Learned Exposure Selection for High Dynamic Range Image Synthesis

Shane Segal

A thesis submitted to the

Faculty of Graduate Studies

in partial fulfilment of the requirements

for the degree of

Master of Science

Graduate Program in

Electrical Engineering and Computer Science

York University

Toronto, Ontario, Canada

January, 2021

Abstract

High dynamic range (HDR) imaging is a photographic technique that captures a greater range of luminance than standard imaging techniques. Traditionally accomplished by specialized sensors, HDR images are regularly created through the fusion of multiple low dynamic range (LDR) images that can now be captured by smartphones or other consumer grade hardware. Three or more images are traditionally required to generate a well-exposed HDR image. This thesis presents a novel system for the fast synthesis of HDR images by means of exposure fusion with only two images required. Experiments show that a sufficiently trained neural network can predict a suitable exposure value for the next image to be captured, when given an initial image as input. With these images fed into the exposure fusion algorithm, a high-quality HDR image can be quickly generated.

Acknowledgements

I would like to thank my co-supervisors, Professor Michael S. Brown and Professor Marcus Brubaker. The patience and support shown to me throughout the process of researching and developing the work in this thesis, and throughout my Masters studies was immense. In spite of challenges with scheduling around the busy lives of everyone involved, I was able to learn an incredible amount from them – both in regards to technical information, and what it means to be an academic researcher. I am very grateful for the chance to study under them and learn from their expertise.

I am also grateful for the experience and advice from my friends and peers at the lab and others in EECS. Without their advice and support, my journey to get here would have been much longer and more arduous. The work ethic and love of the field they display every day is inspiring. The collaborative environment they foster in the lab is always a fun and productive environment to learn and work in.

I would also like to thank my partner, my family and my friends for the support, advice, companionship and entertainment they provided. They were always able to keep me on track and focused, and most importantly lift my spirits when needed. I truly do not know where I would be without them and their almost limitless support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

High dynamic range imaging attempts to capture a wider range of the luminance present in a scene than a non-specialized camera can typically capture. While the human eye is not directly comparable to a digital imaging sensor, the human visual system can capture a range of more than five magnitudes of brightness instantaneously, and over eight magnitudes after a short interval of adaption [1, 2], while non-specialized imaging sensors and camera systems are limited to a far smaller range [3]. This causes traditionally captured scenes to appear washed out, overly dark, or to contain regions where the scene content is lost entirely due to clipping of the pixel values beyond the range of the image sensor. In the early days of photography, the dynamic range of cameras was too small to capture the details of many common scenes. Early photographers would combine negatives using a "dodge-and-burn" method in order to generate an image that contains light and dark elements that could not be simultaneously captured using available methods [4]. Single capture HDR imaging requires sensors with much greater low-light sensitivity [5] than those seen in smartphones or consumer-grade cameras; bracketing and other multi-exposure techniques are used instead [6].

These techniques generally capture between three and five low dynamic range images across a fixed range of exposures to generate a single HDR photo through a process referred to as *exposure fusion* [7]. This thesis shows that such a large range is often not required and as little as two images may be sufficient when the exposure values are carefully selected. The proposed system utilizes a convolutional neural network to learn a function that - given an initial image with a fixed exposure value as input - outputs a prediction for the exposure setting for the second captured image that will produce the best exposure fusion result.

In many mobile camera systems, the maximum number of input images used in an exposure fusion system is limited to five, due primarily to the large time requirements of capturing each additional image, as well as diminishing returns in the quality of the fused result. This thesis also introduces a system for the fast pregeneration of HDR images and their associated statistics. This enables efficient training without the runtime costs associated with on-demand image generation via exposure fusion.

1.1 Problem statement and challenges

The goal of this thesis is to show that it is feasible and practical to generate high-quality, subjectively pleasing, tone-mapped images through exposure fusion using only two input images. One of the input images is based on the camera's auto-exposure optimized for the sensors dynamic range. This is referred to as exposure value 0 (or EV_0). The selection process for the second image is performed by a convolutional neural network given the initially captured EV_0 image as input. The network then predicts the exposure value for the second image to be captured.

While a simple system that predicts the next image to capture based purely upon the brightness or other basic statistics of the EV_0 image is possible, a more

sophisticated system is needed to make better predictions. This motivates the choice of a convolutional network, as they have been shown to have much more expressive power and can effectively learn complicated mappings between their inputs and outputs [8]. Unfortunately, it is also well-known that interpreting what a trained neural network has learned is a hard and active research area [9]. A system capable of distinguishing between different types of scenes must be used in order to determine the required EV of the second image, as the choice is highly scene dependent. While it is often challenging to discern the exact characteristic traits that are being learned by the network in order to make its predictions, it can be hypothesized that brightness, contrast and higher level semantic features are all important factors in determining the EV to be predicted.

The traditional exposure fusion process is outlined in Figure 1.1, and illustrates the length of time required to simply capture all the images at their various exposure values needed for classical exposure fusion. The end result of the HDR imaging process are images with a wider range of exposure than can be regularly captured with a traditional camera sensor. As previously mentioned, the human eye has a much higher dynamic range than a common camera can capture, and as such many photographs do not capture the full scope of the scene as experienced by human perception [1]. Common HDR imaging techniques seek to alleviate this mismatch by combining sequences of low dynamic range images into a composite high dynamic range image by means of exposure fusion and tone mapping. This commonly requires anywhere from three to five LDR images. This work introduces a system that produces high-quality, tone-mapped images via exposure fusion with only two input LDR images required.

As can be seen in Figure 1.1, three to five individual image captures must be made before an HDR image can be produced. While smartphone speeds are always

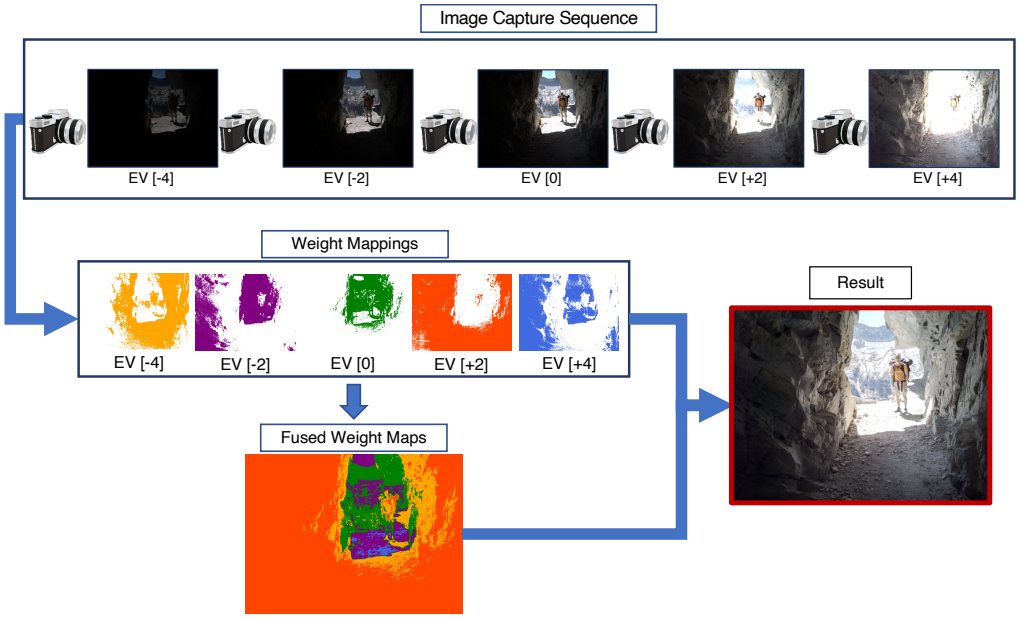

Figure 1.1: The traditional exposure fusion process requires the sequential capture of images at different exposures. As lower exposure values necessarily rely on longer exposure times on fixed aperture mobile imaging systems, the time required to capture the sequence of differently exposed images is strictly larger than the time it would take to simply capture a burst of images at a higher EV. The exposure fusion process consists of creating weight maps based on several "well-exposedness" factors, and then combining them over multiple scales using Laplacian and Gaussian Pyramids, and then performing colour correction on the final result. While the longest part of this process is the image capture, Exposure Fusion is $O(n)$ where n is the number of input images.

increasing, individual image captures have been shown to take over 1200 ms [10]. The time requirements introduce issues such as ghosting, where objects in the scene move from one captured frame to the next, causing a motion blur effect on the scene. Image registration methods are often used to align the input frames of the image, adding to the time requirements [11] before the tone-mapped image can be displayed to the user. Many scenes are simply not feasible for use with exposure fusion, as the subjects to be captured are often gone or changed entirely before the capture sequence has completed.

The experiments performed during the course of this thesis show that it is feasible to generate high quality HDR images from only two input images. This is in contrast to most systems present in mobile phones and consumer-grade cameras that commonly require three or more images. Without an ability to analyze the content of the scene, three to five images must be taken to ensure that the full dynamic range is captured.

While it is true that the system detailed in this thesis is able to produce high-quality HDR images with a wide variety of input images, it is still the case that there exist certain pathological scenes that this method would not be able to address. Consider a scene with both severely underexposed and overexposed regions. This might be an image containing a very bright outdoor area, and a very dimly lit indoor scene. To properly create an exposure fusion image of this scene, at least three images captured with different exposure values would be needed to generate a properly exposed image. However, this occurs rarely in the dataset used here, and remains uncommon in general. As such this is a situation that is not addressed in this work. While pathological scenes cannot be fully avoided, the ideas presented in these thesis are still effective and capable of producing a 60% reduction in the input images required while still producing results of comparable quality, as shown in Section 4.1.6.

In order for two images to produce an aesthetically pleasing tone-mapped and fused image, each source image must provide a suitably large set of pixels that are well-exposed in their source, and underexposed in the other. Using the high dynamic range radiance images in the HDR+ [12] dataset, a set of synthetically bracketed source images were created that span the possible exposure values. The pairwise combinations of those images were fed as input to the exposure fusion algorithm and from there it was determined that many high-quality tone-mapped and fused images

can be recreated from as few as two LDR images if they are chosen carefully. When given the initial captured image, the secondary bracketed image that will produce the highest quality tone-mapped and fused image is highly scene-dependent, and lends itself to a learning-based approach. This thesis therefore describes a system which estimates the function that predicts the best secondary image given the initially captured image as input.

1.2 Outline of Work

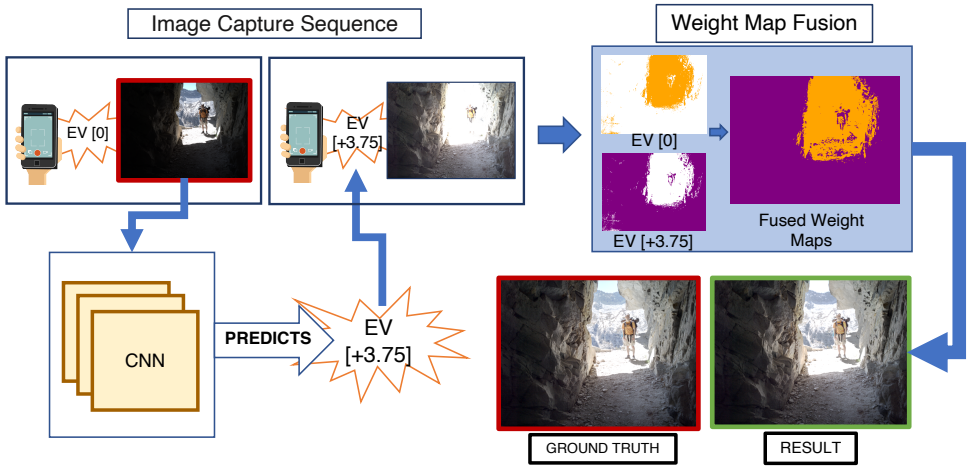

Figure 1.2: An initial image is captured at the camera-default exposure value, and then fed to the EV prediction network, that in turn returns a prediction of the best secondary EV choice for the next image to be captured. The camera then captures the secondary image, and the result is fed through the exposure fusion algorithm to produce a high quality tone-mapped image.

The system described in Figure 1.2 operates under the hypothesis that many multi-capture HDR systems are given more information than strictly necessary to generate a high quality fusion HDR image. As classical imaging systems have no easy and fast method to analyze the contents of an image, they are forced to capture more

input images than strictly required to consistently produce a perceptibly improved HDR image. The growing popularity of machine learning and its corresponding increase in mobile devices affords an opportunity to make smarter choices about the inputs to the exposure fusion algorithm in common scenarios.

The system detailed in this thesis is based on the hypothesis that many multi-capture HDR systems are over-specified in the number of separate exposure photographs that they require in order to produce a reasonable facsimile of a real HDR image. Experiments performed in Section 4.1 below show this to be the case. Why then do real-world systems require such a time-intensive image capturing process? Experiments show that it is only feasible to synthesize HDR photos from two images when scene knowledge is incorporated by means of the neural network. Without scene knowledge, selecting among possible exposure values by chance does not produce good results when restricted to the EV_0 image and an additional selection. In the presented system, this scene knowledge is incorporated by means of a pre-trained convolutional neural network that analyzes a captured image, and determines at what exposure value to capture the secondary image with.

The challenge of collecting high quality data for training machine learning algorithms for HDR applications is well documented [13, 14]. There are no large scale datasets for HDR imaging analogous to ImageNet [15], or MS-COCO [16]. As such, many authors have created compilation datasets of various small sources as seen in [17]. The HDR+ dataset described in [12] is composed of 3640 images, and are of sufficient quality for the task at hand. As the success of training deep learning systems is highly dependent upon having enough training data, the relative sparsity of available HDR datasets made HDR+ highly suitable. Additional details about the preparation of the training data are provided in Section 3.3.

Chapter 2

Background

This chapter presents a literature review of HDR synthesis algorithms and learning based approaches on LDR to HDR image generation. As the presented work in this thesis depends on convolutional neural networks and efficient inference, a brief overview of work in this area is detailed as well.

HDR imaging in general purpose photography is primarily motivated by narrowing the gap between human visual perception and the dynamic range of camera sensors. In the earliest days of photography, dynamic range was achieved through compositing separate negatives onto a final image [4]. The need for imaging systems that can capture a wider dynamic range was evident throughout the development of photographic technology, and methods to extend the dynamic range of available techniques were developed since the invention of photography [18].

Although the human visual system and cameras are too different to compare in most ways, it is commonly regarded as true that the dynamic range of the human eye is much greater than that of a camera, covering a range of more than eight orders of magnitude in luminance [2, 1] while most camera systems are only able to capture a subset of that range [3]. This thesis defines dynamic range in the context

of imaging systems as exposure value differences, where a standard difference is defined as the base-2 logarithm of the f-number of the given optical system, as seen in [19], $EV = \log_2(\frac{N^2}{t})$, where $N = f/D$, f is the focal length of the lens and D is the effective aperture of the system. The model of a camera being considered in this work is that of a smartphone camera. Smartphone cameras generally have fixed f-numbers [20], although continued advances in smartphone camera technology might challenge this assumption. As such, a difference of one EV corresponds to a doubling of the amount of light onto the image sensor.

2.1 Learning-based approaches

Machine learning has become a popular technique in recent years, across a wide variety of sub-fields within computer science and in the wider scientific and engineering community, including extensive adoption in computer vision, with many applications in computational photography, classification, semantic segmentation, and many more.

2.1.1 HDR imaging in the deep-learning regime

There have been many deep learning based approaches applied to various parts of the HDR imaging stack. It has been most commonly applied in single image HDR synthesis, where the missing luminance information is inferred and inpainted by a convolutional neural network as in [21, 22]. There has been work on fully convolutional models that aim to align and fuse LDR images into a single HDR photo, reducing ghosting and other motion artifacts using deep optical flow, e.g [23].

Work has been done on reverse tone mapping using deep neural networks [24]. The authors implement HDR image synthesis by hallucinating various LDR exposure images, and fusing them using conventional tone-mapping and image fusion

algorithms. This approach overcomes the issue of small datasets seen in other approaches by expanding the available data for training.

2.1.2 HDR inpainting

Single image HDR reconstruction methods have been developed in [21] that apply image inpainting techniques to fill in regions where clipping has occurred in the input LDR image. These techniques have improved in recent years; however they are fundamentally limited by the nature of deep image inpainting — which is to hallucinate with learned deep features the missing image information. While the hallucinated information can be perceptually pleasing, it is by definition not a true expansion of dynamic range. For this reason, this work seeks to explore methods that rely only on information that is in the original images for the final output image.

2.1.3 HDR burst photography

Of particular importance to the work presented in this thesis is the synthesis of HDR images from a sequence of burst-photographs [12]. This technique takes a sequence of successive images of similar exposure, and combines them using a modified version of the algorithm provided in [7]. There are several innovations in this work.

The authors highlight many of the traditional challenges of HDR imaging in mobile devices. The size of the aperture is limited by the small physical dimensions of the enclosing device; traditional exposure bracketing and image fusion can take long enough that ghosting becomes a significant issue, especially with the number of bracketed images required and the inherent processing constraints. They also point out that flash photography is quite unpopular among photographers for the often aesthetically upsetting effect on the colour of the scene.

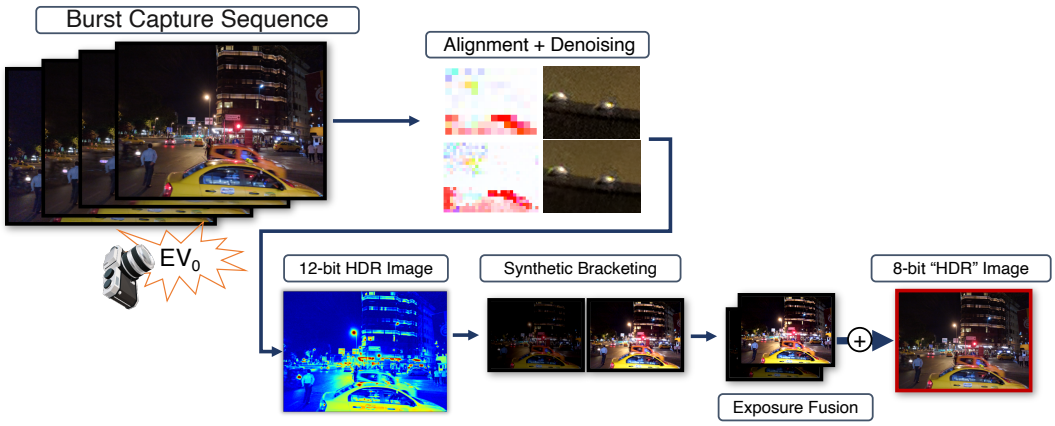

Figure 2.1: The HDR+ pipeline begins with the rapid capture of a sequence of similarly exposed images. These images are then aligned and merged before denoising takes place. The multiple captures of a similar scene allow for greater noise reduction and a higher bit depth that can then be exploited to generate a tone-mapped HDR image. After alignment and denoising, the merged frame is synthetically bracketed into two images through gamma adjustments, and then fed through a modified version of the Exposure Fusion algorithm. Images adapted from [12].

They address these problems by means of burst mode photography, as seen in Figure 2.1. These burst images are captured at the same exposure value, and aligned and merged into a reference frame with high dynamic range. The input burst images are purposely underexposed so as not to clip any regions. An individual underexposed image would be quite noisy in darker regions, but they are able to be denoised effectively by acting on small, merged regions of the image. Each separate region to be denoised is modelled as an individual signal-independent noise model, with the variance of the noise estimated through the root mean squared error of the tile.

To compress the dynamic range of the merged and aligned image in order to make it suitable for use on standard displays, they make use of a modified local tone mapping operation called exposure fusion [7]. They create synthetic exposures from

the intermediate HDR image by means of gain and gamma correction (although the exact steps that make up this process are not elaborated on by the authors), and then fuse them as if they were traditionally bracketed images. This results in images that have noticeably better exposure in the dark and bright regions.

2.2 Tone mapping

Digital Photography employs tone-mapping procedures that map images with a high dynamic range into a smaller range space that is more suitable for use on a display with limited range. Tone-mapping operators are available that optimize for different tasks; producing an aesthetically pleasing image, or reproducing as much detail from the source image in the tone-mapped image as possible [25].

Early examples of tone mapping date back to the 1970's and the Retinex algorithm [26] in the context of understanding colour constancy in the human visual system, in which the authors describe an algorithm for exposure fusion and tone mapping in terms of an electro-mechanical circuit. As digital photography became more common, approaches using more sophisticated methods such as Poisson editing [27] became common in the early 2000's [28]. This method works by efficiently solving the Poisson equation over a modified gradient field of the luminance.

Tone mapping is usually done through either global or local methods. Global tone-mapping operators work through application of a spatially invariant function that does not consider local neighbourhoods in the transformation of the image, yet are conceptually simple and computationally cheap. Local tone-mapping operations include [29, 30], which use neighbourhood-based methods in order to selectively adapt the output dependent on the local luminance or colour values. By contrast, these methods are much more computationally expensive yet produce perceptually better results, especially in more complicated scenes[31]. While Tone Mapping is

roughly divided into local and global methods, recent advances in computer vision and machine learning have muddied the distinction [24] by introducing a "reverse tone-mapping operator" discussed previously in 2.1.1 that estimates the true HDR image from a single LDR source.

2.3 Exposure fusion

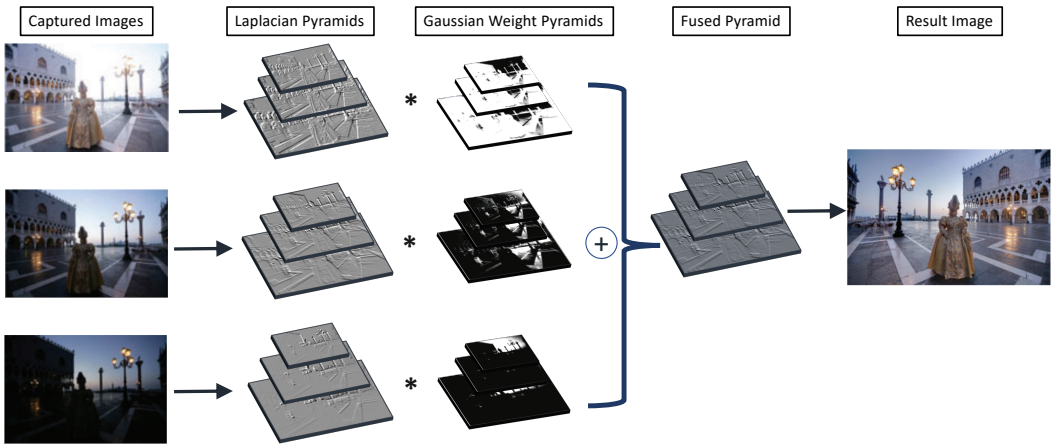

Figure 2.2: The successive steps of exposure fusion are: capturing of the bracketed input photos, processing into weight maps, and then fusion through a Laplacian Pyramid. Images adapted from [7].

Exposure fusion is an algorithm that is highly related to tone-mapping, although has slight differences that prevent it from being included in the category. The algorithm was introduced by [7]. Exposure fusion is simpler in some ways than many tone mapping algorithms. It skips the steps of computing an intermediate HDR image and then tone-mapping, and instead fuses the input LDR images directly into a new LDR image composed of the well exposed regions of each input. The

result is a high quality image that appears to have been tone-mapped from an HDR input.

The work in [7] defines a set of quality measures that they use to select among the n input pixels from each of the n LDR input images. The quality measures were designed to encourage well-exposed, bright and colourful regions in the resulting image. These quality measures are used to form weight maps, where pixels that are present in the map must appear in the final image. These weight maps and their pyramidal representations can be found in Figure 2.2.

To encourage higher weighting around higher contrast regions, they apply a LoG (Laplacian of Gaussian) filter to a grey-scaled version of each input image, and take the magnitude of the filter response as a measure of contrast.

Highly saturated colours are commonly seen as desirable, and when a longer exposure time is used during the capturing process, colours can become desaturated and undergo clipping. To form a measure for saturation, the authors take the pixel-wise standard deviation across all colour channels.

The final quality measure used in exposure fusion is well-exposedness. As the intensity value of each pixel ranges from 0 to 1, pixels with intensity values near either end of the range are either under or over-exposed and might have undergone clipping at the high end of the dynamic range or increased noise in the lower end. The pixels in each channel of the image are assumed to be normally distributed with $\sigma = 0.2$, $\mu = 0.5$ and as such the well-exposedness is defined as the probability of seeing the given pixel value given a normally distributed set of pixels. As the measure deals with colour images, this measure is implemented as the expectation across the colour channels.

These various quality measures are combined using an exponential weighting function W as follows:

$$W_{i,j,k} = (C_{i,jk})_C^{\omega} \times (S_{i,jk})_S^{\omega} \times (E_{i,j,k})_E^{\omega},$$

where C, S, E are the measures for contrast, saturation, and exposure respectively, and $\omega_{[C,S,E]}$ are the corresponding exponential weights.

In order to fuse the input images, weight maps are constructed for each input image using W. They are normalized such that the channel-wise sum is 1 at each pixel location, denoted as $\hat{W}$. The final image is seen as the weighted sum of the input images according to:

$$R_{i,j} = \sum_{k=1}^{N} \hat{W}_{i,jk} I_{i,jk},$$

for N input images.

To minimize seams between patches from different source images, the images are decomposed into a Laplacian pyramid [32], where each level contains the successive difference image of each blurred and scaled image according to its level in the pyramid. Blending occurs at each level of the pyramid, which naturally smoothes seams and keeps natural edges as a natural edge is generally the last distinct change in intensity at the higher levels of the pyramid.

The weight maps generated from each separately exposed image contribute in varying amounts to the final output image. This can be seen in Figure 2.3, where each colour in the fused output image can be traced directly back to the weight map of the same colour.

Exposure fusion has been widely used in several applications, and was chosen for use in both the HDR+ work [12] and the work presented herein, as compared to other similar algorithms [33, 29] that take several LDR images and produce an HDR image, less camera information is required during computation and the visual

results are subjectively superior, as well as not requiring an additional tone-mapping procedure, since exposure fusion operates purely on LDR images. Subjectively, images produced with methods from [33, 29] are of lower quality, with less realistic colours.

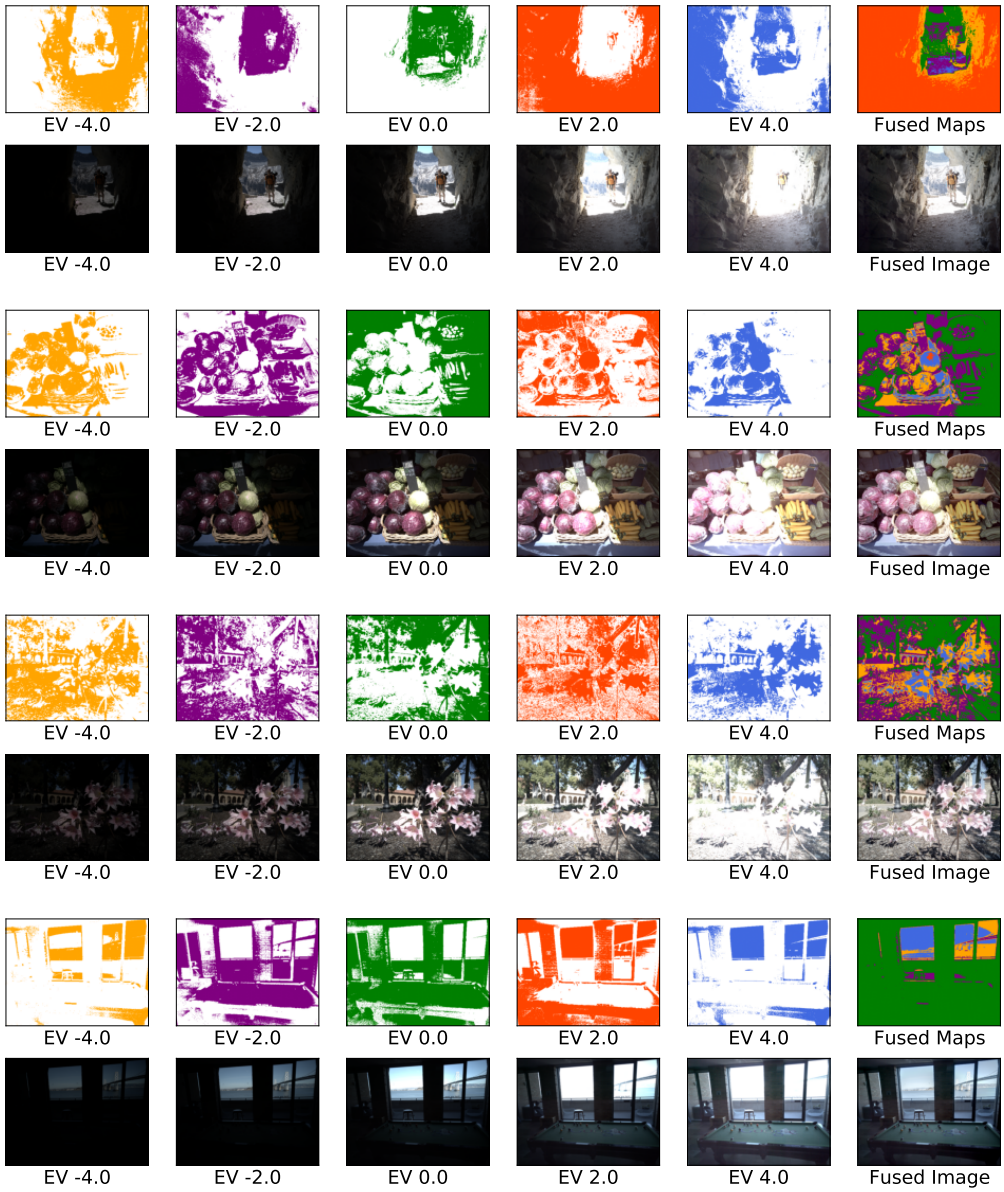

Figure 2.3: Each coloured image is a representation of the weight maps that are used to fuse the various exposure-valued images. Notice that many of the images are largely comprised of source pixels from two images. Each weight map contributes a fractional value to the final image in the Exposure Fusion algorithm, but are shown in a binarized form here for ease of illustration. The fused image components show where each pixel in the final image comes from in the source images.

Chapter 3

Methodology

The technical goal of this thesis is to show how to create a system that learns to estimate exposure value that will generate the highest quality tone-mapped HDR image when combined with an initially captured image of the same scene. The objective function is initially framed as a reconstruction loss. The reconstruction loss is computed between the ground-truth image and the output of the exposure fusion function given the baseline EV_0 image and an additional image with a predicted exposure value. However, the prediction of the exposure value is dependent upon the exposure fusion function, which is non-differentiable. As such, the problem is re-framed in terms of classification, where the correct class is the EV that produces the best two-image reconstruction as compared to the ground-truth. This allows for an easily differentiable objective function, and provides the opportunity for efficiency gains in training time. Taking advantage of the opportunity provided by this shift, the vast majority of the data and statistics needed for training are computed ahead of time, and reused for training different iterations of the system.

3.1 Reconstruction error

The highest quality is defined as the image that most closely matches the gold-standard image of the same scene, which is the image generated from the exposure fusion of five images, matching the greatest number of input images commonly seen in existing exposure fusion systems. The gold-standard image generation is explained in more depth in section 3.3.2.

To find the best second exposure image that will produce an exposure-fused image closest to the gold-standard ground-truth image, a reconstruction error between the gold-standard and reconstructed image is used as the objective function to be minimized. The reconstruction error, R, is formulated in terms of the mean squared error between the gold-standard image, $\mathbf{X}$, and the image generated from the predicted exposure values, $F(\mathbf{I}_0, \mathbf{I}_\varepsilon)$:

$$R(\mathbf{X}, \varepsilon) = \|\mathbf{X} - F\left(\mathbf{I}_0, \mathbf{I}_\varepsilon\right)\|^2, \tag{3.1}$$

where F is the exposure fusion function that takes as input a number of images, $\mathbf{I}_\varepsilon$, with exposure values $\varepsilon \in EV = \{-6, -5.75, \ldots, 5.75, 6\}$. R is computed for each colour channel, with the final result being the mean over the channels. An image with EV ε is denoted as $\mathbf{I}_\varepsilon$, with e.g $\mathbf{I}_0$ being the EV_0 image. The exposure value, ε, is predicted by the neural network, given the initial EV_0 image, $\mathbf{I}_0$, as input. This reconstruction error is minimized through the selection of an exposure value, ε that produces the fused image, $F(\mathbf{I}_0, \mathbf{I}_\varepsilon)$, that is closest to the ground truth as measured by the root mean squared error. The system chosen to select this scalar value, ε, is a convolutional neural network, trained to minimize the reconstruction error, R. The network outputs the probabilities of each possible exposure value, and the prediction of ε is taken as the one associated with the largest probability.

For any given ground-truth image, $\mathbf{X}$, the reconstruction error is determined solely by the chosen exposure value, ε, of the secondary image.

The objective function for the network then becomes the minimization of the reconstruction error, which attempts to predict the optimal exposure value, ε^*, for a given image:

$$\varepsilon^*(\mathbf{I}_0) = \text{argmin}_{\varepsilon \in EV} \|\mathbf{X} - F(\mathbf{I}_0, \mathbf{I}_\varepsilon)\|^2. \tag{3.2}$$

The synthesis of the gold-standard image, $\mathbf{X}$, as well as the predicted image, $F(\mathbf{I}_0, \mathbf{I}_\varepsilon)$, are computationally expensive. While $\mathbf{X}$ can be cached for reuse with every iteration over the exposure values of a given image, the predicted image, $F(\mathbf{I}_0, \mathbf{I}_\varepsilon)$, is determined by both the baseline EV_0 image and the predicted EV, ε. This is expensive to compute during the training process as the exposure fusion images are generated on-demand. The primary method explored to alleviate this time constraint was to employ a pregeneration scheme for the different gold-standard images and the various combinations of exposure fused images, using the scheme discussed in Section 3.3.

3.2 Classification error

The time constraints involved in computing the reconstruction based loss directly make it difficult to optimize directly. Instead, a classification based scheme is employed as a proxy for the reconstruction error, where the network is trained to predict the correct class as the optimal EV, $\varepsilon^*(\mathbf{X})$.

Notice that there is a finite and discrete number of possible image exposures, and that the total number of these combinations is small enough that it is reasonable to generate all of them and their associated statistics in advance. An initial choice

for the output of the network is simply to predict the exposure value, ε. While this seems reasonable at first glance, the exposure values are discrete elements and as such the network can not differentiate through them. Instead, a scheme is devised to encode each exposure value as a separate class, where each class represents the distance in reconstruction space from the optimal fused image with a given fused image, $F(\mathbf{I}_0, \mathbf{I}_\varepsilon)$.

As an example, consider the case with one source HDR image and the associated two-image exposure fusion-reconstructions. For exposure values $[-2, -1, 1, 2]$, the corresponding root mean squared errors are 19.75, 16.32, 10.37, and 14.91.

Clearly, the optimal EV choice in this case is EV+1, as that corresponds to the lowest reconstruction error. This is considered the true class for the image. The optimal EV choice is determined for each sample in the dataset and stored for future reuse.

During training, the network emits a prediction for the optimal class, which is then scored by the negative log-likelihood function:

$$\mathcal{L}(\varepsilon) = -\log P\left(\varepsilon = \varepsilon^*\left(\mathbf{I}_0\right) \mid \mathbf{I}_0; \mathbf{X}\right). \tag{3.3}$$

By reformulating the reconstruction loss as a classification problem, and using the negative log-likelihood function to score it, the induced loss landscape has some differences from that of the initial objective function. However, experimental results show that this objective function leads to learning and generalizes well. While the classification loss performs well and is easier to train, the reconstruction loss is the most natural formulation as the exposure values that are being predicted are not truly separate classes, but are instead points along a continuum defined by the aperture and shutter speed of the given imaging system.

3.3 Data generation and augmentation

The HDR+[12] dataset contains 3640 sequences of burst-captured images, along with an associated linear-raw DNG file of the fused and aligned frames, as well as a final LDR image with additional colour and white-balance corrections that make up the final portion of their pipeline. For the purpose of this thesis, the final image as well as the input burst frames are all discarded. The DNG images have a high dynamic range, with a bit depth slightly below 12, which is below the full dynamic range that could be seen from a specialist camera or other methods, but still provides a wide enough dynamic range for synthetic exposure image generation.

3.3.1 Synthetic exposure image generation

The aligned and merged DNG source files are used to generate both the ground-truth gold-standard HDR images and the LDR input images of varying exposure values. The gold-standard images are referred to as such because they are the highest quality HDR images that can be generated from the input LDR images. The processes for generating the gold-standard and input images are similar in most aspects. In order to generate the input LDR images from a given source file, a constant exposure is assumed, and as the HDR+ images are in a linear RAW space, the synthetic exposures are easily calculated as:

$$E_{\mathbf{S}} = \{\mathbf{S} \cdot 2^{\varepsilon - b} | \varepsilon \in EV\} \tag{3.4}$$

$$b = \log_2 \frac{N^2}{t}, \tag{3.5}$$

where E_S is the set of synthetic exposure images, S is the source image, EV is the set of candidate synthetic exposure values, N is the f-number, t is the exposure time,

and b is the baseline exposure of each image, used to ensure that the EV_0 as defined by the software camera pipeline is correct, despite malformed metadata present in the HDR+ dataset.

The range of EVs chosen is from $[-6, \ldots, +6]$. Wider ranges were considered, but not used as the physical plausibility of such exposure values is slim, and experimental results showed that images outside this range are extremely unlikely to contribute towards a well-exposed output image. In addition, the 12-bit range of the raw images in the dataset limit exposure values beyond this range.

The intermediate DNG files are captured with inconsistent metadata, a problem that has been observed with images captured using the commercial version of the HDR+ burst capture process as found in the Google Nexus 6P smartphone [12]. Part of the missing information is an inconsistent measure of exposure time. As all source images were captured using a burst mode where each burst takes a constant amount of time, and are then aligned and fused using the same process, this thesis assumes that the adjustment by subtracting the baseline exposure value from each image is valid, which is born out by increased accuracy in experiments detailed in Chapter 4. While EV_0 is always a relative measure, the software camera pipeline used in this work determines the EV_0 settings from the metadata associated with the raw images, which has shown itself to be inconsistent.

As can be seen in Figure 3.1, the distribution of EVs in the HDR+ dataset is heavily skewed to the negative end of the spectrum and has no exposure values in the $1 - 3$ range. This is not representative of most images, and was potentially introducing a source of bias in the data. To correct this, the distance from EV0.0 was calculated using (3.4), and then added to the image in order to have a properly exposed DNG image from which to generate the synthetic images.

As can be seen in Figure 3.2, the corrected images show a more balanced level of

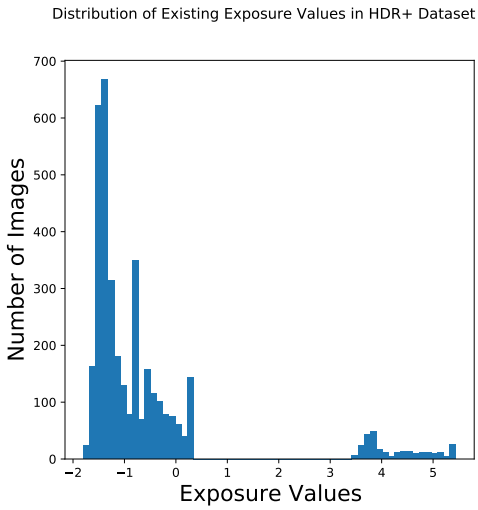

Figure 3.1: The EV distribution of the HDR+ Dataset, showing malformed metadata. Notice that the large majority of exposure values are clustered around -1.5, with a small number of very high exposure values present in the data.

exposure. There exist outliers in terms of calculated EV, but these were determined to be caused by faulty or incomplete metadata, or in the case of several of the outliers, from being computer-generated synthetic images used for testing the original HDR+ algorithm. These were pruned from the dataset as being unrepresentative of the kinds of images that are the focus of this work.

3.3.2 Gold standard image generation

In order to implement a reasonable supervised loss for training of the convolutional neural network, an attainable ground-truth image is necessary. The original HDR image is not a suitable candidate as it represents an unattainable goal for an exposure fusion system to match. The synthetic bracketing process that is used to generate the input images with various exposure values is inherently lossy, as the process takes an HDR image with a bit depth of 12 to several 8-bit sRGB images. As such, there would be a permanent margin between the best possible solutions in the dual

Figure 3.2: A selection of images showing the calculated EV present in the RAW file on the left, and then the adjusted EV_0 image in the middle.

image HDR scenario explored here and the original HDR image.

A possible choice for the ground-truth images is to simply use the final image in the pipeline from [12], but this would not be conducive to learning as that pipeline implements several colour adjustment procedures outside the scope of this thesis. Another obvious choice is to use a standard method for rendering the DNG file into a LDR image, such as that found in [34]. Unfortunately, this does not provide a suitable ground-truth either, as the tone-mapping procedure operates in a framework that assumes a full-bit depth image, as opposed to the Exposure Fusion algorithm by [7], which by design assumes discrete exposure-valued source images.

While it is unclear if images produced through tone-mapping via [34] will always be perceptually improved over those produced by Exposure Fusion, the ground-truth image should at least in principal be obtainable through the same methods as the two-photo HDR images produced by the method described here. This leads to the design choice of producing the gold-standard images through the exposure fusion of five synthetic exposure images generated as in Subsection 3.3.1.

To produce the gold standard ground-truth images, it suffices to choose five synthetic exposure images generated during the process show in Subsection 3.3.1. Five images are chosen as it is common in some exposure fusion systems to choose anywhere from three to five images [7], and they are centred at the origin. In this case, the synthetic exposures chosen have values: $[-3, -1, 0, 1, 3]$, and as such the ground-truth image is simply:

$$\mathbf{X} = F(\mathbf{I}_{-3}, \mathbf{I}_{-1}, \mathbf{I}_0, \mathbf{I}_1, \mathbf{I}_3). \tag{3.6}$$

3.3.3 Training Image Combinatorial Generation

Due to the computational requirements of fusing images during training, a scheme for pre-generation of the fused input images and associated statistics was devised. As the dataset used here was small enough that a brute-force data generation step would take less than 12 hours of computation time and occur only once, a sufficient set of statistics to compute during the generation phase was determined. All statistics that can be generated cheaply on-demand during training were left out of this step.

The statistics that were calculated ahead of time include the root mean squared error (RMSE), the peak signal to noise ratio (PSNR), the structural similarity index (SSIM), and the learned perceptual image path similarity (LPIPS) metric [35]. The RMSE was chosen as it corresponds directly to the reconstruction error used in the online version of the training algorithm. PSNR is closely related to the RMSE, and is a robust estimator for image reconstruction quality, as well as being a commonly used metric in image reconstruction tasks. Similarly, the SSIM aims to provide a reconstruction metric that is more closely aligned to human perception than a simpler metric like the PSNR. LPIPS was chosen as it shows greater performance as a perceptual metric for image similarity than both SSIM and PSNR [35], and works by utilizing layers of a neural network pretrained on ImageNet [15] as a judge of perceptual similarity. While RMSE is not generally considered the best metric for perceptual image quality, its use as the scoring metric for performance of the network is considered valid, as experiments showed that performance was highly correlated between metrics (see Subsection 4.1.6).

In order to efficiently generate the images and associated statistics, a highly parallel system was employed to generate all combinations of the input images and corresponding statistics.

For any given source HDR image, $\mathbf{S}$, the output data, $D(\mathbf{S})$, is a tuple of the set

of all exposure images, $E_{\mathbf{S}}$, the gold-standard or ground-truth image $\mathbf{X}$, the set of all possible 2-image inputs to the exposure fusion algorithm $F_{E\times E}$, and the statistics and metadata for each image, M:

$$D(\mathbf{S}) = (\mathbf{E_S}, \mathbf{X}, F_{E\times E}, M). \tag{3.7}$$

This necessarily generates a large volume of data, but allows for easy and fast computation of additional statistics after the fact, and allows for comparison with oracle methods that can be used as a baseline to measure against.

3.4 Summary of data generation

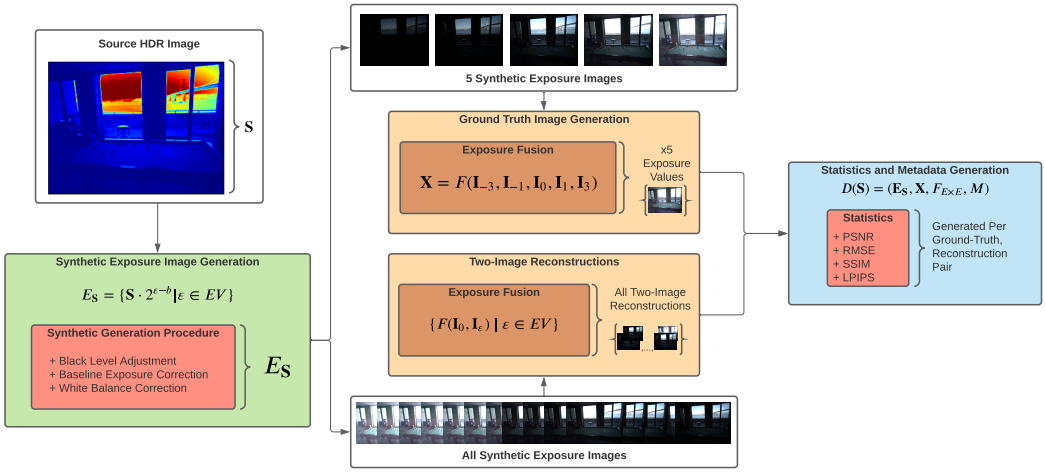

Figure 3.3: This system diagram details the different stages of the data generation algorithm. The synthetically exposed images are used for the generation of the ground-truth images and the two-image fusions. The associated statistics are generated between each ground-truth image and the corresponding two-image reconstructions.

In summary, the data generation system as described in Section 3.3 is composed of several key steps, and diagrammed in Figure 3.3. The end goal of the system is to produce the ground-truth images and the associated statistics for each of the

possible two-input exposure fusion reconstructions, D. As such, the requirements graph for D is constructed and computed on demand in a distributed fashion.

Each source image, $\mathbf{S}$, has an associated set of synthetically exposed images, $E_{(\mathbf{S})}$, a ground-truth image, $\mathbf{X}$, and metadata and statistics, M. As each ground-truth image depends on five synthetically exposed images, they are generated on-demand as needed, and saved for later re-use in computing the statistics of each image. The ground-truth image is then generated by applying the exposure fusion function, F to the required synthetically exposed images.

The reconstruction images are generated in an almost identical process, where the required inputs are simply the synthetically exposed images from:

$$\{F(\mathbf{I}_0, \mathbf{I}_\varepsilon) | \varepsilon \in EV\}. \tag{3.8}$$

The associated statistics are then computed between the ground-truth image and each of the reconstructed images.

To increase efficiency and avoid difficulties related to contested memory and program resources, the full set of source images are batched and dispatched across CPU cores and GPU's where appropriate CUDA or OpenCL implementations exist, as is the case for the statistics functions. The computation of the statistics functions are also batched in order to maximize bandwidth when transferring data to and from the GPU's.

Chapter 4

Experiments

To evaluate the performance of the system developed, a series of experiments were performed, including training with different errors to investigate the best performing metrics on this task, as well as different network architectures. Due to the choice to pre-generate the data instead of computing what was needed on demand during training, options for comparison to an oracle method were made available. Models were trained with the reconstruction error being computed as the RMSE, the PSNR, the SSIM, and LPIPS as described by [35]. Of the initial 3640 source HDR images in the HDR+ dataset, 20% (728 images) were randomly selected and reserved for the test set, and a further 10% of the 2912 images were reserved as a validation set used during the training process.

While subjective experiments on human participants were not conducted, it is an interesting topic for future research. A potential experimental setup for this would be recording the stated preferences of the test subjects in a double blind manner, where both the interviewer and subject would be unaware of the true labels of the presented pair of ground-truth and fused images.

Baseline comparisons were made against several simpler methods as checks that

the performance achieved by this system is at a high enough level to justify the additional complexity of the implementation, and in the case of the oracle method, the kind of performance that could be achieved in an impractical but optimal scenario where both the initial exposure and the secondary exposure could be selected out of all possible options — a scenario that when implemented on a physical system with time constraints would clearly undermine the goal of a practical and fast HDR system. Nonetheless, the oracle method provides a useful and interesting comparison against the fully trained network.

Additional baselines were created with the goal of simulating systems currently in use in consumer cameras. One such is picking a pair of images with an EV E, and then using the $[-E, 0, +E]$ tuple as input to exposure fusion. Another simple baseline measure is simply picking a single EV image regardless of the input. This is denoted as "FixedChoice" in Table 4.3. Slightly more complex is the histogram baseline, which is simply a fully connected network fed with the histogram of the EV_0 image as input.

In Chapter 3, the motivation for choosing the classification based loss function over a purely reconstruction loss was described. Justifying experiments for this choice will be described below. The reconstruction based model was unable to match the performance of the classification based error, in terms of top-k accuracy, training time, and complexity. Variations on the reconstruction error were also explored.

Several training optimization techniques were tested for efficacy, along with hyper-parameter optimizations. With each of the different network architectures, dropout[36] probabilities were adjusted, batch normalization was added, and in cases where the network was collapsing the number of parameters near the end, additional fully connected layers were added.

All of the following models were assessed with the RMSE, PSNR, SSIM, and LPIPS metrics [35]. Each metric was recorded per image, and the mean and standard deviation of those results is reported in Table 4.1.

4.1 Architecture experimentation and evaluation

During development, different architectures were considered and tested. A very commonly used backbone for vision problems is the ResNet architecture [37], and was one of the first options considered. The ResNet architecture can be configured with various numbers of residual layers, and is commonly used in variants with 18, 34, 50, 101, and 152 hidden layers. Even the smallest of the ResNet models is significantly larger than the others models in use here. A compelling motivation for the system presented here is the fast inference time and the potential for use on mobile devices, necessitating much smaller models than ResNet. In that vein, both MobileNet [38, 39] and SqueezeNet [40] were considered. A very simple convolutional architecture was also implemented and tested as another simple comparison against the more sophisticated backbones. Each network must output a probability; thus a softmax layer is applied at the end of each network architecture. The remainder of the subsection considers ResNet, SqueezeNet, MobileNet, and the reconstruction model results in detail, and the performance of the various models considered can be seen in Table 4.1.

4.1.1 ResNet

The ResNet architecture was first introduced in 2015 by [37]. The residual skip connections introduced allowed for deeper models to be trained more efficiently by requiring only the residuals between layers to be propagated through the network, leading to smaller gradient adjustments needed in order to update weights deeper

Metric	ResNet	SqueezeNet	MobileNet	Reconstr.
RMSE ↓	10.86 ± 6.02	12.30 ± 8.67	11.90 ± 5.47	15.42 ± 4.95
PSNR ↑	28.88 ± 5.50	28.24 ± 6.01	27.42 ± 3.66	24.80 ± 2.75
SSIM ↑	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.96 ± 0.02
LPIPS ↓	0.04 ± 0.03	0.05 ± 0.03	0.06 ± 0.03	0.05 ± 0.02
Model Size (MB)	140.20	8.51	35.37	11.91

Table 4.1: This table shows the performance of the ResNet, SqueezeNet, MobileNet, and Reconstruction models, as well as the associated model sizes in megabytes. The results are averaged across the test dataset and are reported as: score ± standard deviation. ↑ and ↓ indicate that higher and lower scares are preferred for that metric respectively. Notice that while the ResNet model slightly outperforms the MobileNet and SqueezeNet models, the size of the model reduces the practicality in the proposed use cases of mobile camera systems. The SqueezeNet model has a slightly higher RMSE than MobileNet, and the precision of the results with MobileNet is much higher when comparing the standard deviation.

in the model. This helps to alleviate issues such as that of vanishing gradients, where the gradient update at some time steps is so small that it can be lost among floating point errors, leading to little reduction in the loss function. The ResNet models introduced a higher level of performance and accuracy than seen previously in computer vision models.

This increase in accuracy came at a cost of large model weight and an increase in training time. For the models trained during these experiments, a ResNet-based model was approximately four times larger than others, with a total weight size of 147 MB. The ResNet model slightly outperformed the SqueezeNet and MobileNet models in terms of RMSE and PSNR, although the variance of results was slightly larger.

While the model was clearly learning, the statistics shown in Table 4.1 did not justify the increase in training time and model size. Notice that the RMSE is slightly lower than the RMSE of the SqueezeNet model seen in Table 4.1.

4.1.2 SqueezeNet

The SqueezeNet architecture was introduced by [40] as a network architecture that claims to have AlexNet[41]-level performance in 5% of the model size. This provides immediate benefits to mobile devices and other power limited computers in terms of inference time. With an impressively small model size of under 9 MB, SqueezeNet is extremely well suited to mobile devices.

The SqueezeNet authors introduce "Fire Modules" as an improvement and alternative to the "Inception Modules" from [42]. The Fire modules work by a series of "squeeze-expand" steps, where a set of 1×1 convolutions is applied before a nonlinearity as the squeeze phase, and then expanded by a set of 3×3 convolutional filters. The application of 1×1 filters helps to reduce the parameter count. As well, downsampling only occurs later in the network, with the goal of obtaining larger activation maps from convolutional layers nearer to the output, in an attempt to increase model accuracy.

As can be seen in Table 4.1, the much smaller SqueezeNet performs similarly or better than ResNet, with a total model size of around one tenth the size.

4.1.3 MobileNet

MobileNet[38] is an architecture that is designed around mobile computing, with a focus on computational and power efficiency requirements. MobileNets use depth-wise separable convolutions in order to reduce the total parameter count while retaining reasonable expressiveness in the network. The depth-wise convolutions are much more efficient than standard convolutions, but do not combine input channels, meaning that another operation is required in order to generate new features from across the spatial domain. As such, the depth-wise convolutions are combined with an additional 1×1 convolution that computes a linear combination of the depth-wise

convolutions in order to generate the same features that would be computed by a standard convolution. This combination of depth-wise and 1×1 convolutions lead to highly efficient networks. In addition, the network is designed to specifically take advantage of highly efficient low-level matrix multiplication algorithms referred to as GEMM, found in software such as LAPACK [43].

In the experiments performed here, the MobileNetV2 [39] architectures is used. This architecture offers several benefits over the initial MobileNet architecture. Linear Bottlenecks are introduced to further reduce the dimensionality of layers of the network and therefore the total parameter count. Residual skip connections were also introduced, in order to improve the information flow of the network. The best performing architecture in the experiments performed in this thesis was the MobileNetV2 architecture, in both accuracy and in the ratio of overall model size to performance.

As can be seen in Table 4.1, the accuracy of the MobileNetV2 model is outperformed slightly by SqueezeNet in terms of overall reconstruction error, and also outperforms or matches ResNet, even though ResNet makes use of a far greater number of parameters. However, MobileNet performs better in terms of Top-K classification accuracy, shown in Table 4.2.

4.1.4 Reconstruction Model

In order to test a version of the model that uses a purely reconstruction based loss instead of the classification function defined in (3.2), the MobileNetV2 architecture was trained with the reconstruction based loss. This model uses the same MobileNet architecture as in Section 4.1.3, but with a reconstruction based loss instead. This model utilizes pregeneration of the various exposure fusion images, but computes the reconstruction error on demand, instead of using the classes representing distance

K	Reconstruction	ResNet-18	MobileNetV2	SqueezeNet
1	4.70%	44.89%	41.85%	44.75%
2	8.29%	70.17%	66.99%	64.50%
3	13.67%	80.80%	77.76%	75.41%
4	18.51%	86.46%	86.60%	81.63%
5	23.20%	90.06%	90.88%	85.36%
6	29.56%	92.54%	93.51%	86.05%
7	38.12%	95.03%	95.17%	87.15%

Table 4.2: The Top-K errors of various models compared. The reconstruction based model does not perform as well as might be expected. Instead, the various classification-based models approach a consistently higher baseline accuracy.

from the optimal exposure fusion image synthesis.

This produces some benefits in terms of the simplicity of the model, but the performance falls short of the classification based models. This is possibly due to the higher computational efficiency and ease of training the classification based models, but the exact reason is unknown. While the model is outperformed in terms of the RMSE by the other models in Table 4.1 and the various baseline measures, it does achieve a higher PSNR and lower Perceptual Loss than the conventional methods, as seen in Table 4.3.

4.1.5 Accuracy

The Top-k accuracies of the different models shows similar outcomes to the evaluations in Section 4.1. ResNet performs very well, yet the massive model size and training time requirements make the competing efficient architectures much more compelling, as they deliver very similar performance at a fraction of the cost. The MobileNetV2 architecture outperforms the SqueezeNet architecture in Top-k accuracy, in contrast to the Top-1 accuracy seen in Section 4.1.2.

An important consideration when evaluating these models is the distance in Exposure-space from the optimal prediction that the model produces. A difference

of $0.25 - 0.75$ steps in exposure value produces images that are perceptually quite similar to the optimal prediction target. As such, the percentage of predictions produced by a model that are within N EV steps of the optimal target are reported in Figure 4.1.

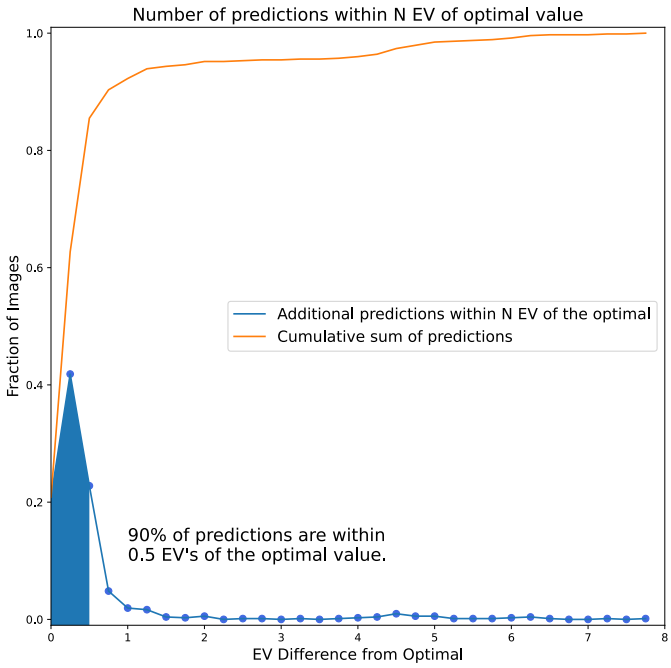

Figure 4.1: This plot shows the percentage of errors that fall within N EV steps of the optimal EV choice. Notice that a 90% of the predictions are within 0.5 of the optimal EV choice.

4.1.6 Image results

The images predicted by the various models are often of high quality, as shown quantitatively in Table 4.1. A selection of predicted image results from the top-performing MobileNet model are shown in Figure 4.2. They are presented next to the ground-truth images for comparison purposes, as well as the EV image predicted

that will be combined with the baseline EV 0 image during the exposure fusion process. The predicted images appear qualitatively very close to the ground-truth images, which are composed of three more images yet appear largely similar in quality.

4.2 Baseline comparisons

The baselines are shown in Table 4.3. This table shows the results of the different baseline EV capture strategies in comparison to the best performing trained model. The strategies detailed below were chosen as reasonable comparisons and minimum standards of performance that the model should meet. As can be seen in Table 4.3, the best performing model significantly outperforms the baseline comparisons.

The conventional strategy works by selecting an exposure value ε, and then computing the exposure fusion image with inputs $F(\mathbf{I}_{-\varepsilon_{\text{conv}}}, \mathbf{I}_0, \mathbf{I}_{\varepsilon_{\text{conv}}})$. The exposure value selected for this baseline method is that which has the lowest reconstruction error across the test set:

$$\begin{aligned} \varepsilon_{\text{conv}} &= \text{argmin}_{\varepsilon \in \text{EV}} \left(\sum_{i}^{\text{HDR+}} \| \mathbf{X}_i - F\left(\mathbf{I}_{-\varepsilon}, \mathbf{I}_0, \mathbf{I}_{\varepsilon}\right) \|^2 \right) \\ \varepsilon_{\text{conv}} &= 3.0 \end{aligned} \tag{4.1}$$

This simulates the traditional exposure fusion process on mobile devices, which typically capture one under-exposed image, a regularly exposed image, and an over-exposed image as input to Exposure Fusion.

The fixed choice strategy simply chooses a constant EV that performs best across the dataset, in the same manner as in Equation 4.1. The image computed by the fixed choice strategy is $F(\mathbf{I}_0, \mathbf{I}_{\varepsilon_{\text{fixed}}})$. For the fixed choice strategy, the exposure value is chosen by:

$$\varepsilon_{\text{fixed}} = \text{argmin}_{\varepsilon \in \text{EV}} \left(\sum_{i}^{\text{HDR+}} \| \mathbf{X}_i - F\left(\mathbf{I}_0, \mathbf{I}_\varepsilon\right) \|^2 \right)$$
$$\varepsilon_{\text{fixed}} = 3.75 \tag{4.2}$$

An upper bound on performance is possible due to the pregeneration of the data and associated statistics. The oracle method computes the optimal EV for each image, $\varepsilon^*(\mathbf{X})$, and looks up this value during inference time. The optimal EV calculation is described in Equation (3.2).

The scores presented below are averaged across the dataset. For the fixed choice and conventional strategies, the exposure values used were selected by examining the dataset for the exposure values that produced the lowest average reconstruction error when the given strategy was used.

Metric	**Model Prediction**	**Fixed EV** $(0 + \varepsilon_{\text{fixed}})$	**Conventional** $(0 \pm \varepsilon_{\text{conv}})$	**Histogram**	**Oracle** $(0 + \varepsilon^*)$
RMSE ↓	11.90 ± 5.47	14.35 ± 10.20	14.48 ± 5.05	18.57 ± 9.33	10.72 ± 5.93
PSNR ↑	27.42 ± 3.66	27.27 ± 6.87	23.14 ± 3.15	23.78 ± 4.23	29.08 ± 5.77
SSIM ↑	0.97 ± 0.02	0.96 ± 0.02	0.94 ± 0.03	0.94 ± 0.04	0.97 ± 0.02
LPIPS ↓	0.04 ± 0.02	0.05 ± 0.04	0.07 ± 0.03	0.06 ± 0.03	0.05 ± 0.02

Table 4.3: The MobileNetV2 model compared against the various baseline comparisons, as well as the oracle method, which acts as an upper limit on the performance of the system. The oracle method selects the best exposure value for a given image through a look-up table. The Fixed EV method uses an EV of 3.75, as that is the EV that most often results in the lowest RMSE. The conventional method uses an EV of 3.0. The histogram model is a simple linear classifier given the stacked colour histograms of the EV_0 images as input. The results as reported are averaged across the test dataset with the standard deviation reported after the ±, and the metric is indicated as being maximized or minimized by ↑ and ↓ respectively.

4.2.1 Histogram Baseline

A simple neural network was trained on the normalized histograms of the EV_0 images. This neural network is composed of two fully connected layers and outputs a set of probabilities over the possible classes of exposure value for each image in the same manner as the ResNet, SqueezeNet, and MobileNet models.

The purpose of this model was to serve as motivation and justification for the use of models of increased complexity. As can be seen in Table 4.3, the histogram model performs well above simply picking an ε at random from the set of possible exposure values, but does not approach the performance of the more complicated models such as MobileNet, ResNet, and SqueezeNet. This suggests the histograms do not provide enough information for the histogram based model to effectively discriminate among the possible exposure values required in order to generate an exposure fused image close to the ground-truth, gold-standard image.

Further work could investigate if a larger number of bins for the histograms could increase the accuracy enough to rival the more complicated models.

4.2.2 Dataset experimentation

There were many modifications made to the original HDR+ dataset from [12]. In the original dataset, each final image is a result of the capture and processing pipeline that transforms the sequence of burst captured images into a final tone-mapped and exposure fused image. For each image, there exists an intermediate image, which is the result of the alignment and colour range expansion process, but has not undergone final exposure fusion and white-balance steps.

During the intermediate image generation, metadata from the original captured images is corrupted. As such, standard methods for determining the brightness level of the merged and aligned image fail. To compensate, the metadata for all of the

images was examined and compared with the input burst images. From this, it was determined that a correction factor would have to be applied. The exposure value of each image was estimated with the shutter speed and f-number information present in the EXIF data.

The image metadata used for this EV calculation is unaffected by the corruption that occurred in the intermediate DNG images, as these values are present and unchanged from the initial burst captured images. The Exposure Values are calculated according to: $EV = \log_2 \frac{N}{t}$, where N is the f-number and t is the shutter speed of the camera. When the differently exposure-valued images are synthesized from the intermediate DNG file, they first have their exposure value shifted by the calculated EV, in order to have a more predictable and reliable baseline exposure value for the image.

This produces reliably better results during image generation. The images computed with the shifted baselines have generated exposure-fused images closer to the gold-standard images. This heavily implies that the original exposure values present in the intermediate merged images are being produced incorrectly from the software camera pipeline as a result of incorrect metadata, as detailed in Section 3.3.

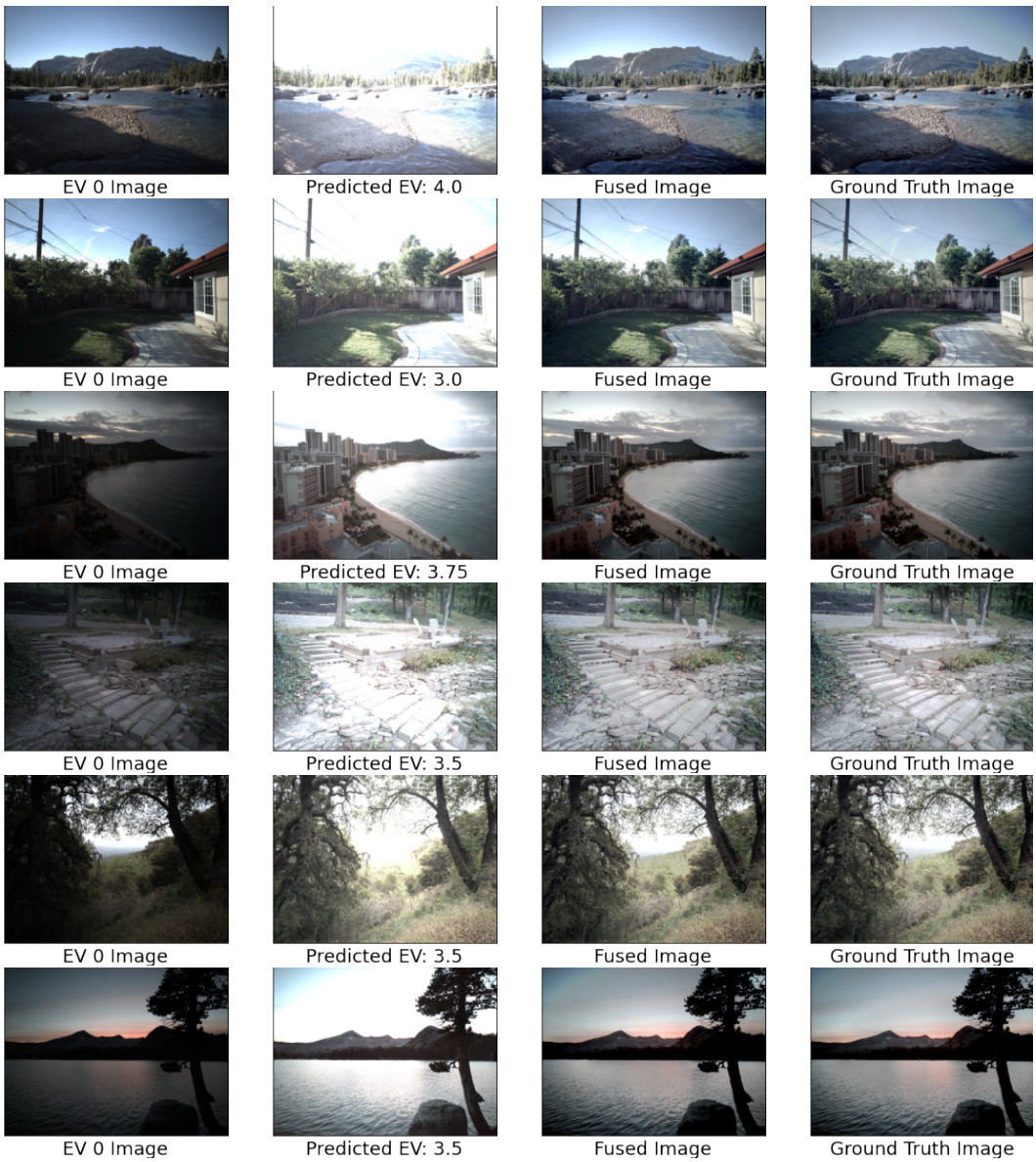

Figure 4.2: Selection of predicted images from the MobileNet-based model. Many images appear qualitatively quite close to the ground-truth.

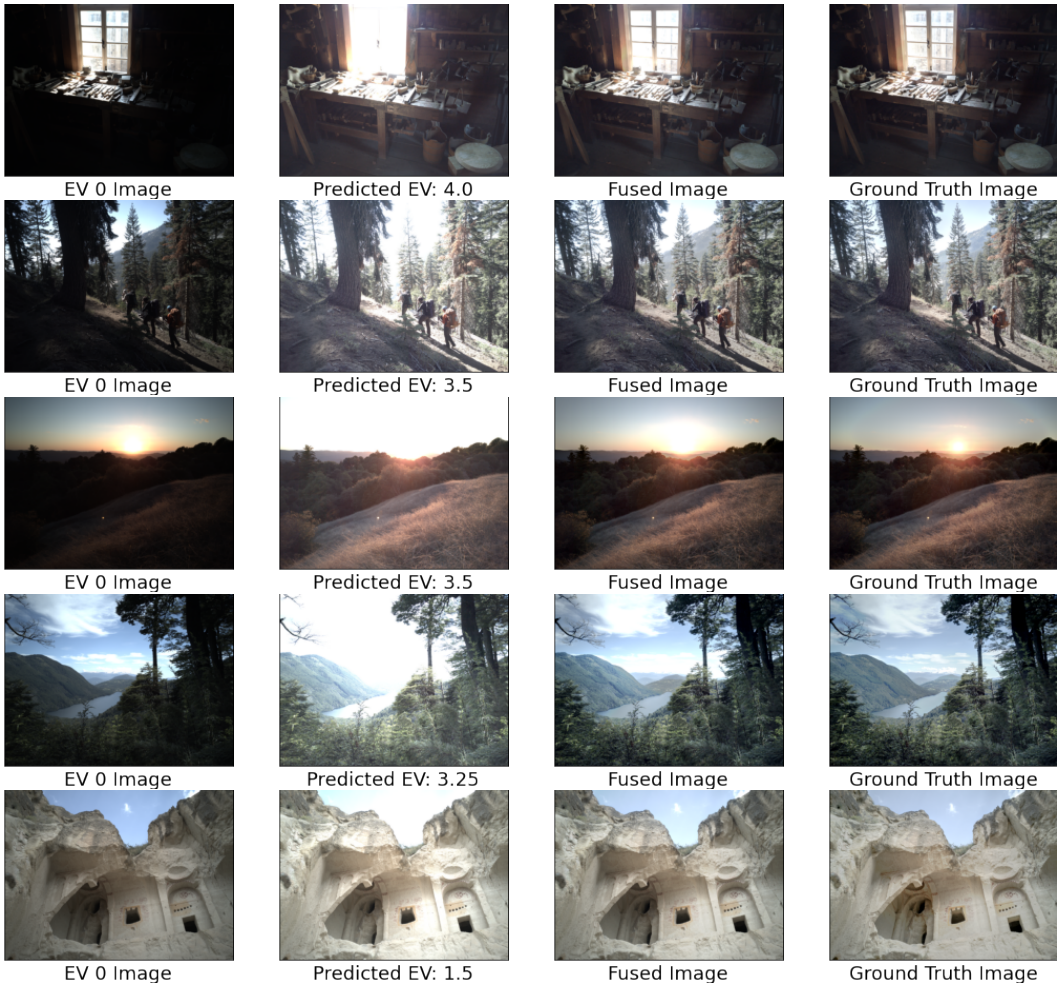

Figure 4.3: Selection of predicted images from the MobileNet-based model. Many images appear qualitatively quite close to the ground-truth.

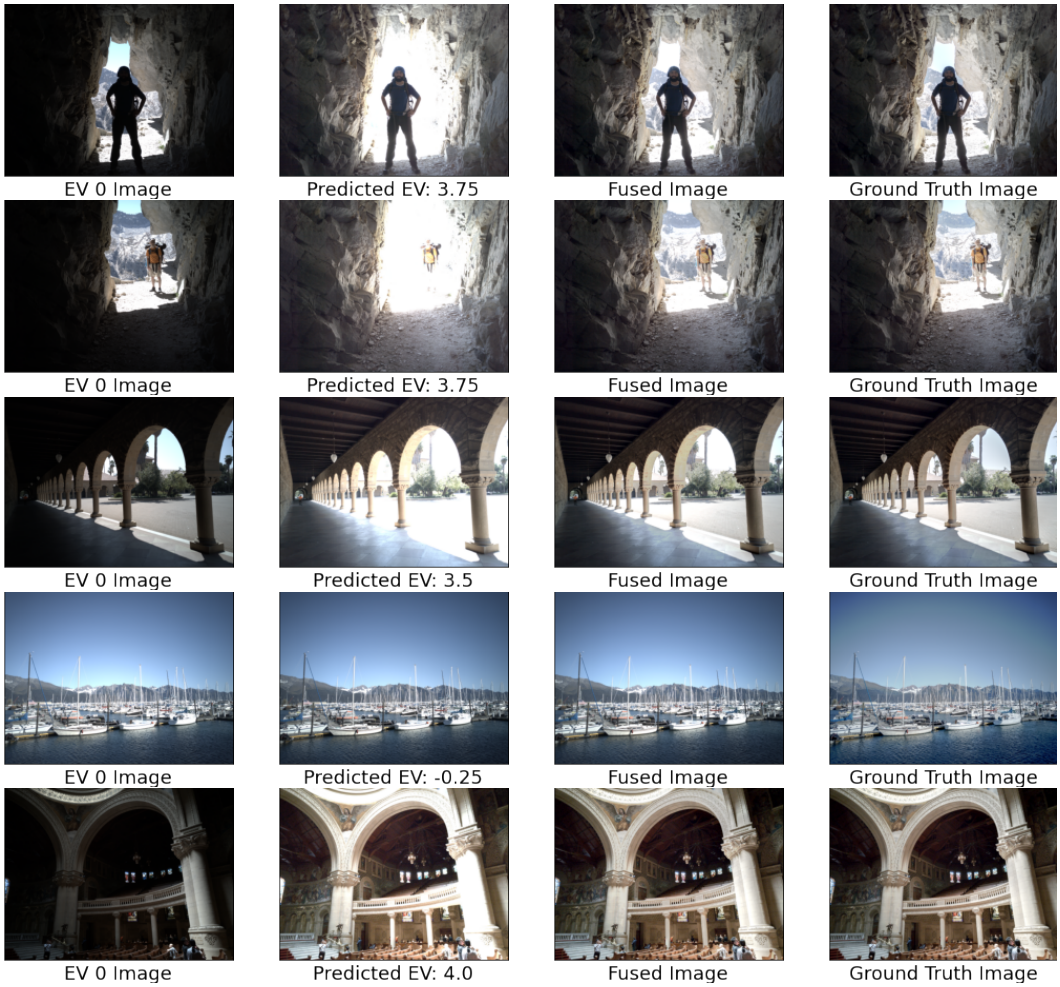

Figure 4.4: Selection of predicted images from the MobileNet-based model. Many images appear qualitatively quite close to the ground-truth.

Chapter 5

Conclusions and Future Directions

This thesis presented a new system for creating HDR images based on exposure fusion using only *two* images. The novelty of this approach was the introduction of a light-weight deep neural network that quickly and reliably predicted the second image in a pair, composed of the baseline (EV 0) image, and one additional image with a different exposure value. We showed that our method provides benefits in terms of computational time, and in user experience, as a requirement for only two image captures introduces smaller chances of error caused by camera motion between image captures.

This work also shows that much of the information in the standard five image captures used in conventional image fusion pipelines is unnecessary. In many fused scenes, major contributions are only made by two images, with much of the rest being small and extraneous to the final result. While the quality of the two image exposure fusion results can not match that of the conventional strategy in every situation, the quality is very close in a wide variety of scenes and lighting conditions.

The main contribution of this thesis is the introduction of a practical system that incorporates those findings into a system that can examine the initially captured image of a scene, and effectively predict the exposure value required in order to produce a high quality, tone-mapped version of the scene through exposure fusion. This system includes a data generation pipeline that can produce many synthetic exposures of a scene from a HDR source image, corrections for malformed metadata in the source images, and a system to train and test a neural network to predict those images. The neural network chosen as the best performing model is the classification network with the MobileNet backbone, as shown in Section 4.1.3 is a fast and efficient network designed for mobile phones and other computationally restricted environments. This provides relatively fast training times, as well as a small total model size.

This work introduced an image generation system that can adjust existing datasets of high dynamic range images into sets of synthetic exposures appropriate for use in training a wide variety of machine learning systems. This system can generate statistics over the whole dataset, and is suitable for extension and modification for future work. Our data generation incorporated information from the image's metadata and was able to correct for incorrectly calibrated exposures of images. Our generated dataset was used to train the different models explored in this thesis, and was a core factor in enabling the fast training and iterative process required to tune and evaluate the system as whole.

5.1 Lessons learned

It is important to carefully analyze the data and metadata associated with a data source by hand. In the course of this thesis, it was discovered that metadata that was used to calculate the baseline exposure values of images was incorrect, and was

negatively affecting the data generation process. This required a significant rework and examination of the generation process in order to identify and understand the error. This involved cross validation and referencing between different stages of the data processing pipeline. The intermediate aligned HDR images in the dataset had to have their metadata cross-referenced with the burst captures that were used to create it, in order to determine which fields in the EXIF data could be relied upon and could be used to infer the missing exposure data that was needed.

Additional lessons learned include the importance of verifying data manually when large software pipelines are employed that are composed of pieces from different environments. Manual examination of the data generation process revealed that high dynamic range images were being down-sampled and corrupted due to data type inconsistencies between software packages, but was silently ignored by the software. Without manual verification of the data, this would not have been discovered and the exposure prediction system would have failed to produce meaningful results.

5.2 Directions for future research

This work showed the possibility of two-image exposure fusion when a learning based approach is used. In most images considered here, there was more than enough information in two images to reproduce exposure-fused images that are very close to the gold-standard images. Future work in this area includes exploring additional optimization of the exposure fusion algorithm itself.

There are many parameters inside the exposure fusion algorithm that have the potential to be optimized with learned parameters, in a scene dependent manner. This can involve the computation of the Laplacian pyramids used for multi-scale blending and merging of the input images. An additional place for optimization

is in the weight-mapping procedure of the algorithm, where a set of hand-picked parameters are used to determine "well-exposedness" factor of the various input images. Such a parameter might benefit from a learned model that can determine whether an image is well-exposed or not in a more sophisticated fashion. It would also be very interesting to see if the entire weight-mapping procedure can be performed by an image segmentation system that would jointly consider each of the images provided as input.

In summary, this work explores reducing the redundancy in the standard exposure fusion process, and does so in an effective manner. It has also left open the possibility for future research in this area to further improve the computational requirements and subjective quality of the results.

Bibliography

[1] B. Hoefflinger, "The Eye and High-Dynamic-Range Vision," in *High-Dynamic-Range (HDR) Vision*, pp. 1–12, 2007.

[2] H. R. Blackwell, "Contrast Thresholds of the Human Eye," *Journal of the Optical Society of America*, vol. 36, pp. 624–643, Nov. 1946.

[3] A. Spivak, A. Belenky, A. Fish, and O. Yadid-Pecht, "Wide-Dynamic-Range CMOS Image Sensors—Comparative Performance Analysis," *IEEE Transactions on Electron Devices*, vol. 56, pp. 2446–2461, Nov. 2009.

[4] S. Aubenas, *Gustave Le Gray, 1820-1884*. Getty Publishing.

[5] V. Schneider, "The High-Dynamic-Range Sensor," in *High-Dynamic-Range (HDR) Vision* (B. Hoefflinger, ed.), pp. 13–56, Springer, 2007.

[6] N. Gelfand, A. Adams, S. H. Park, and K. Pulli, "Multi-exposure imaging on mobile devices," in *ACM International Conference on Multimedia*, pp. 823–826, 2010.

[7] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure Fusion: A Simple and Practical Alternative to High Dynamic Range Photography," *Computer Graphics Forum*, vol. 28, pp. 161–171, Mar. 2009.

[8] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the Expressive Power of Deep Neural Networks," in *International Conference on Machine Learning*, pp. 2847–2854.

[9] Q.-s. Zhang and S.-c. Zhu, "Visual interpretability for deep learning: A survey," vol. 19, no. 1, pp. 27–39.

[10] R. a. Mantiuk, M. Cichowicz, and M. o. a. Smyk, "Implementation of HDR Photographic Pipeline in Mobile Devices," in *Image Analysis and Recognition* (A. Campilho and M. Kamel, eds.), pp. 367–374, Springer.

[11] T. Grosch, "Fast and robust high dynamic range image generation with camera and object movement," *Vision, Modeling and Visualization*, vol. 277284, 2006.

[12] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Computer Graphics*, vol. 35, pp. 1–12, Nov. 2016.

[13] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Transactions on Graphics*, vol. 36, pp. 1–12, July 2017.

[14] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs," in *IEEE International Conference on Computer Vision*, pp. 4724–4732, Oct. 2017.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, pp. 740–755, 2014.

[17] N. K. Kalantari and R. Ramamoorthi, "Deep HDR Video from Sequences with Alternating Exposures," *Computer Graphics Forum*, vol. 38, pp. 193–205, May 2019.

[18] J. J. McCann and A. Rizzi, *The Art and Science of HDR Imaging*, vol. 26. John Wiley & Sons.

[19] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann, May 2010.

[20] T. Steinich and V. Blahnik, "Optical design of camera optics for mobile phones," vol. 1, no. 1-2, pp. 51–58.

[21] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep Tone Mapping Operator for High Dynamic Range Images," *IEEE Trans. on Image Processing*, vol. 29, pp. 1285–1298, 2020.

[22] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Transactions on Graphics*, vol. 36, pp. 178:1–178:15, Nov. 2017.

[23] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, and Y. Zhang, "Deep HDR Imaging via A Non-Local Network," *IEEE Transactions on Image Processing*, vol. 29, pp. 4308–4322, 2020.

[24] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Transactions on Graphics*, vol. 36, pp. 1–10, Nov. 2017.

[25] X. Cerda-Company, C. A. Parraga, and X. Otazu, "Which tone-mapping operator is the best? A comparative study of perceptual quality," *Journal of the Optical Society of America*, vol. 35, p. 626, Apr. 2018.

[26] E. H. Land and J. J. McCann, "Lightness and Retinex Theory," *Journal of the Optical Society of America*, vol. 61, pp. 1–11, Jan. 1971.

[27] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *SIGGRAPH*, pp. 313–318, July 2003.

[28] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Transactions on Graphics*, vol. 21, pp. 249–256, July 2002.

[29] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *SIGGRAPH*, pp. 369–378, Aug. 1997.

[30] J. W. Lee, R.-H. Park, and S. Chang, "Local tone mapping using the K-means algorithm and automatic gamma setting," *IEEE Transactions on Consumer Electronics*, vol. 57, pp. 209–217, Feb. 2011.

[31] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Computers & Graphics*, vol. 32, pp. 330–349, June 2008.

[32] P. Burt and E. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communications*, vol. 31, pp. 532–540, Apr. 1983.

[33] M. Robertson, S. Borman, and R. Stevenson, "Dynamic range improvement through multiple exposures," in *International Conference on Image Processing*, vol. 3, pp. 159–163, 1999.

[34] D. Coffin, "DCRAW: Decoding raw digital photos in linux." `https://www.dechifro.org/dcraw/`.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, June 2018.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016.

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861 [cs]*, Apr. 2017.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," pp. 4510–4520.

[40] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv:1602.07360 [cs]*, Nov. 2016.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, May 2017.

[42] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in

IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, June 2015.

[43] D. Padua, "LAPACK," in *Encyclopedia of Parallel Computing*, pp. 1005–1006, Springer Science & Business Media, 2011.