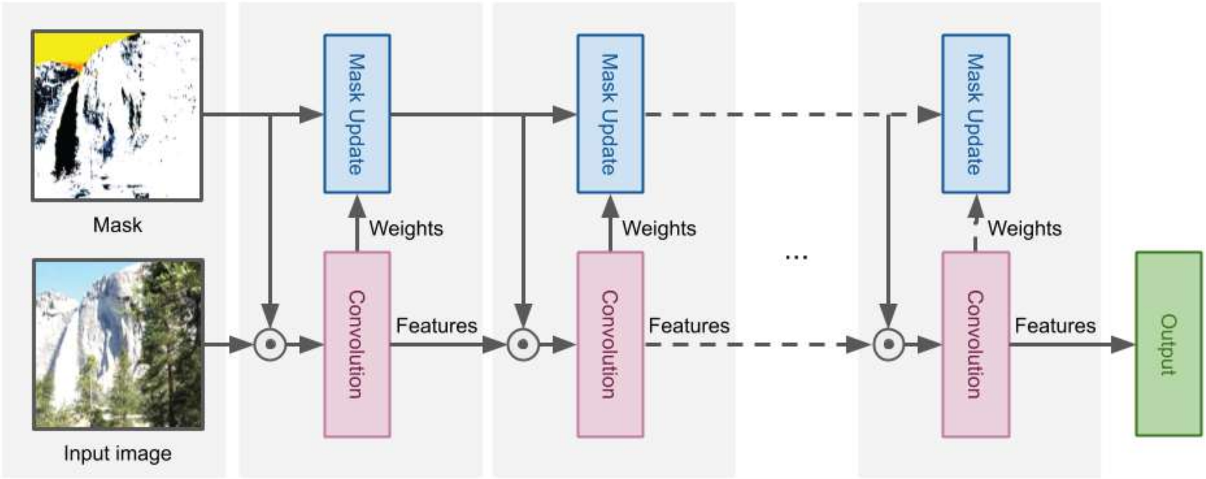

Figure 7 – Illustration of the proposed feature masking mechanism. The features at each layer are multiplied with the corresponding mask before going through the convolution process. The masks at each layer are obtained by updating the masks at the previous layer using Equation 3.4.

where W_l and b_l refer to the weight and bias of the current layer, respectively. Moreover, ϕ_l is the activation function and $*$ is the standard convolution operation.

We compute the masks at each layer by applying the convolutional filter to the masks at the previous layer (See Figure 8 for visualization of some of the masks). The basic idea is that since the features are computed by applying a series of convolutions, the same filters can be used to compute the contribution of the valid pixels in the features. However, since the masks are in the range $[0, 1]$ and measure the percentage of the contributions, the magnitude of the filters is irrelevant. Therefore, we normalize the filter weights before convolving them with the masks as follows:

$$M_{l+1} = \left(\frac{|W_l|}{\|W_l\|_1 + \epsilon}\right) * M_l, \tag{3.4}$$

where $\|\cdot\|_1$ is the l_1 function and $|\cdot|$ is the absolute operator. Here, $|W_l|$ is a $\mathbb{R}^{H\times W\times C}$ tensor and $\|W_l\|_1$ is a $\mathbb{R}^{1\times 1\times C}$ tensor. To perform the division, we replicate the values of $\|W_l\|_1$ to obtain a tensor with the same size as $|W_l|$. The constant ϵ is a small value to avoid division by 0 ($\epsilon = 10^{-6}$ in our implementation).

It is important to mention that a couple of recent approaches have proposed strategies to overcome similar issues in image inpainting task (YU et al., 2019; LIU et al., 2018). Specifically, Liu et al. (2018) propose to modify the convolution process to only apply the filter to the pixels with valid information. Unfortunately, this approach is specially designed for cases with binary masks. However, the masks in our application are soft and, thus, this method is not applicable. Yu et al. (2019) propose to multiply the features at each layer with a soft mask, similar to our feature masking strategy. The key difference is that their mask at each layer is learnable, and it is estimated using a small network from

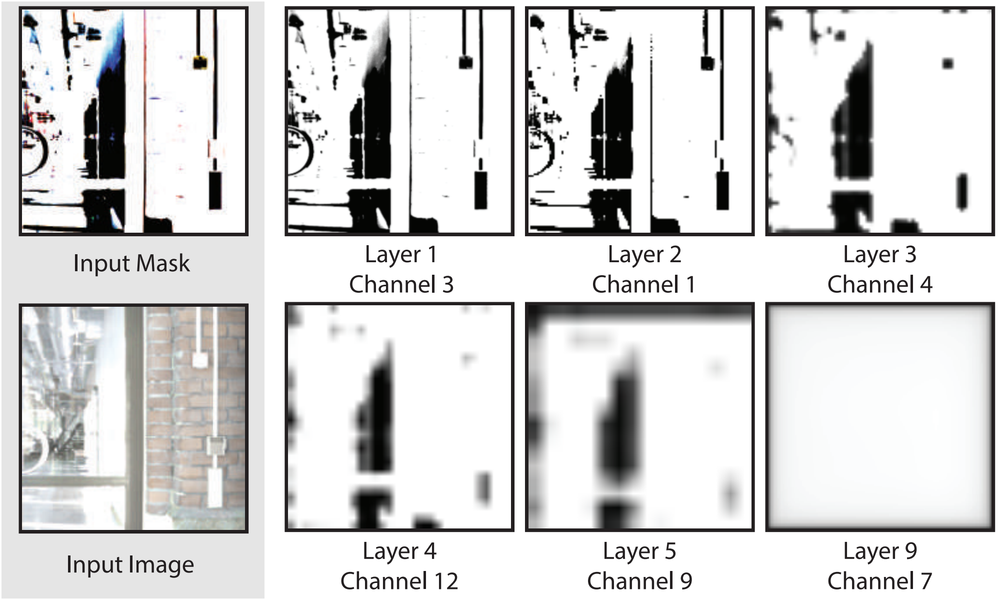

Figure 8 – On the left, we show the input image and the corresponding mask. On the right, we visualize a few masks at different layers of the network. Note that, as we move deeper through the network, the masks become blurrier and more uniform. This is expected since the receptive field of the features become larger in the deeper layers.

the features in the previous layer. Because of the additional parameters and complexity, training this approach on limited HDR images is difficult. Therefore, this approach is not able to produce high-quality HDR images (see Section 4.5).

3.3 LOSS FUNCTION

The choice of the loss function is critical in each learning system. Our goal is to reconstruct an HDR image by synthesizing plausible textures in the saturated areas. Unfortunately, using only pixel-wise loss functions, as utilized by most previous approaches, the network tends to produce blurry images (Figure 18). Inspired by the recent image inpainting approaches (YANG et al., 2017; LIU et al., 2018; HAN et al., 2019), we train our network using a VGG-based perceptual loss function. Specifically, our loss function is a combination of an HDR reconstruction loss L_r and a perceptual loss L_p, as follows:

$$L = \lambda_1 L_r + \lambda_2 L_p \tag{3.5}$$

where $\lambda_1 = 6.0$ and $\lambda_2 = 1.0$ in our implementation. We define these terms using an informal hyper-parameter search on 103 validation images. We did not perform a systematic grid search due to the high computational cost.

3.3.1 Reconstruction Loss

The HDR reconstruction loss is a simple pixel-wise l_1 distance between the output and ground truth images in the saturated areas. Since the HDR images could potentially have large values, we define the loss in the logarithmic domain. Otherwise, the network would focus heavily in the high luminance values, leading to underestimation of important differences in the lower range of luminaces. This formulation for the reconstruction loss function is motivated by the Weber-Fechner law (FECHNER, 1860). This law states that the response of the human visual system (HVS) is close to the logarithmic in large areas of the luminance range, which implies a logarithmic relationship between physical luminance and the perceived brightness.

Given the estimated HDR image $\hat{Y}$ (in the log domain) and the linear ground truth image H, the reconstruction loss is defined as:

$$L_r = \|(1 - M) \odot (\hat{Y} - \log(H + 1))\|_1. \tag{3.6}$$

Here, the multiplication by $(1-M)$ ensures that the loss is computed only in the saturated areas. This avoids the network to unnecessarily learn the non-saturated regions since we take the information for these areas from the linear input image (as defined in Equation 3.1).

3.3.2 Perceptual Loss

Our perceptual term is a combination of the VGG and style loss functions (JOHNSON; ALAHI; FEI-FEI, 2016) as follows:

$$L_p = \lambda_3 L_v + \lambda_4 L_s. \tag{3.7}$$

In our implementation, we set $\lambda_3 = 1.0$ and $\lambda_4 = 120.0$. These terms were also determined by performing a hyper-parameter search on a validation set. The VGG loss function L_v evaluates how well the features of the reconstructed image match with the features extracted from the ground truth. This allows the model to produce textures that are perceptually similar to the ground truth. This loss term is defined as follows:

$$L_v = \sum_l \|\phi_l(\mathcal{T}(\tilde{H})) - \phi_l(\mathcal{T}(H))\|_1 \tag{3.8}$$

where ϕ_l is the feature map extracted from the l^{th} layer of the VGG network. Moreover, the image $\tilde{H}$ is obtained by combining the information of the ground truth H in the well-exposed regions and the content of the network's output $\hat{Y}$ in the saturated areas using the mask M, as follows:

$$\tilde{H} = M \odot H + (1 - M) \odot \hat{Y}. \tag{3.9}$$

We use $\tilde{H}$ in our loss functions to ensure that the supervision is only provided in the saturated areas. Finally, $\mathcal{T}(\cdot)$ in Equation 3.8 is a function that compresses the range to $[0, 1]$. Specifically, we use the μ-law function, a commonly-used range compressor in audio processing, which is differentiable and therefore suitable for our learning system. This function is defined as:

$$\mathcal{T}(H) = \frac{\log(1 + \mu H)}{\log(1 + \mu)}, \tag{3.10}$$

where μ is a parameter defining the amount of compression. In our implementation, we set $\mu = 500$, which produces good results in our experiments. Since VGG is trained with LDR images from the ImageNet dataset (DENG et al., 2009), this process is performed in order to ensure that the input to the network is similar to the ones that it has been trained on.

The style loss in Equation 3.7 (L_s) captures style and texture by comparing global statistics with a Gram matrix (GATYS; ECKER; BETHGE, 2015) collected over the entire image. Specifically, the style loss is defined as:

$$L_s = \sum_l \|G_l(\mathcal{T}(\tilde{H})) - G_l(\mathcal{T}(H))\|_1, \tag{3.11}$$

where $G_l(X)$ is the Gram matrix of the features in layer l and is defined as follows:

$$G_l(X) = \frac{1}{K_l}\phi_l(X)^T\phi_l(X). \tag{3.12}$$

Here, K_l is a normalization factor computed as $C_l H_l W_l$. Note that, the feature ϕ_l is a matrix of shape $(H_l W_l) \times C_l$ and, thus, the Gram matrix has a size of $C_l \times C_l$. In our implementation, we use the VGG-19 (SIMONYAN; ZISSERMAN, 2015) network and extract features from layers `pool1`, `pool2` and `pool3`.

As we show in Figure 18, the proposed perceptual loss function is essential for hallucinating details and synthesizing realistic texture in the saturated areas, as opposed to a simple pixel-wise l_1 loss function.

3.4 INPAINTING PRE-TRAINING

Deep learning methods usually require large-scale training datasets, however large-scale HDR image datasets are currently not available, which makes training our system a difficult task. Existing techniques (EILERTSEN et al., 2017) overcome this limitation by pre-training their network on simulated HDR images that are created from standard image datasets like the MIT Places (ZHOU et al., 2014). They then fine-tune their network on real HDR images. Unfortunately, our network is not able to learn to synthesize plausible textures with this training strategy (see Figure 17), as the saturated areas are typically in

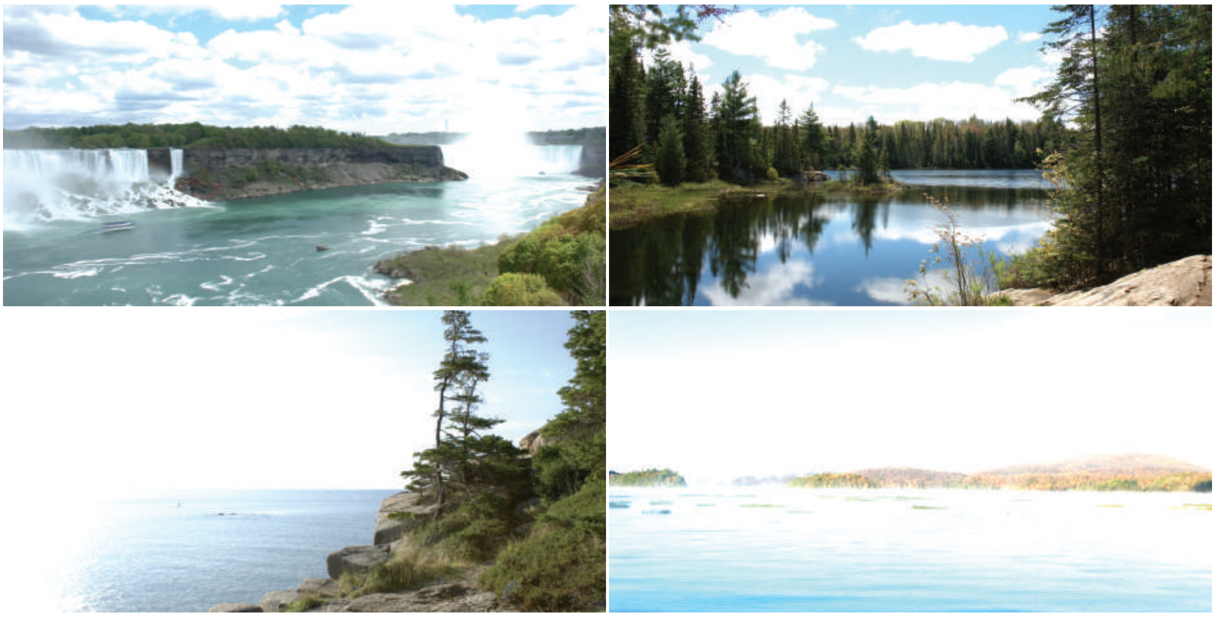

Figure 9 – The saturated areas are typically in textureless regions, such as sky, clouds and water. As a result, the network is not able to hallucinate plausible textures using solely the HDR images for training. We propose to pre-train our network in an auxiliary task to encourage the network to synthesize textures.

the bright and smooth regions, such as sky, clouds and water (see Figure 9) and therefore do not provide enough supervision for texture synthesis.

To address this problem, we propose to pre-train our network on image inpainting tasks. Intuitively, during inpainting, the masks are obtained randomly and therefore the missing areas are diverse. As a result, our network leverages a large number of training data to learn an appropriate internal representation that is capable of synthesizing visually pleasing textures. In the HDR fine-tuning stage, the network adapts the learned representation to the HDR domain to be able to synthesize HDR textures. We follow Liu et al.'s approach 2018 and use their loss function and mask generation strategy during pre-training. Note that we still use our feature masking mechanism for pre-training, but the input masks are binary. We fine-tune the network on real HDR images using the loss function, discussed in Section 3.3.

One major problem is that the majority of the bright areas in the HDR examples are smooth and textureless and the patches containing textures are scarce. Therefore, during fine-tuning, the network adapts to these types of patches and, as a result, has difficulty producing textured results (see Figure 17). In the next section, we discuss our strategy to select textured and challenging patches.

3.5 PATCH SAMPLING

Our goal is to select the patches that contain texture in the saturated areas. We perform this by first computing a score for each patch and then choosing the patches with a high

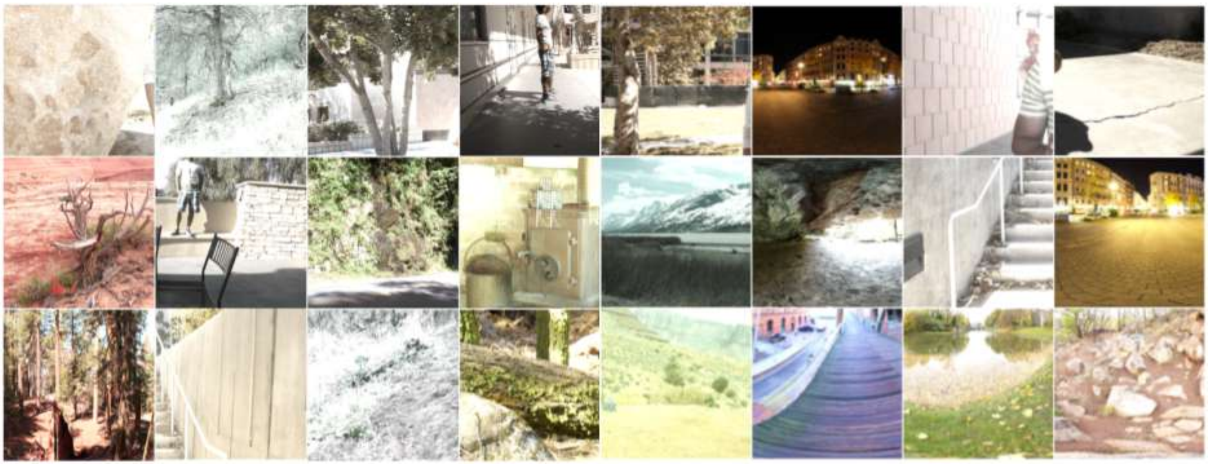

Figure 10 – A few example patches selected by our patch sampling approach. These are challenging examples as the HDR images corresponding to these patches contain complex textures in the saturated areas.

score. The main challenge here is finding a good metric that properly detects the textured patches. One way to do this is to compute the average of the gradient magnitude in the saturated regions. However, since our images are in HDR and can have large values, this approach can detect a smooth region with bright highlights as textured.

To avoid this issue, we propose to first decompose the HDR image into base and detail layers using a bilateral filter (DURAND; DORSEY, 2002). We use the average of the gradients (using the Sobel operator) of the detail layer in the saturated areas as our metric to detect the textured patches. We consider all the patches with a mean gradient above a certain threshold (0.85 in our implementation) as textured, and the rest are classified as smooth. Since the detail layer only contains variations around the base layer, this metric can effectively measure the amount of textures in an HDR patch. Figure 10 shows example of patches selected using this metric. As shown in Figure 17, this simple patch sampling approach is essential for synthesizing HDR images with sharp and artifact-free details in the saturated areas. The summary of our patch selection strategy is listed in Algorithm 1.

Algorithm 1 Patch Sampling

```
procedure PatchMetric($H$, $M$)
    $H$: HDR image, $M$: Mask
    $\sigma_c = 100.0$                          ▷ Bilateral filter color sigma
    $\sigma_s = 10.0$                           ▷ Bilateral filter space sigma
    $I$ = RgbToGray($H$)
    L = log($I$ + 1)
    B = bilateralFilter($L, \sigma_c, \sigma_s$)                ▷ Base Layer
    D = L - B                                         ▷ Detail Layer
    $G_x$ = getGradX($D$)
    $G_y$ = getGradY($D$)
    G = abs($G_x$) + abs($G_y$)
    N = $G \odot (1 - M)$        ▷ Computes the metric only in the saturated areas
    return mean(N)
end procedure
```

3.6 SUMMARY

In this chapter we discuss each of the building blocks of our system. Specifically, we discuss the feature masking mechanism and the automatic mask updating process. This mechanism alleviates the artifacts caused by conditioning the convolutional layers on the saturated pixels. Moreover, we discuss the changes we made to the VGG-based perceptual loss function in order to adapt it to the HDR reconstruction task. We propose to train the system in two stages where we pre-train the network on inpainting before fine-tuning it on HDR generation. To encourage the network to synthesize textures, we propose a sampling strategy to select challenging patches in the HDR examples.

Chapter 4

EXPERIMENTS AND RESULTS

In this chapter we present several examples verifying the quality of the HDR reconstructions on both synthetic (Section 4.2) and real camera LDR images (Section 4.3). We then present a study of performance of our system (Section 4.4). Here, we only compare our method to other state of the art learning-based approaches. This is because Eilertsen et al. (2017) demonstrated that this class of methods outperform the non-learning-based approaches by a wide margin in several scenarios. Specifically, we compare our approach against three existing learning-based single image HDR reconstruction approaches of Endo, Kanamori e Mitani (2017), Eilertsen et al. (2017), and Marnerides et al. (2018). We use the source code provided by the authors to generate the results for each of these approaches. We then show the result of several ablations we performed to show the effectiveness of each component of our system (Section 4.5). Finally, we show some failure cases of our approach (Section 4.6) to motivate future explorations. All images that we show in this section have been tone-mapped (exposure reduction following by a gamma correction) for display purposes.

4.1 IMPLEMENTATION

We start by discussing some implementations details. We implement our network in PyTorch (PASZKE et al., 2019), but write the data pre-processing, data augmentation, and patch sampling code in C++ for performance. We implement the feature masking mechanism using the existing standard convolutional layer in PyTorch, however it can be improved both in time and space using custom layers. Nevertheless, the entire network inference on a 512 × 512 image takes 300ms on a single NVIDIA GTX 1080Ti GPU with 11GB of video memory and has competitive performance to the existing learning-based single image HDR reconstruction approaches (see Section 4.4).

4.1.1 Architecture

We use a network with U-Net architecture (RONNEBERGER; FISCHER; BROX, 2015) similar to the one used in Isola et al. (2017), as shown in Figure 11. We use the feature masking strategy in all the convolutional layers and up-sample the features in each layer in the decoder using nearest neighbor method. All the encoder layers use Leaky ReLU activation function (MAAS; HANNUN; NG, 2013). On the other hand, we use ReLU (NAIR; HINTON, 2010) in all the decoder layers, with the exception of the last one, which has a linear activation function. We use skip connections between all the encoder layers and their corresponding decoder layers to concatenate the corresponding feature maps. The Table 1 contains a more detailed specification of our network architecture.

Table 1 – Our network architecture. **k** is the kernel size, **s** the stride, **p** the padding, **chns** is the number of input and output channels for each layer, **input** denotes the input of each layer with + meaning the features concatenation, and layers starting with "nnup" perform 2x nearest neighbor upsampling. All convolutional layers (**conv_x**) refer to the proposed feature masking mechanism.

Layer	**k**	**s**	**p**	**chns**	**input**
conv_1	7	2	3	3/64	LDR image
conv_2	5	2	2	64/128	conv_1
conv_3	5	2	1	128/256	conv_2
conv_4	3	2	1	256/512	conv_3
conv_5	3	2	1	512/512	conv_4
conv_6	3	2	1	512/512	conv_5
conv_7	3	2	1	512/512	conv_6
nnup1				512/512	conv_7
conv_8	3	1	1	1024/512	nnup1 + conv_6
nnup2				512/512	conv_8
conv_9	3	1	1	1024/512	nnup2 + conv_5
nnup3				512/512	conv_9
conv_10	3	1	1	1024/512	nnup3 + conv_4
nnup4				512/512	conv_10
conv_11	3	1	1	768/256	nnup4 + conv_3
nnup5				256/256	conv_11
conv_12	3	1	1	384/128	nnup5 + conv_2
nnup6				128/128	conv_12
conv_13	3	1	1	192/64	nnup6 + conv_1
nnup7				64/64	conv_13
conv_14	3	1	1	67/3	nnup7+input

4.1.2 Dataset

We use different datasets for each training step (pre-training and fine-tuning) of our method. We summarize each of them and the pre-processing methods (when necessary) in the following sections.

4.1.2.1 Inpainting pre-training

For the image inpainting step, we use the MIT Places (ZHOU et al., 2014) dataset with the original train, test, and validation splits. We choose Places for this step because it

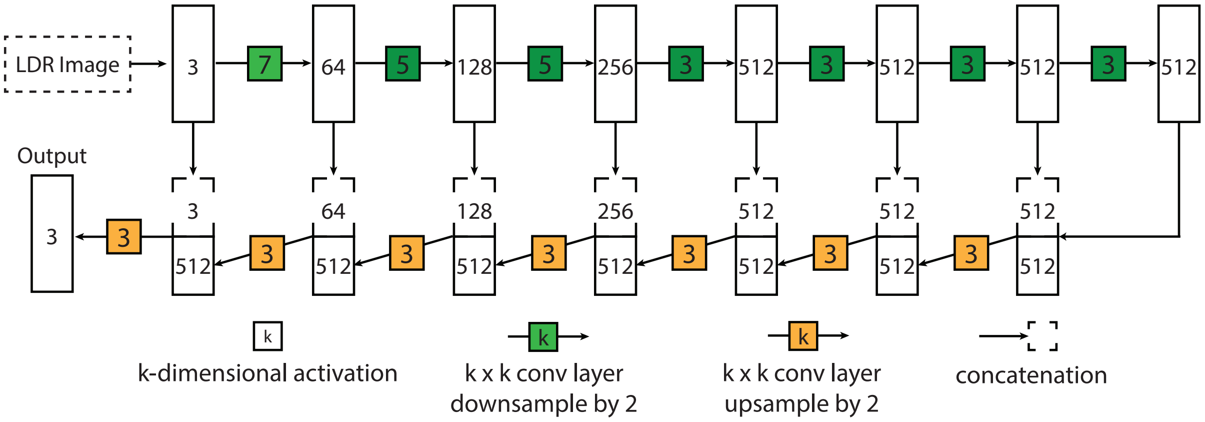

Figure 11 – The proposed network architecture. The model takes as input the RGB LDR image and outputs an HDR image. We use a feature masking mechanism in all the convolutional layers.

contains a large number of scenes ($\sim 2.5M$ images) with diverse textures. We use the method of Liu et al. 2018 to generate masks of random streaks and holes of arbitrary shapes and sizes.

4.1.2.2 HDR fine-tuning

For the HDR fine-tuning step, we collect approximately 2,000 HDR images from 735 HDR images and 34 HDR videos collected from various sources. We generate the input LDR images for our system following the approach by Eilertsen et al. 2017.

Specifically, for each HDR image in the dataset, Eilertsen et al. 2017 setup a virtual camera that simulates several cameras attributes such as exposure (selected so that 5-15% of the total number of pixels are saturated), camera curve, white balance and noise. The different camera curves are approximated using a parametric function in form of a sigmoid,

$$f(H) = (1+\sigma)\frac{H^n}{H^n+\sigma} \tag{4.1}$$

The parameters n and σ are selected to fit the mean of the database of camera curves collected by Grossberg e Nayar (2003b), where $n = 0.9$ and $\sigma = 0.6$ gives a good fit as shown by Eilertsen et al. 2017. For randomly selecting the camera curves in the training data synthesis, these parameters are drawn from normal distributions around the fitted values, $n \sim \mathcal{N}(0.9, 0.1)$ and $\sigma \sim \mathcal{N}(0.6, 0.1)$. Moreover, the noise in the virtual camera is simulated by injecting an additive Gaussian noise with variance sampled in the range $[0, 0.01]$. The virtual camera shots with several randomly selected attributes to obtain the images used as input for the network.

Along with the aforementioned virtual camera augmentation we also employ other standard augmentation strategies. Specifically, augmentation in terms of colors is accomplished in the HSV color space. Here, we modify the hue and saturation channels by

Table 2 – Numerical comparison in terms of mean square error (MSE) and HDR-VDP-2 (MANTIUK et al., 2011) against existing learning-based single image HDR reconstruction approaches.

Method	MSE	HDR-VDP-2
Endo, Kanamori e Mitani (2017)	0.0390	55.67
Eilertsen et al. (2017)	0.0387	59.11
Marnerides et al. (2018)	0.0474	54.31
Ours	**0.0356**	**63.18**

adding a random perturbation $\tilde{h} \sim \mathcal{N}(0,7)$ and $\tilde{s} \sim \mathcal{N}(0,0.1)$ respectively. We also apply random flipping with a probability of 0.5 and extract 250 random patches of size 512×512 for each image. The processed linear images represent the ground truth data, while the inputs for training are clipped at 1. Finally, we select a subset of these patches using our patch selection strategy (Section 3.5). We also discard patches with no saturated content, since they do not provide any source of learning to the network. Each of the remaining pairs are the training instances for our method. All the data pre-processing, data augmentation, and patch sampling were done offline in C++.

Our final training dataset is a set of 100K input and corresponding ground truth patches. This data augmentation approach is responsible for a training model that generalizes well to a wide range of cameras (see Section 4.3).

4.1.3 Training

We initialize our network using the Xavier approach (GLOROT; BENGIO, 2010) and train it on image inpainting task until convergence. We then fine-tune the network on HDR reconstruction. We train the network with a learning rate of 2×10^{-4} in both stages. However, during the second stage, we reduce the learning rate by a factor of 2.0 when the optimization plateaus. The training process is performed until convergence. Both inpainting and HDR fine-tuning stages are optimized using gradient descent with momentum, employing the ADAM (adaptive moment estimation) optimizer (KINGMA; BA, 2015) with the default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and mini-batch size of 4. The entire training takes approximately 11 days on a machine with an Intel Core i7, 16GB of memory, and an Nvidia GTX 1080 Ti GPU with 11GB of video memory.

4.2 RESULTS FOR SYNTHETIC IMAGES

We begin by quantitatively comparing our approach against the other methods in terms of mean squared error (MSE) and HDR-VDP-2 (MANTIUK et al., 2011) in Table 2. The errors are computed on a test set of 75 randomly selected HDR images, with resolutions ranging from 1024×768 to 2084×2844. We generate the input LDR images using various

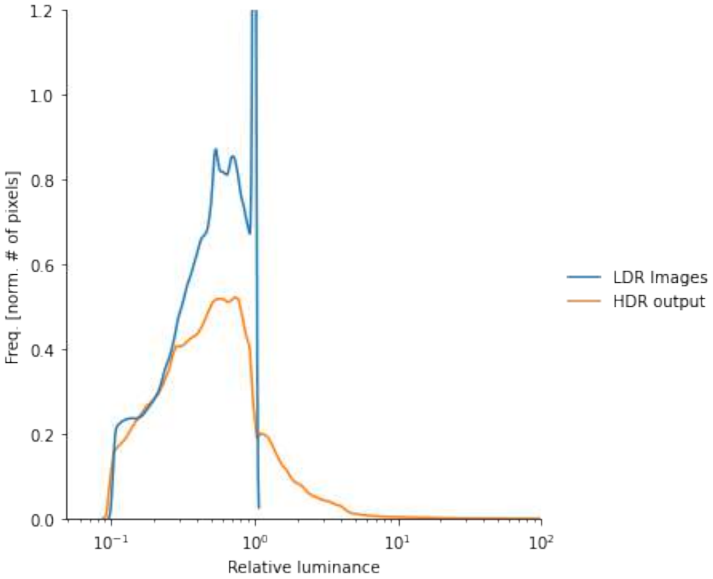

Figure 12 – Histogram of the test set images of the resulting high dynamic range images of our approach compared to the range of input LDR images. Notice that our method is able to reconstruct and therefore expand the range of the LDR images (clipped at relative luminance = 1 due to sensor saturation).

camera curves and exposures, similar to the approach by Eilertsen et al. (2017). We compute the MSE values on the gamma corrected images and HDR-VDP-2 scores are obtained on the linear HDR images. As seen, our method produces significantly better results, which demonstrate the ability of our network to accurately recover the full range of luminance. To further demonstrate the ability of our network to reconstruct the range of luminance, we show in Figure 12 the resulting dynamic range of our approach on the test set LDR images. As can be seen, our method reconstructs the missing luminance of the LDR images (clipped at relative luminance = 1 due to sensor saturation).

Next, we compare our approach against the other methods on five challenging scenes in Figure 13. Overall other approaches are not able to synthesize texture and produce results with blurriness, discoloration, and checkerboard artifacts. However, our approach can effectively utilize the information in the non-saturated color channels and the contextual information to synthesize visually pleasing textures. It is worth noting that although our approach has been trained using a perceptual loss, it can still properly recover the bright highlights. For example, our results in Figure 13 (fifth row) are similar to Eilertsen et al. (2017) and better than Endo, Kanamori e Mitani (2017) and Marnerides et al. (2018).

We also demonstrate that our approach can consistently generate high-quality results on images with different amount of saturated areas in Figure 14. As can be seen, the results of all the other approaches degrade quickly by increasing the percentage of the saturated pixels in the input LDR image. On the other hand, our approach is able to produce high-quality results with sharp details and bright highlights in all the cases.

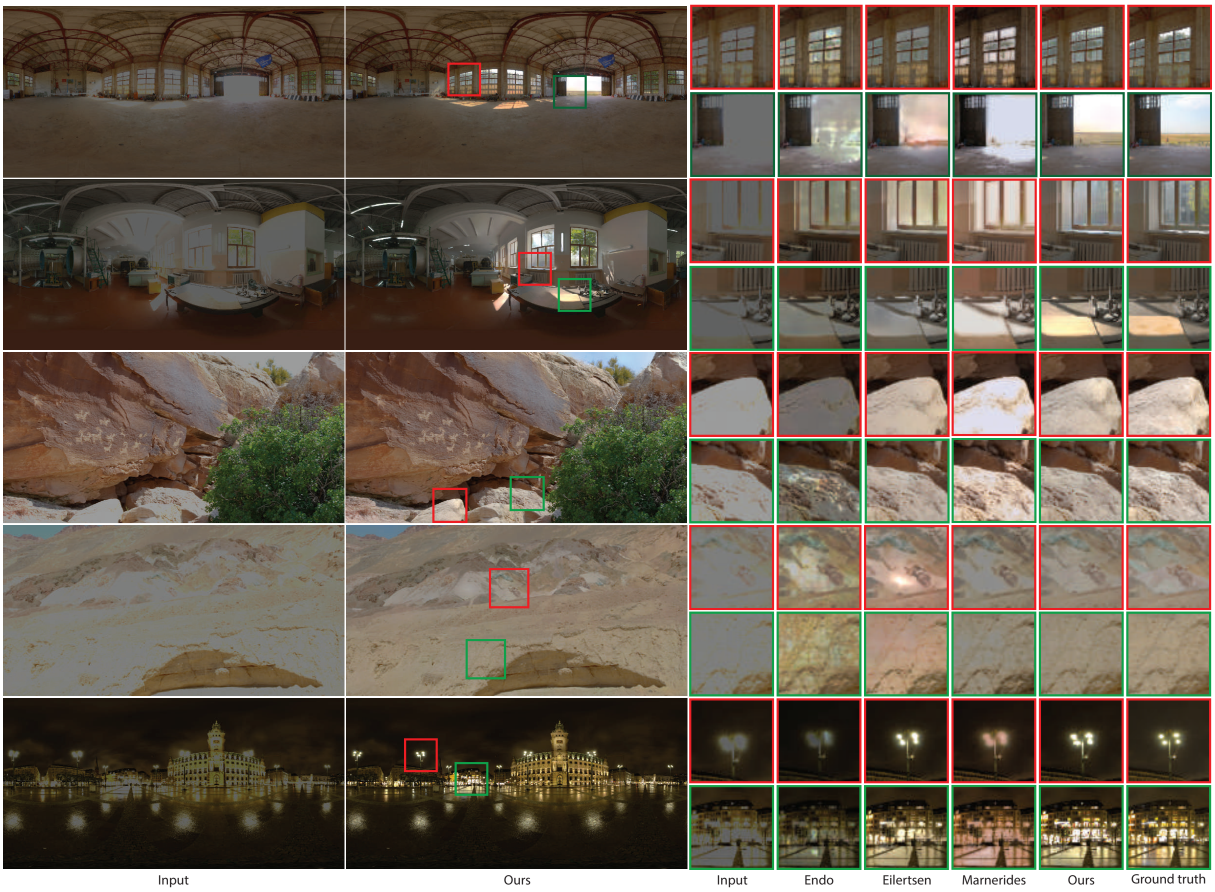

Figure 13 – We compare our method against state-of-the-art approaches of Endo, Kanamori e Mitani (2017), Eilertsen et al. (2017), and Marnerides et al. (2018) on a diverse set of synthetic scenes. Our method is able to synthesize textures in the saturated areas better than the other approaches (rows one to four), while producing results with similar or better quality in the bright highlights (fifth row).

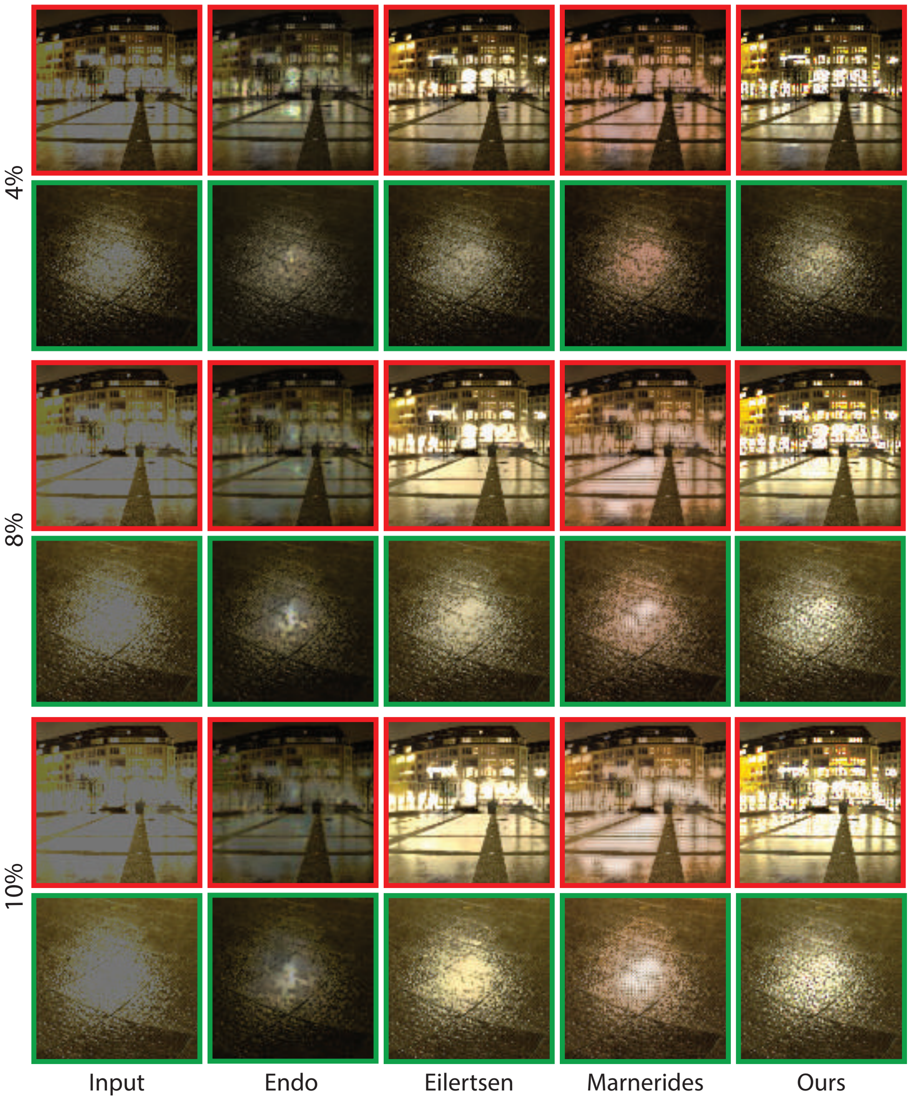

Figure 14 – We compare the performance of the proposed method against previous methods for various amounts of saturated areas. The numbers indicate the percentage of the total number of pixels that are saturated in the input. Although our method slightly degrades as the saturation increases, we consistently present better results than the previous methods.

4.3 RESULTS FOR REAL IMAGES

We show the generality of our approach by producing results on a set of real images taken in a variety of situations, captured with various standard cameras, in Figure 15. Specifically, the first three images in the left are from Google HDR+ dataset (HASINOFF et al., 2016), captured with a variety of Google's smartphones, such as Nexus 5/6/5X/6P, Pixel, and Pixel XL. The image in the last column is captured using a DSLR camera Canon 5D Mark IV. These cameras provide more realistic scenarios than the synthetic images we discussed in Section 4.2.

As we can see, all the other approaches are not able to properly reconstruct the saturated regions, producing results with discoloration, blurriness and color shift, as indicated by the red arrows. On the other hand, our method is able to properly increase the dynamic range by synthesizing realistic textures in the saturated areas. The examples demonstrate that our method is able to generalize to different cameras and also to produce textures that match the context in a wide range of situations and objects, such as walls, stones and sand.

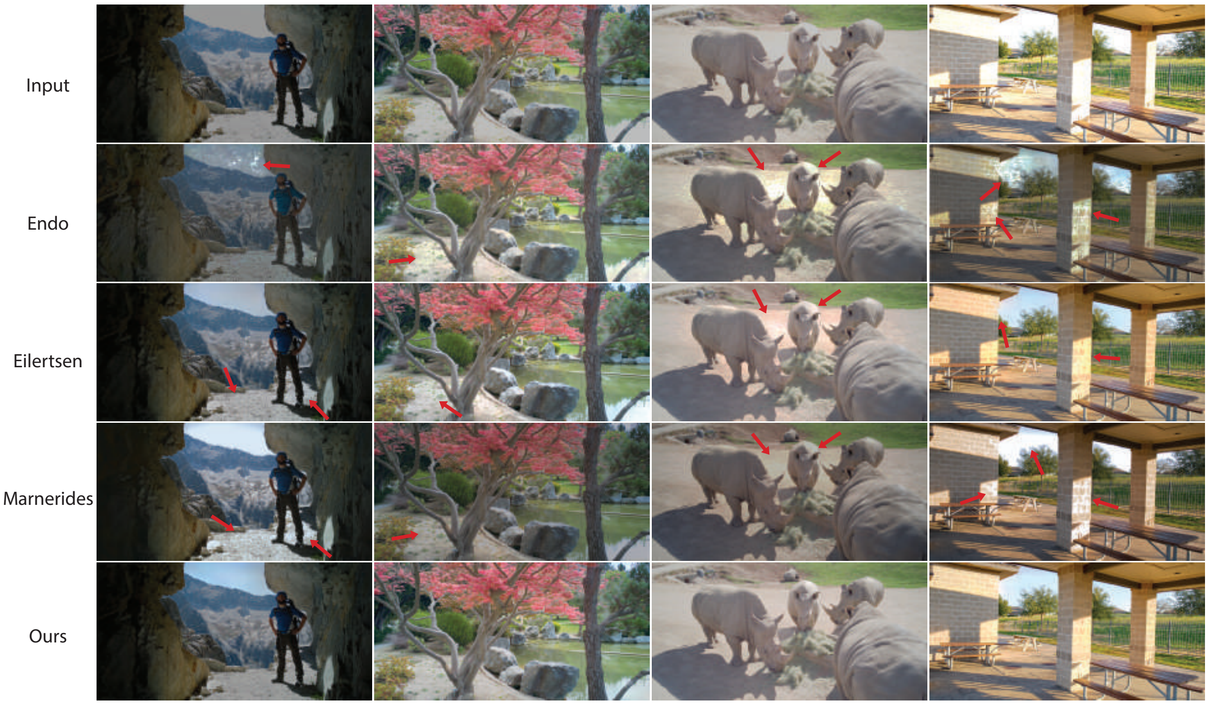

Figure 15 – Comparison against state-of-the-art approaches on images captured by standard cameras. Plausible reconstructions of textured areas can be made of walls, stones and sand. Zoom in to the electronic version to see the differences across the images.

Table 3 – We evaluate the effectiveness of our masking and pre-training strategies by comparing against other alternatives in terms of MSE and HDR-VDP-2 (MANTIUK et al., 2011). Here, SConv, GConv, IMask, and FMask refer to standard convolution, gated convolution (YU et al., 2019), only masking the input image, and our full feature masking approach, respectively. Moreover, Inp. pre-training and HDR pre-training correspond to our proposed pre-training on inpainting and HDR reconstruction tasks, respectively.

Method (Masking + Pre-training)	**MSE**	**HDR-VDP-2**
SConv + HDR pre-training	0.0402	58.43
SConv + Inp. pre-training	0.0374	60.03
GConv + HDR pre-training	0.0398	53.32
GConv + Inp. pre-training	0.1017	43.13
IMask + HDR pre-training	0.0398	58.39
IMask + Inp. pre-training	0.0369	61.27
FMask + HDR pre-training	0.0393	58.81
FMask + Inp. pre-training (Ours)	**0.0356**	**63.18**

4.4 PERFORMANCE COMPARISON

We now provide a timing comparison of our approach against the other methods. We obtain the numbers reported in Table 4 by running each model on a CPU Intel Core i5-6500 CPU @ 3.20GHz x 4. On average, our approach produces a single frame with resolution of 1024 in 5.6 seconds. As we can see, our method is faster than all previous methods. Comparing to the method of Endo, Kanamori e Mitani (2017), Eilertsen et al. (2017), and Marnerides et al. (2018), our approach is 30, 1.5, and 3 times faster, respectively. Our approach is also significantly faster than the three other methods for images with resolutions of 512 × 512. Notice that we implement the Feature Masking using the existing standard convolutional layer in PyTorch. Therefore, we could achieve even higher speed up using custom layers implemented using native PyTorch modules. We leave for future work to explore methods for speeding up our approach to enable real-time inference (see Section 5.1).

Table 4 – Timing comparison on CPU with the methods of Endo, Kanamori e Mitani (2017), Eilertsen et al. (2017), and Marnerides et al. (2018) on inputs with different resolutions.

Method	**512 × 512**	**1024 × 512**
Endo, Kanamori e Mitani (2017)	62.8s	154.5s
Eilertsen et al. (2017)	4.1s	7.9s
Marnerides et al. (2018)	7.9s	16.9s
Ours	**2.3s**	**5.6s**

4.5 ABLATION STUDIES

4.5.1 Feature Masking

We begin comparing our feature masking strategy against several other approaches in Table 3. Specifically, we compare our method against standard convolution (SConv), gated convolution (YU et al., 2019) (GConv), and the simpler version of our masking strategy where the mask is only applied to the input (IMask). For completeness, we include the result of each method with both inpainting and HDR pre-training. As seen, our masking strategy is considerably better than the other methods. It is worth noting that unlike other methods, the performance of gated convolution with inpainting pre-training is worse than HDR pre-training. This is mainly because gated convolution estimates the masks at each layer using a separate set of networks which become unstable after transitioning from inpainting pre-training to HDR fine-tuning.

We also visually compare our feature masking method against standard convolution in Figure 16. Standard convolution produces results with checkerboard artifacts (top) and halo and blurriness (bottom), while our network with feature masking produces considerably better results. These artifacts happen because standard convolutional layers treat all input pixels as valid ones and apply the same filter to the entire image. However, our input LDR image contains invalid information in the saturated areas. Since meaningful features cannot be extracted from the saturated contents, naïve application of standard convolution introduces ambiguity during training and leads to visible artifacts, as we discussed in Section 3.2. Moreover, we visually compare our approach against other masking strategies in Figure 17. Note that, for each masking strategy, we only show the combination of masking and pre-training that produces the best numerical results in Table 3, i.e., gated convolution (GConv) with HDR pre-training and input masking (IMask) with inpainting pre-training. Gated convolution is not able to produce high frequency textures in the saturated areas. Input masking performs reasonably well, but still introduces noticeable artifacts. Our feature masking method, however, is able to synthesize visually pleasing textures. This is helpful since it removes the response from completely saturated contents and the weak signals can propagate through the network more efficiently. Figure 13 shows an example of this case where the outside area in the green inset is completely saturated. However, using the weak signals, our approach is able to recover the sky, grass, and concrete floor properly.

4.5.2 Inpainting Pre-training

Here, we study the effect of the proposed inpainting pre-training step by comparing it against the commonly-used synthetic HDR pre-training in Table 3 and Figure 17. As seen, our pre-training (“FMask + Inp. pre-training (Ours)”) performs better than HDR pre-training (“FMask + HDR pre-training”) both numerically and visually. Specifically,

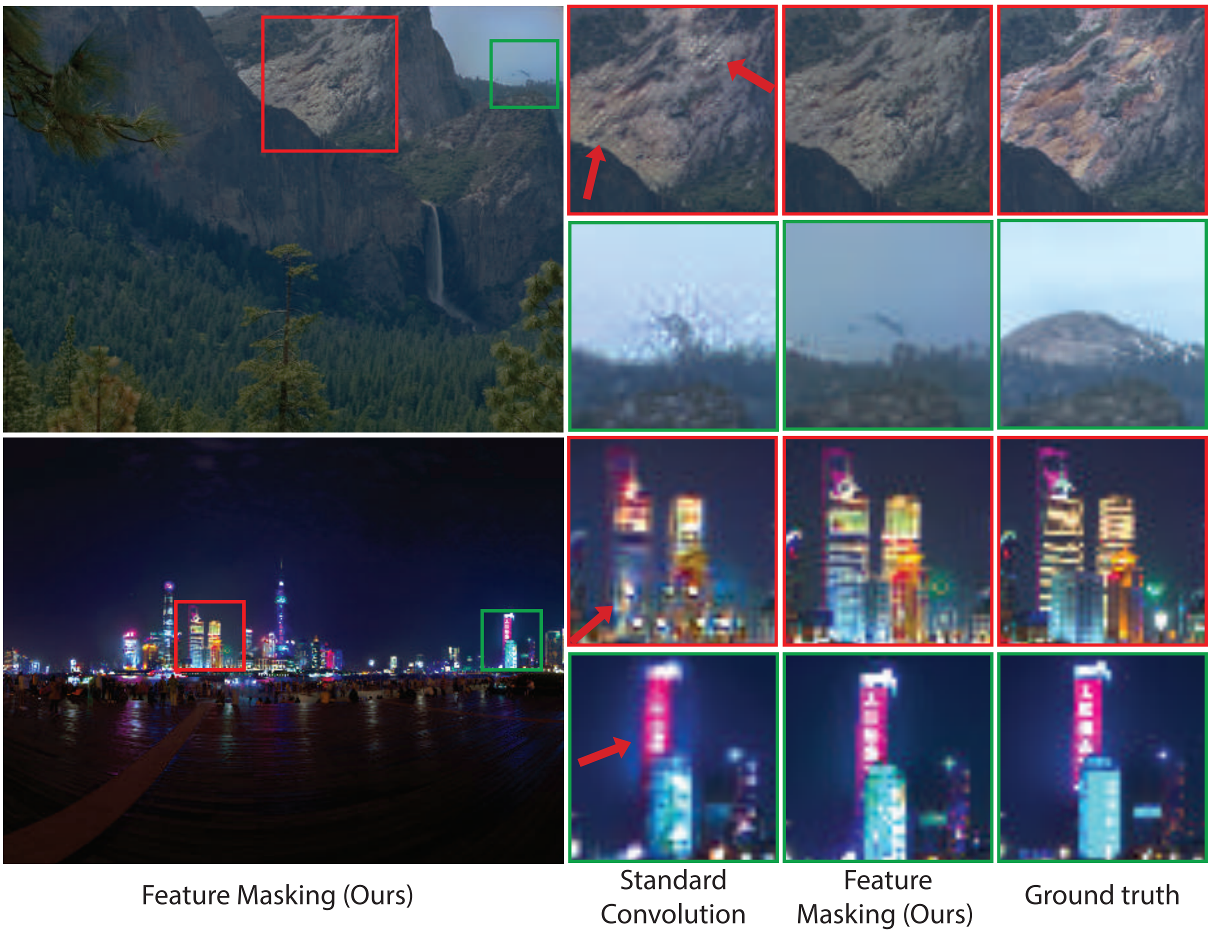

Figure 16 – In regions with both saturated and well-exposed content (boundaries of sky and mountain and bright building lights), the response of the invalid saturated areas in standard convolution dominates the feature maps. Therefore, the network cannot properly utilize the content of the valid regions, introducing high frequency checkerboard artifacts (top row) and blurriness and halo (bottom row). Our approach suppresses the features from the saturated content and allows the network to synthesize the image using the well-exposed information.

as shown in Figure 17, our network using inpainting pre-training is able to learn better features and synthesizes sharp textures in the saturated areas.

4.5.3 Patch Sampling

We show our result without patch sampling (Section 3.5) to demonstrate its effectiveness in Figure 17. As seen, by training on the textured patches (ours), the network is able to synthesize textures with more details and fewer objectionable artifacts.

4.5.4 Loss Function

Finally, we compare the proposed perceptual loss function against a simple pixel-wise (l_1) loss. Specifically, we train our network with only a simple pixel-wise (l_1) loss and with

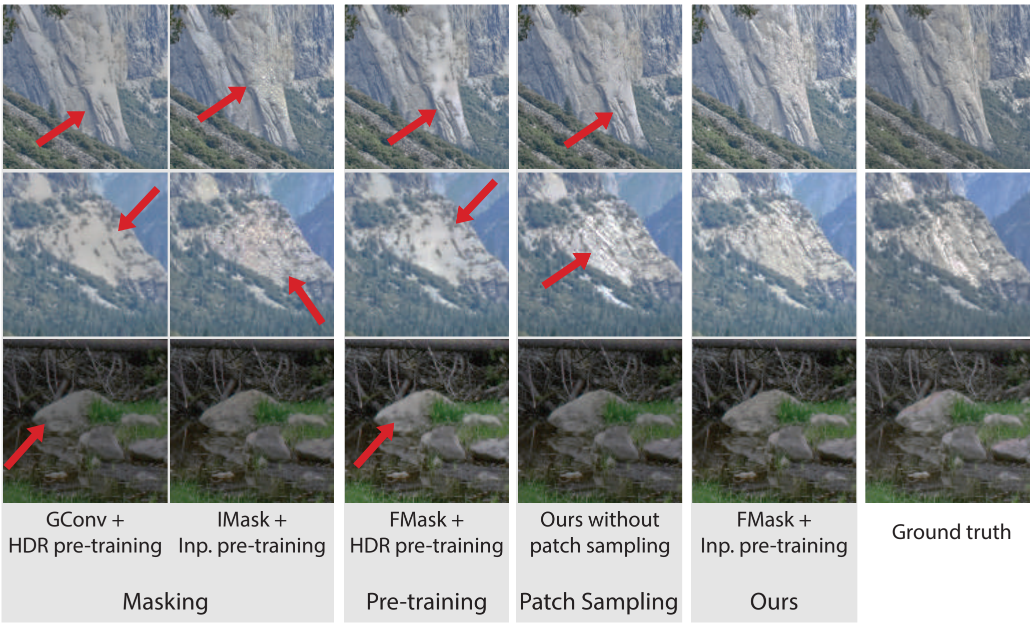

Figure 17 – From left to right, we compare our method against two other masking strategies as well as a pre-training method, and evaluate the effect of patch sampling. Here, GConv, IMask, and FMask refer to gated convolution (YU et al., 2019), only masking the input image, and our full feature masking method, respectively. Moreover, Inp. pre-training refers to our proposed pre-training on inpainting task.

the purposed perceptual loss function. As seen in Figure 18, using only the pixel-wise loss function our network tends to produce blurry images, while the network trained using the proposed perceptual loss function can produce visually realistic textures in the saturated regions. Notice that the pixel-wise (l_1) loss is fundamental for the network reconstructing the dynamic range in the output image. Since the perceptual loss is computed to the range-compressed images, this term is responsible for forcing the neural network to synthesize textures and details while the pixel-wise loss for increasing the range of luminance in the final image. For this reason, in our experiments, we did not evaluate the result of training our method using only perceptual loss function.

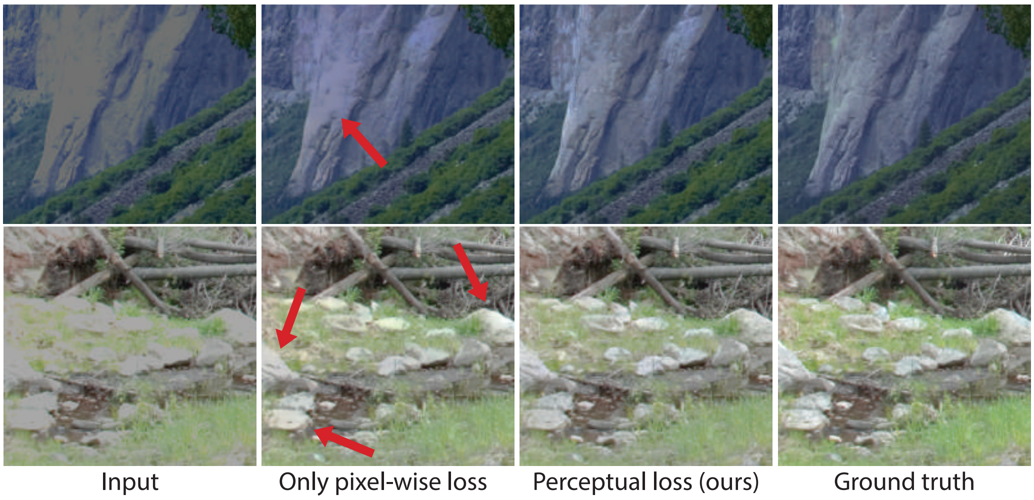

Figure 18 – We compare the results of our network trained with only a pixel-wise loss (l_1) and the proposed perceptual loss. Using the perceptual loss function, our network can synthesize visually realistic textures, while the network trained with only a pixel-wise loss produces blurry results.

4.6 FAILURE CASES

Single image HDR reconstruction is a notoriously challenging problem. Although our method can recover the luminance and hallucinate textures, it is not always able to reconstruct all the details. One of such cases is shown in Figure 19, where our approach fails to reconstruct the wrinkles on the curtain. Nevertheless, our result is still better than the other approaches as they overestimate the brightness of the window and produce blurry results. Moreover, as shown in Figure 20, when the input lacks sufficient information about the underlying texture, our method could potentially introduce patterns that do not exist in the ground truth image. Despite that, our result is still comparable to or better than the other approaches. Additionally, in some cases, our method reconstructs the saturated areas with an incorrect color, as shown in Figure 21. It is worth noting that the network reconstruct the building in blue since trees and skies are usually next to each other in the training data. As seen, other approaches also reconstruct parts of the building in blue color.

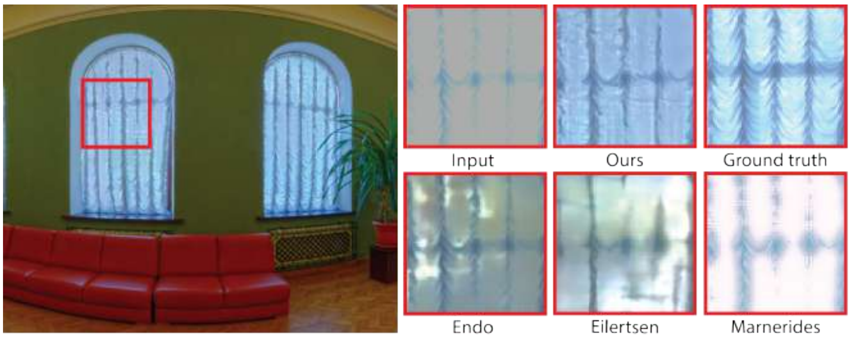

Figure 19 – Our method fails to reconstruct the wrinkles on the curtain, but it is still better than the other approaches.

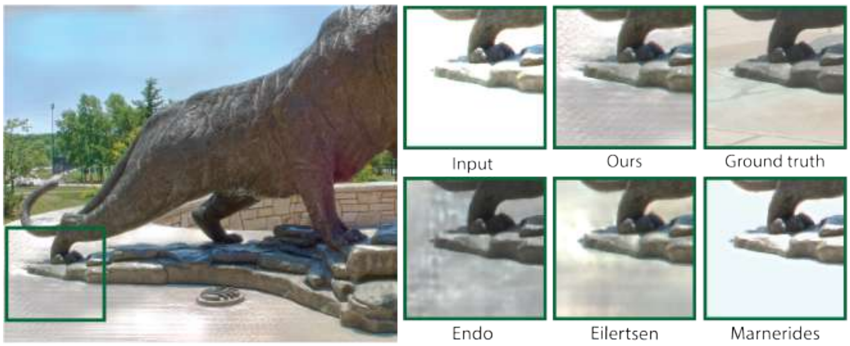

Figure 20 – Our method introduces textures that are not in the ground truth, however, our result is still comparable to or better than previous methods.

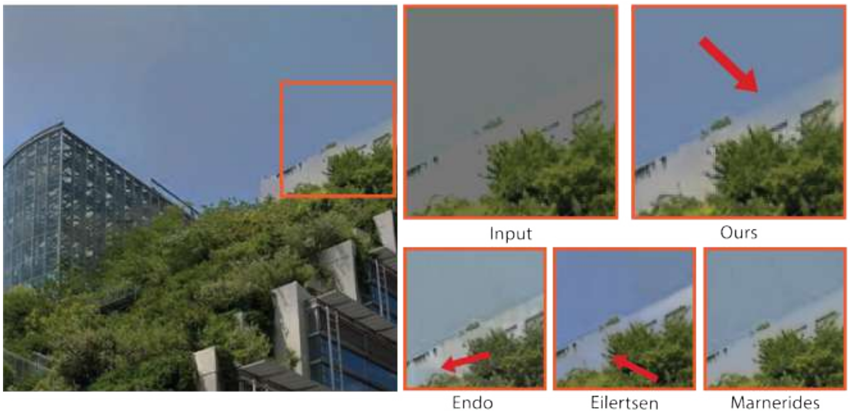

Figure 21 – In this case, our method incorrectly reconstructs the building with sky color. However, other approaches suffer from the same issue and reconstruct the building with blue color. Note that, the previous two examples are synthetic, but this one is real for which we do not have access to the ground truth image.

Chapter 5

CONCLUSIONS

In this dissertation, we presented a novel learning-based system for single image HDR reconstruction using a convolutional neural network. To alleviate the artifacts caused by conditioning the convolutional layer on the saturated pixels, we proposed a feature masking mechanism with an automatic mask updating process. We showed that this strategy reduces halo and checkerboard artifacts caused by standard convolutions. Moreover, we proposed a perceptual loss function that is designed specifically for the HDR reconstruction application. By minimizing this loss function during training, the network is able to synthesize visually realistic textures in the saturated areas. We further proposed to train the system in two stages where we pre-train the network on inpainting before fine-tuning it on HDR generation. To encourage the network to synthesize textures, we proposed a sampling strategy to select challenging patches in the HDR examples. Our model can robustly handle saturated areas and can reconstruct high-frequency details in a realistic manner. We showed quantitatively and qualitatively comparisons that our method outperforms previous methods on both synthetic and real-world images.

5.1 FUTURE WORK

We believe that learning-based approaches to HDR reconstruction have great potential for future exploration. Therefore, in this section, we list some directions for future work that we think are promising.

Temporal Coherence Although our network can be used to reconstruct an HDR video from an LDR video, our result is not temporally stable. This is mainly because we synthesize the content of every frame independently. In the future, it would be interesting to address this problem through temporal regularization (EILERTSEN; MANTIUK; UNGER, 2019; CHEN et al., 2019).

Performance We also would like to experiment with the architecture of the network to increase the efficiency of our approach and reduce the memory footprint. These optimizations could be performed by using, for instance, specific architectural building blocks (SANDLER et al., 2018; IANDOLA et al., 2016), quantization (WU et al., 2016; JACOB et al., 2018; LIN; TALATHI; ANNAPUREDDY, 2016), or network pruning (HAN et al., 2015; MOLCHANOV et al., 2016; SRINIVAS; BABU, 2015), to name a few methods.

Loss Function In this work, we did not consider generative adversarial networks (GANs) as a loss function for training our approach due to stability issues in the results and difficulty posed for training GANs. Image inpainting methods based on such generative models attempt to address these issues by proposing several changes to the generator and discriminator and with a specific formulation of the loss. However, the quality of textures produced by these approaches is still far from looking realistic. Our application is even more constrained than image inpainting (soft mask, per channel saturation, as opposed to binary mask) and, thus, using GAN is a greater challenge. Of course, this is an interesting problem to solve, but the solution is not straightforward. Based on the recent advances in both unconditional (KARRAS et al., 2020; KARRAS; LAINE; AILA, 2019) and conditional (PARK et al., 2019) image synthesizes using GANs, we believe it is worthwhile to investigate the interplay between these ideas and the HDR reconstruction.

Dataset Finally, in this work, we argue that a fundamental limitation of the HDR imaging datasets is the lack of diversity and the limited number of images available. To overcome this limitation, we train our method in two stages. In the first stage, we train our system on a large-scale set of images for the inpainting task. In the next step, we fine-tune this network on the HDR reconstruction task using a set of simulated LDR and their corresponding ground truth HDR images. We also utilize several data augmentation techniques, as purposed in previous work (EILERTSEN et al., 2017), to further regularize our model. Nevertheless, we believe that it would be beneficial for the community to purpose a large-scale HDR imaging dataset, which we leave for future work. This dataset could be employed for both benchmarking HDR image synthesis methods and training learning-based approaches.

REFERENCES

AKYÜZ, A. O. Photographically guided alignment for hdr images. In: *Eurographics (Areas Papers)*. [S.l.: s.n.], 2011. p. 73–74.

BANTERLE, F.; LEDDA, P.; DEBATTISTA, K.; CHALMERS, A. Inverse tone mapping. In: ACM. *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*. [S.l.], 2006. p. 349–356.

BAU, D.; STROBELT, H.; PEEBLES, W.; WULFF, J.; ZHOU, B.; ZHU, J.; TORRALBA, A. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, v. 38, n. 4, 2019.

BIST, C.; COZOT, R.; MADEC, G.; DUCLOUX, X. Tone expansion using lighting style aesthetics. *Computers & Graphics*, Elsevier, v. 62, p. 77–86, 2017.

CHAUDHARI, P.; SCHIRRMACHER, F.; MAIER, A.; RIESS, C.; KÖHLER, T. Merging-isp: Multi-exposure high dynamic range image signal processing. *arXiv preprint arXiv:1911.04762*, 2019.

CHEN, C.; CHEN, Q.; DO, M. N.; KOLTUN, V. Seeing motion in the dark. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 3185–3194.

CHEN, Q.; KOLTUN, V. Photographic image synthesis with cascaded refinement networks. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 1511–1520.

DEBEVEC, P. A median cut algorithm for light probe sampling. In: ACM. *ACM SIGGRAPH 2005 Posters*. [S.l.], 2005. p. 66.

DEBEVEC, P. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In: *ACM SIGGRAPH 2008 classes*. [S.l.: s.n.], 2008. p. 1–10.

DEBEVEC, P.; HAWKINS, T.; TCHOU, C.; DUIKER, H.-P.; SAROKIN, W.; SAGAR, M. Acquiring the reflectance field of a human face. In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. [S.l.: s.n.], 2000. p. 145–156.

DEBEVEC, P.; MALIK, J. Recovering high dynamic range images. In: *Proceeding of the SPIE: Image Sensors*. [S.l.: s.n.], 1997. v. 3965, p. 392–401.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2009. p. 248–255.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DIDYK, P.; MANTIUK, R.; HEIN, M.; SEIDEL, H.-P. Enhancement of bright video features for hdr displays. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum.* [S.l.], 2008. v. 27, n. 4, p. 1265–1274.

DONG, C.; LOY, C. C.; HE, K.; TANG, X. Learning a deep convolutional network for image super-resolution. In: FLEET, D.; PAJDLA, T.; SCHIELE, B.; TUYTELAARS, T. (Ed.). *Computer Vision – ECCV 2014.* Cham: Springer International Publishing, 2014. p. 184–199. ISBN 978-3-319-10593-2.

DONG, C.; LOY, C. C.; HE, K.; TANG, X. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 38, n. 2, p. 295–307, 2015.

DURAND, F.; DORSEY, J. Fast bilateral filtering for the display of high-dynamic-range images. In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques.* [S.l.: s.n.], 2002. p. 257–266.

EDEN, A.; UYTTENDAELE, M.; SZELISKI, R. Seamless image stitching of scenes with large motions and exposure differences. In: IEEE. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).* [S.l.], 2006. v. 2, p. 2498–2505.

EILERTSEN, G.; KRONANDER, J.; DENES, G.; MANTIUK, R. K.; UNGER, J. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, ACM, v. 36, n. 6, p. 178, 2017.

EILERTSEN, G.; MANTIUK, R.; UNGER, J. Single-frame regularization for temporally stable cnns. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* [S.l.: s.n.], 2019.

ENDO, Y.; KANAMORI, Y.; MITANI, J. Deep reverse tone mapping. *ACM Transactions on Graphics (TOG)*, v. 36, n. 6, p. 177–1, 2017.

FECHNER, G. Elements of psychophysics (holt, rhinehart & winston, new york). 1860.

GALLO, O.; GELFANDZ, N.; CHEN, W.-C.; TICO, M.; PULLI, K. Artifact-free high dynamic range imaging. In: IEEE. *2009 IEEE International Conference on Computational Photography (ICCP).* [S.l.], 2009. p. 1–7.

GATYS, L. A.; ECKER, A. S.; BETHGE, M. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

GATYS, L. A.; ECKER, A. S.; BETHGE, M. Image style transfer using convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* [S.l.: s.n.], 2016. p. 2414–2423.

GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS).* [S.l.: s.n.], 2010. p. 249–256.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS).* [S.l.: s.n.], 2014. p. 2672–2680.

GROSSBERG, M. D.; NAYAR, S. K. Determining the camera response from images: What is knowable? *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 25, n. 11, p. 1455–1467, 2003.

GROSSBERG, M. D.; NAYAR, S. K. What is the space of camera response functions? In: IEEE. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* [S.l.], 2003. v. 2, p. II–602.

HAFNER, D.; DEMETZ, O.; WEICKERT, J. Simultaneous hdr and optic flow computation. In: IEEE. *2014 22nd International Conference on Pattern Recognition.* [S.l.], 2014. p. 2065–2070.

HAJISHARIF, S.; KRONANDER, J.; UNGER, J. Adaptive dualiso hdr reconstruction. *EURASIP Journal on Image and Video Processing*, Springer, v. 2015, n. 1, p. 41, 2015.

HAJSHARIF, S.; KRONANDER, J.; UNGER, J. Hdr reconstruction for alternating gain (iso) sensor readout. In: *Eurographics, Strasbourg, France, April 7-11, 2014.* [S.l.: s.n.], 2014.

HAN, S.; POOL, J.; TRAN, J.; DALLY, W. Learning both weights and connections for efficient neural network. In: *Advances in neural information processing systems.* [S.l.: s.n.], 2015. p. 1135–1143.

HAN, X.; WU, Z.; HUANG, W.; SCOTT, M. R.; DAVIS, L. S. Finet: Compatible and diverse fashion image inpainting. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* [S.l.: s.n.], 2019. p. 4481–4491.

HASINOFF, S. W.; SHARLET, D.; GEISS, R.; ADAMS, A.; BARRON, J. T.; KAINZ, F.; CHEN, J.; LEVOY, M. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, ACM, v. 35, n. 6, p. 192, 2016.

HEIDE, F.; STEINBERGER, M.; TSAI, Y.-T.; ROUF, M.; PAJĄK, D.; REDDY, D.; GALLO, O.; LIU, J.; HEIDRICH, W.; EGIAZARIAN, K. et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (TOG)*, ACM New York, NY, USA, v. 33, n. 6, p. 1–13, 2014.

HEO, Y. S.; LEE, K. M.; LEE, S. U.; MOON, Y.; CHA, J. Ghost-free high dynamic range imaging. In: SPRINGER. *Asian Conference on Computer Vision.* [S.l.], 2010. p. 486–500.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *Science*, American Association for the Advancement of Science, v. 313, n. 5786, p. 504–507, 2006.

HU, J.; GALLO, O.; PULLI, K.; SUN, X. Hdr deghosting: How to deal with saturation? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* [S.l.: s.n.], 2013. p. 1163–1170.

IANDOLA, F. N.; HAN, S.; MOSKEWICZ, M. W.; ASHRAF, K.; DALLY, W. J.; KEUTZER, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

IIZUKA, S.; SIMO-SERRA, E.; ISHIKAWA, H. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, ACM New York, NY, USA, v. 35, n. 4, p. 1–11, 2016.

ISOLA, P.; ZHU, J.-Y.; ZHOU, T.; EFROS, A. A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 1125–1134.

JACOB, B.; KLIGYS, S.; CHEN, B.; ZHU, M.; TANG, M.; HOWARD, A.; ADAM, H.; KALENICHENKO, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 2704–2713.

JACOBS, K.; LOSCOS, C.; WARD, G. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, IEEE, v. 28, n. 2, p. 84–93, 2008.

JINNO, T.; OKUDA, M. Motion blur free hdr image acquisition using multiple exposures. In: IEEE. *2008 15th IEEE International Conference on Image Processing*. [S.l.], 2008. p. 1304–1307.

JOHNSON, J.; ALAHI, A.; FEI-FEI, L. Perceptual losses for real-time style transfer and super-resolution. In: SPRINGER. *European Conference on Computer Vision (ECCV)*. [S.l.], 2016. p. 694–711.

KALANTARI, N. K.; RAMAMOORTHI, R. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (TOG)*, v. 36, n. 4, p. 144–1, 2017.

KANG, S. B.; UYTTENDAELE, M.; WINDER, S.; SZELISKI, R. High dynamic range video. In: ACM. *ACM Transactions on Graphics (TOG)*. [S.l.], 2003. v. 22, n. 3, p. 319–325.

KARRAS, T.; LAINE, S.; AILA, T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 4401–4410.

KARRAS, T.; LAINE, S.; AITTALA, M.; HELLSTEN, J.; LEHTINEN, J.; AILA, T. Analyzing and improving the image quality of stylegan. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 8110–8119.

KHAN, E. A.; AKYUZ, A. O.; REINHARD, E. Ghost removal in high dynamic range images. In: IEEE. *2006 IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2006. p. 2005–2008.

KIM, S. Y.; OH, J.; KIM, M. Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. *arXiv preprint arXiv:1909.04391*, 2019.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2015.

KOVALESKI, R. P.; OLIVEIRA, M. M. High-quality reverse tone mapping for a wide range of exposures. In: IEEE. *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*. [S.l.], 2014. p. 49–56.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2012. p. 1097–1105.

KRONANDER, J.; GUSTAVSON, S.; BONNET, G.; UNGER, J. Unified hdr reconstruction from raw cfa data. In: IEEE. *IEEE international conference on computational photography (ICCP)*. [S.l.], 2013. p. 1–9.

LANDIS, H. Production-ready global illumination. *SIGGRAPH Course Notes*, sn, v. 16, n. 2002, p. 11, 2002.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LEE, S.; AN, G. H.; KANG, S.-J. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, IEEE, v. 6, p. 49913–49924, 2018.

LEE, S.; AN, G. H.; KANG, S.-J. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 596–611.

LI, X.; CHEN, Y.; JIANG, H.; ZHAO, H. Multi-exposure high dynamic range image synthesis with camera shake correction. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *AOPC 2017: 3D Measurement Technology for Intelligent Manufacturing*. [S.l.], 2017. v. 10458, p. 104580I.

LIN, D.; TALATHI, S.; ANNAPUREDDY, S. Fixed point quantization of deep convolutional networks. In: *International conference on machine learning*. [S.l.: s.n.], 2016. p. 2849–2858.

LIU, G.; REDA, F. A.; SHIH, K. J.; WANG, T.-C.; TAO, A.; CATANZARO, B. Image inpainting for irregular holes using partial convolutions. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 85–100.

LUZARDO, G.; AELTERMAN, J.; LUONG, H.; PHILIPS, W.; OCHOA, D.; ROUSSEAUX, S. Fully-automatic inverse tone mapping preserving the content creator's artistic intentions. In: IEEE. *2018 Picture Coding Symposium (PCS)*. [S.l.], 2018. p. 199–203.

MAAS, A. L.; HANNUN, A. Y.; NG, A. Y. Rectifier nonlinearities improve neural network acoustic models. In: *Proceedings of International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2013. v. 30, n. 1, p. 3.

MADDEN, B. C. Extended intensity range imaging. *Technical Reports (CIS)*, p. 248, 1993.

MANTIUK, R.; KIM, K. J.; REMPEL, A. G.; HEIDRICH, W. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics (TOG)*, ACM, v. 30, n. 4, p. 40, 2011.

MARNERIDES, D.; BASHFORD-ROGERS, T.; HATCHETT, J.; DEBATTISTA, K. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2018. v. 37, n. 2, p. 37–49.

MCGUIRE, M.; MATUSIK, W.; PFISTER, H.; CHEN, B.; HUGHES, J. F.; NAYAR, S. K. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, IEEE, v. 27, n. 2, p. 32–42, 2007.

MIN, T.-H.; PARK, R.-H.; CHANG, S. Histogram based ghost removal in high dynamic range images. In: IEEE. *2009 IEEE International Conference on Multimedia and Expo*. [S.l.], 2009. p. 530–533.

MITSUNAGA, T.; NAYAR, S. K. Radiometric self calibration. In: IEEE. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. [S.l.], 1999. v. 1, p. 374–380.

MOLCHANOV, P.; TYREE, S.; KARRAS, T.; AILA, T.; KAUTZ, J. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2010. p. 807–814.

NING, S.; XU, H.; SONG, L.; XIE, R.; ZHANG, W. Learning an inverse tone mapping network with a generative adversarial regularizer. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2018. p. 1383–1387.

OH, T.-H.; LEE, J.-Y.; TAI, Y.-W.; KWEON, I. S. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 37, n. 6, p. 1219–1232, 2014.

OORD, A. v. d.; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A.; KAVUKCUOGLU, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

OU, Y.; AMBALATHANKANDY, P.; IKEBE, M.; TAKAMAEDA, S.; MOTOMURA, M.; ASAI, T. Real-time tone mapping: A state of the art report. *arXiv preprint arXiv:2003.03074*, 2020.

PARK, T.; LIU, M.-Y.; WANG, T.-C.; ZHU, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 2337–2346.

PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L. et al. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2019. p. 8024–8035.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. Language models are unsupervised multitask learners. *OpenAI Blog*, v. 1, n. 8, p. 9, 2019.

REMPEL, A. G.; TRENTACOSTE, M.; SEETZEN, H.; YOUNG, H. D.; HEIDRICH, W.; WHITEHEAD, L.; WARD, G. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. In: ACM. *ACM Transactions on Graphics (TOG)*. [S.l.], 2007. v. 26, n. 3, p. 39.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. *International Conference on Medical Image Computing and Computer-assisted Intervention*. [S.l.], 2015. p. 234–241.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 4510–4520.

SEN, P.; KALANTARI, N. K.; YAESOUBI, M.; DARABI, S.; GOLDMAN, D. B.; SHECHTMAN, E. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics (TOG)*, v. 31, n. 6, p. 203–1, 2012.

SERRANO, A.; HEIDE, F.; GUTIERREZ, D.; WETZSTEIN, G.; MASIA, B. Convolutional sparse coding for high dynamic range imaging. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2016. v. 35, n. 2, p. 153–163.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2015.

SRINIVAS, S.; BABU, R. V. Data-free parameter pruning for deep neural networks. *arXiv preprint arXiv:1507.06149*, 2015.

TIAN, C.; XU, Y.; ZUO, W.; ZHANG, B.; FEI, L.; LIN, C.-W. Coarse-to-fine cnn for image super-resolution. *IEEE Transactions on Multimedia*, IEEE, 2020.

TOCCI, M. D.; KISER, C.; TOCCI, N.; SEN, P. A versatile hdr video production system. In: ACM. *ACM Transactions on Graphics (TOG)*. [S.l.], 2011. v. 30, n. 4, p. 41.

TOMASZEWSKA, A.; MANTIUK, R. Image registration for multi-exposure high dynamic range image acquisition. Václav Skala-UNION Agency, 2007.

TSIN, Y.; RAMESH, V.; KANADE, T. Statistical calibration of ccd imaging process. In: IEEE. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. [S.l.], 2001. v. 1, p. 480–487.

TURSUN, O. T.; AKYÜZ, A. O.; ERDEM, A.; ERDEM, E. The state of the art in hdr deghosting: A survey and evaluation. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2015. v. 34, n. 2, p. 683–707.

WANG, L.; WEI, L.-Y.; ZHOU, K.; GUO, B.; SHUM, H.-Y. High dynamic range image hallucination. In: EUROGRAPHICS ASSOCIATION. *Proceedings of the 18th Eurographics Conference on Rendering Techniques*. [S.l.], 2007. p. 321–326.

WARD, G. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of graphics tools*, Taylor & Francis, v. 8, n. 2, p. 17–30, 2003.

WU, J.; LENG, C.; WANG, Y.; HU, Q.; CHENG, J. Quantized convolutional neural networks for mobile devices. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 4820–4828.

WU, S.; XIE, S.; RAHARDJA, S.; LI, Z. A robust and fast anti-ghosting algorithm for high dynamic range imaging. In: IEEE. *2010 IEEE International Conference on Image Processing*. [S.l.], 2010. p. 397–400.

WU, S.; XU, J.; TAI, Y.-W.; TANG, C.-K. Deep high dynamic range imaging with large foreground motions. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 117–132.

XU, Y.; NING, S.; XIE, R.; SONG, L. Gan based multi-exposure inverse tone mapping. In: IEEE. *2019 IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2019. p. 1–5.

YAN, Q.; GONG, D.; SHI, Q.; HENGEL, A. v. d.; SHEN, C.; REID, I.; ZHANG, Y. Attention-guided network for ghost-free high dynamic range imaging. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 1751–1760.

YANG, C.; LU, X.; LIN, Z.; SHECHTMAN, E.; WANG, O.; LI, H. High-resolution image inpainting using multi-scale neural patch synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 6721–6729.

YANG, X.; XU, K.; SONG, Y.; ZHANG, Q.; WEI, X.; LAU, R. W. Image correction via deep reciprocating hdr transformation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2018. p. 1798–1807.

YAO, S. Robust image registration for multiple exposure high dynamic range image synthesis. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Image Processing: Algorithms and Systems IX*. [S.l.], 2011. v. 7870, p. 78700Q.

YU, J.; LIN, Z.; YANG, J.; SHEN, X.; LU, X.; HUANG, T. S. Free-form image inpainting with gated convolution. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 4471–4480.

ZHANG, J.; LALONDE, J.-F. Learning high dynamic range from outdoor panoramas. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 4519–4528.

ZHANG, R.; ISOLA, P.; EFROS, A. A. Colorful image colorization. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 649–666.

ZHAO, H.; SHI, B.; FERNANDEZ-CULL, C.; YEUNG, S.-K.; RASKAR, R. Unbounded high dynamic range photography using a modulo camera. In: IEEE. *2015 IEEE International Conference on Computational Photography (ICCP)*. [S.l.], 2015. p. 1–10.

ZHOU, B.; LAPEDRIZA, A.; XIAO, J.; TORRALBA, A.; OLIVA, A. Learning deep features for scene recognition using places database. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2014. p. 487–495.

ZIMMER, H.; BRUHN, A.; WEICKERT, J. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2011. v. 30, n. 2, p. 405–414.