Master in Computational Colour and Spectral Imaging (COSI)

Multi-Attention SKFHDRNet For HDR Video Reconstruction

Master Thesis Report

Presented by

Ehsan Ullah

and defended at the

Norwegian University of Science and Technology

September 2022

Academic Supervisor(s): Professor Marius Pedersen
Host Supervisor: Kjartan Sebastian Waaseth
Jury Committee:

1. Prof. Philippe Colantoni, Université Jean Monnet, France
2. Dr. Mika Flinkman, University of Eastern Finland, Finland

Submission of the thesis: 10th August 2022
Day of the oral defense: 1st September 2022

Abstract

High dynamic range (HDR) video reconstruction is a very challenging task, especially from sequence of frames with alternating exposures. A convenient approach to generating HDR video is to acquire a sequence of images with alternating exposures using conventional camera systems and reconstruct the missing content or details at each frame. Sadly, conventional methods are typically slow and incapable of dealing with complex examples. Current learning-based techniques usually align low dynamic range (LDR) input sequences using optical flow by estimating flows between neighboring frames. The aligned LDR images are then merged to produce final HDR output. However, due to noise in the under-exposed regions and missing content in the over-exposed regions, precise alignment and fusion is still a big challenge which results in an unappealing ghosting artifacts.

In this research work, we propose a learning-based approach to address the issue of HDR video reconstruction with alternating exposures. Our approach have three main stages, the first stage perform alignment of neighbouring frames to the current frame by estimating the flows between them, the second stage is composed of multi-attention modules and pyramid cascading deformable (PCD) alignment module to refine previously aligned features and extract only relevant information from the neighbouring frames in relation to the reference frame. The final stage perform the merging which takes the features extracted with the multi-attention guided and PCD alignment module as input and estimates the final HDR scene relying on a series of dilated selective kernel fusion residual dense blocks (DSKFRDBs) with the global residual learning strategy allowing the network to fill the over-exposed regions with rich details.

The whole network is trained in an end-to-end fashion for estimating HDR video using publicly available HDR video datasets, with simulated limitations of conventional digital cameras. We employ L1 and a combined L_1MS$-$SSIM loss function to minimize the error between the estimated and original HDR images. We demonstrate the performance of our method on a number of HDR test datasets achieving better alignment and hallucinating details in the over-exposed regions in most cases from the recent state of the art methods and having a smaller number of network parameters than the state of the art methods.

Acknowledgment

- I express my gratitude to Almighty Allah for all the opportunities, hardships, and strength that have enabled me to complete the thesis. I went through a lot during my thesis, both academically and in terms of my personality.
- I would like to express my profound gratitude to my supervisors, Prof. Marius Pedersen, and Mr. Kjartan Sebastian Waaseth for their valuable advice, understanding, and patience, but most importantly for their encouragement and kindness in assisting me in completing my thesis. I am very lucky and honoured by being under their kind supervision.
- I would like to show my deepest gratitude to all my family members who supported me throughout my master's till the end of my thesis.
- I want to express my sincere gratitude to all of my dear COSI classmates who stuck with me and supported me no matter what. In particular, I want to thank Akib Jayed Islam, Tawsin Uddin Ahmed, and Sanam Nisar for their unwavering support throughout tough moments.
- Lastly, Many thanks to EU and all the COSI coordinators for giving me an opportunity to grow and flourish.

Acronyms

CCD: Charged Coupled Device
CIE: Commission internationale de l'éclairage
CMOS: Complementary Metal-Oxide Semiconductor
CRT: Cathode Ray Tube
CRF: Camera Response Function
CD: Pyramid cascading deformable alignment
DLP: Digital Light Processing
DMD Digital Micromirror Device
DLP: Digital Light Processing
DSKFRDBs: Dilated Selective Kernel Fusion Residual Dense Blocks
DSLR: Digital Single-Lens Reflex Camera
fc: fully connected
HDR: High Dynamic Range
HDRI: High Dynamic Range Imaging
HVS: Human Visual System
IQA: Image Quality Assesment
JNDs: Just Noticeable Differences
LED: Light Emitting Diode
LDR: Low Dynamic Range
MDTA: Multi-Dconv Head 'Transposed'
MLP: Multi-Layer Perceptron
OLED: Organic Light-Emitting Diode
PSNR: Peak Signal-to-Noise Ratio
SDR: Standard Dynamic Range
SKF: Selective Kernel Fusion
SA: Self-Attention
TMO's: Tone-Mapping Operators
iTMO's: inverse Tone-Mapping Operators
VDP: Visual Difference Predictor
VQM: Video Quality Metric

Contents

1 Introduction

Camera devices are designed to mimic or perform similar tasks like human visual system (HVS) by capturing surrounding scene information for higher level processing. Viewing a physical scene on a display device which has been captured by camera should give resemblance or evoke a same response as seeing the scene physically using our HVS. For a number of reasons, this is, however, extremely rarely the case and most of the time, the acquired image has inconsistent color and brightness values. Additionally, another well known difference is the mismatch of dynamic range information in many scenes. There is more visual information available in the scene than what can be captured and reproduced since the camera and display are unable to simultaneously cover the wide range of luminance values that the HVS can detect. For instance, one must choose between correctly exposing the background or foreground while trying to capture an image in a dark indoor environment in front of a bright window, as the other information is lost in the dark or saturated image regions, respectively. The human eye register both the foreground and background at the same time with less difficulty.

Clearly camera has limitations in terms of comparing to the HVS. This limitation can be resolved by HDR imaging methods where information in both dark and bright region of an image is recovered with the visual results matching or surpassing the HVS dynamic range. This thesis work will present technical research contribution in the HDR imaging pipeline (specifically in inverse tone mapping section) see Fig. 2.3. First, a short introduction to HDR imaging is provided in this chapter. Then the thesis research context, Motivation, Research gaps, Thesis outline and then contributions are briefly explained.

1.1 Display referred image representation

The big amount of currently used digital images are captured and stored using 8-bit integer values, offering $2^8 = 256$ distinct levels for differentiating the intensity of each color channel in a pixel. Low dynamic range formats like JPEG, PNG, TIFF, and others are constructed according to the limitations of the display devices and

accommodate according to the capabilities of the imaging device with a minimum care for the loss of visual information which the imaging device cannot display. As a result, these formats are usually regarded as device-referred, which is also known as output-referred. Clearly, the relation of device referred image representations is limited to real photometric properties of the scene. Due to that, it is very difficult to reproduce the scene with full details and a high level of realism across display devices that have significantly varied contrast specifications, absolute lowest and peak luminance values, and color gamuts.

1.2 Scene referred image representation

A simple solution to this problem may be found in the scene-referred encoding of pictures, which encodes the actual photometric qualities of the scenes that are being represented. It then depends on a specific device to convert data from a common representation, by directly correlating physical luminance or spectral radiance values , into a format that can be easily handle by that device. This provides a best possible solution of utilizing HDR content since only the device knows all of the information relating to its constraints and limitations for showing the content in a more appropriate way. HDR file formats are example of scene-referred encoding since they often represent either luminance or spectral radiance. Rather than being gamma corrected and ready to display encoded pixel values Mantiuk (2015). The difficulty of accurately representing scene-referred images comes when considering the amount of quantization error that can be compensated. Regarding display referred image formats, the pixel precision is directly related to the reproduction capabilities of the display devices that are being used as targets. The precision regarding scene-referred image representations should not be related to any specific imaging technology. However, the big issue with scene referred image representation is that it requires huge storage space which restricts the usage in commercial products.

1.3 Dynamic range compression

In order to display the content according to the capabilities of the display devices. The Luma value L should be encoded and then gamma corrected, $l = \mathrm{L}^{1/\gamma}$, by performing non-linear correction of the linear luminance L. The dynamic range is compressed by the gamma value, which is typically in the range between 1.8 - 2.8 Mantiuk (2015). The non-linearity of cathode ray tube (CRT) displays was the reason for the original purpose of this correction, but it is also applied to current displays by simulating the non-linearity. In order to make the range of encoded

values more perceptually linear, the correction additionally accounts for a similar HVS non-linearities by considering the range of LDR image intensities. So when encoding an image with the limited precision provided by 8 bits, the quantization errors caused by rounding off to the nearest representable value would be viewed as equally big across the range of pixel values. Utilizing the gamma-offset-gain model, the gamma correction method may be further expanded by taking the display and viewing environment conditions into consideration Mantiuk (2015). Before the process of encoding, digital cameras often calibrate the captured images in-camera. The non-linear calibration, also known as the camera response function (CRF), can take multiple forms and achieve desired calibration/tone-mapping results depending on the brand and model of the camera. For instance, one camera may use a larger dynamic range compression to expose more of the RAW pixels recorded by the sensor, while another output a better reproduction of contrast. Most recent digital single-lens reflex camera (DSLR) cameras have the option to directly access the linear RAW sensor read-out so that it may be processed for display in post-processing, giving customers additional flexibility.

The RAW format is composed of a wider dynamic range than the display-referred 8-bit image with a storage capacity of about 12–14 bits. HDR images and videos, in contrast to the LDR formats, should not be transferred directly to a display or imaging devices. Instead, scene referred calibration is done by measuring linear relative luminance, which relates pixel values to the physical luminance in the image that was acquired. The linearity of pixel values, in addition to the large dynamic range and precision, is the most crucial characteristic of HDR content. Either the linear RAW format data may be used in algorithms for creating HDR images\videos from ordinary cameras, or the non-linear transformation employed by the CRF has to be calculated and inverted. On the other hand, it is more challenging to establish an absolute calibration of the pixels. It depends on the imaging sensor itself as well as a broad range of camera settings, such as exposure time, aperture, gain, etc. Using a device that measures a reference luminance white point within the captured scene and then scaling the relative luminances of the HDR image to correspond with the measurement for providing absolute calibration.

Having distinct display and scene calibration domains, preparing an HDR image for display, commonly known as tone-mapping, entails not only compressing the dynamic range but also converting from a scene-referred to a display-referred format. Fig. 1.1 illustrates the impact of gamma correction which perform compression when transforming to a display-referred format. The gamma compression reduces the dynamic range enabling more shadow and highlights to be reproduce. Tone-mapping operators (TMO's) amplify this process, making the overall image content more visible, by producing a perception of the scene that is more closely aligned with that of HVS.

(a) *Linear.* **(b)** *Gamma-adjusted.* **(c)** *Tone mapped.*

Figure 1.1: *Represents and compares the differences between scene-referred linear values (a), gamma-adjusted display-referred pixels with $\gamma = 2.2$ (b), a tone mapped image using Reinhard et al. (2002) technique (c). The tone mapping technique significantly compressed the dynamic range while retaining the local contrast in the image.*

From all the above discussion, It is clear that there is a huge amount of legacy content with the data represented in 8 bits considering the capabilities of conventional display system. Current HDR display system cover far wider luminance range with a wide color gamuts. Directly showing the legacy content stored in 8 bits will produce contouring and quantization artifacts. For faithfully representing low dynamic range content in HDR displays, Inverse tone mapping is required which recover details in the over exposed and under-exposed regions of the images by increasing the dynamic range of the content either by combining a set of images with multiple exposures or applying learning-based techniques to recover lost details in bright and dark regions due to compression.

1.4 Research context

It is obvious that the future needs for the creation/reconstruction, distribution, and display of HDR images and videos will increase as its application grows. This thesis is a contribution to the HDR video reconstruction from LDR images of the HDR imaging pipeline. Utilizing and enabling compressed dynamic range LDR images in these applications is an interesting problem due to immense popularity of HDR applications and the limited supply of HDR image and video content with addition

of huge amount of legacy content which is captured by conventional cameras need proper prior mapping before it is faithfully represented in HDR displays. Researcher come up with different approaches for this goal, known as inverse tone-mapping operators (iTMOs) and learning based image reconstruction techniques. These, however, are rather limited since they only increase dynamic range without really recovering the lost information in some of the challenging LDR images. Furthermore, compared to HDR still images, HDR video reconstruction is less researched, which is the major focus of this thesis work. We approach to the problem of HDR video reconstruction for recovering details in saturated and underexposed areas of an LDR image sequences using current deep learning techniques. It is challenging to employ LDR images in a wide range of HDR applications, but currently, the deep learning technique represents a major improvement over past methods. Which will pave the way for more utilization of LDR images in future HDR applications.

1.5 Motivation

From the introduction, big portion of currently used digital images and videos are stored using 8-bit integer values where each pixel color is represented by 256 values. Representing original scene with limited number of integer values is primarily because of storage constraints and display capabilities. By doing this, visual information is lost in under and over-exposed regions considering both the cases of images and videos. The optimal solution to HDR video reconstruction is to use the inexpensive conventional cameras capturing LDR sequences by alternating the exposure of each frame using software solutions. Fig.1.2 shows the sequences of frames where the exposure is alternated between frames in such a way that large range of luminance values is registered.

Figure 1.2: *represents conventional off the shelf camera frames capture with alternating exposures for reconstruction of single HDR frame by making use of the dynamic range from both exposures for registering large dynamic range.*

The HDR video can then be reconstructed by recovering the missing HDR details at each frame, from the neighboring images with different exposure. Additionally,

it is also important to temporally align the frames for reconstructing a high-quality HDR video. Few methods have been proposed to address the HDR video reconstruction problem using alternating exposure which is discussed in great detail in the chapter 2. The prior research work have limitations. First they are slow in reconstruction of HDR video. Second they are struggling in recovering details in challenging examples. Third, due to the temporal aspects of the video flickering and frame alignment is still a challenging issue in reconstructing high quality video. In addition to that, Current learning-based video reconstruction methods have large network parameters making it non feasible for commercial use cases.

1.6 Research gap

This thesis will answer some of the limitations of the previous work mentioned in Sec. 1.5.

1. **[R1]**: Reconstruction of HDR video requires proper alignment of frames to remove the ghosting artifacts which is depicted in Fig 1.3. Previous methods produce ghosting artifacts in challenging test samples.This research work will provide a unique solution to this research problem discussed in Chapter 3.

Figure 1.3: *represents conventional camera consecutive frames with alternating exposures highlighted in orange. Ghosted frames is highlighted in red. Indicates the importance of image alignment in HDR video reconstruction.*

2. **[R2]**: Recovering details in under and over-exposed regions is still a big challenge in HDR video reconstruction see (Fig. 1.4). Prior methodologies struggle to recover rich information from over-exposed and under-exposed regions of LDR images. This thesis work will introduce a unique strategy in recovering details in under and over-exposed regions using a learning-based method.

Figure 1.4: *Represents conventional camera consecutive frames with alternating exposures highlighted in red. The over-exposed frames have regions where all the information is lost while in case low exposure frames, the information is suppressed in shadow regions of the frames. The goal of the the HDR reconstruction method is to recover the lost details in both under and over-exposed regions from LDR frames.*

3. **[R3]**: Previous learning-based methods have large network parameters. This increases the model inference time and become harder to deploy in interactive real-time commercial purposes. Robust and efficient learning-based method having less network parameters while achieving state of the art results in HDR video reconstruction is presented to resolve this issue.

1.7 Thesis outline

The dissertation organization and overall structure are as follows. In Chapter 2, the overall processing pipeline from acquisition to display for HDR images and videos is presented with the inverse tone mapping part discussed in more detail where physical, non learning and learning based image\video reconstruction methodologies are discussed.

The proposed approach methodology is presented in detail in Chapter 3, which is divided and discussed in multiple sections, each perform specific task on resolving issues related HDR video reconstruction. The general idea of the research work is presented, outlining the problem and what is the approach of the thesis to overcome it. Overall, the architecture of our proposed methodology for HDR video reconstruction is given along with the training process. Datasets for training and testing and its post-processing is presented. L_1 and a combined L_1MS$-$SSIM loss are explained. The prior work is briefly introduced in this research work with its limitations by comparing their performance on multiple test datasets with our proposed method variants in Chapter 4. The results are presented along with a discussion about different experiments done in our thesis work both quantitatively

and qualitatively. The result section in Chapter 4 illustrates how Multi-Attention SKFHDRNet variants performed against prior HDR reconstruction techniques. Initial ablation study is presented in Chapter 5 with the aim of continuing it in a future work.The limitations of our proposed method along with the prior research work is discussed in Chapter 5. Further conclusion and future work is presented in Chapter 6.

1.8 Contributions

The major contributions of this thesis for HDR video reconstruction is as follows:

- Introduction of Multi-Attention blocks with the goal of proper image alignment by extracting rich information both spatial and channel-wise. This contribution will address research gap **[R1]** in Section 1.6.
- We also applied soft attention by adapting to the scale of the information in the input frames for improved alignment. This contribution is also in relation to research gap **[R1]** in Section 1.6.
- We introduce Pyramid cascading deformable alignment module for further refining the alignment process. This contribute to the research gap **[R1]** in Section 1.6.
- For effective final HDR video reconstruction we employ robust dilated selective kernel fusion residual dense blocks (DSKFRDBs) in the merge network for recovering details in over-exposed regions. This contribution is aligned to the research gap **[R2]** in Section 1.6.
- Our proposed model have less network parameters from prior learning-based techniques. This contribution is aligned to research gap **[R3]** in section 1.6
- Lastly, The proposed method training is done by using single L_1 loss and a combined $\mathrm{L_1MS{-}SSIM}$ loss for guiding the optimization algorithm by learning more refined network weight parameters for HDR video reconstruction.

Our Multi-Attention SKFHDRNet proposed method showed a substantial improvement over existing techniques and makes it possible to use LDR frames in a reconstruction of HDR video.

2 Background

This article will discuss different sections of the processing pipeline from acquisition to display for HDR images and videos, and will discuss the inverse tone mapping part in more detail with proposed solution.

2.1 High dynamic range

The HVS's greater dynamic range than that of off the shelf cameras and display devices provides a natural motivation for the development of methods for capturing and displaying HDR images, which can better mimic the sensation of viewing a real scene. The most popular method for creating HDR images is to combine a set of images that were taken at various exposure times since a camera sensor has limited range of luminance that can be captured, as shown in Figure 2.1d.

Details and information in the dark or bright areas in the images or videos are captured by adjusting the exposure time. For instance, details in the dark regions of the image is retained or captured by setting a long exposure time while by doing this information in the bright image regions is lost due to sensor over saturation. Similarly, information in the bright regions of image are retained or captured by setting the exposure time shorter. By doing this it will disappear the information in darker regions of image with addition of noise and quantization error. Information both in dark and bright regions of image which are outside of the capabilities of the sensor can be recovered by combining images with different exposures. Combining images with multiple exposure retain large amount of missing details with an overall increase of dynamic range. On the other hand, Images and videos captured with a high dynamic range (HDR) imaging method, include pixels that represents a far wider variety of colors and brightness levels than those provided by images with an industry standard dynamic range. HDR is an important technology with the capability of significantly increasing the overall quality of the visual information, provides more realism to the audience.

2.1.1 Definition

Illuminance, or incoming light from the environment, is reflected according to the characteristics of the surface material at a particular place on a surface in a scene. When the light is registered as it falls on a pixel in a camera sensor, The integrated outgoing light across a region in a particular direction is measured. SI unit, Candela per square meter (cd/m^2) is specifically used for measuring luminance in a scene or on a screen. The display manufacturing sector frequently referred to this unit as nit (1 nit = 1 cd/m^2). In Fig. 2.1 (a), the typical luminances for some objects are illustrated as a reference for the range of observable luminance values. The dynamic range is the ratio or measure of the difference between the smallest and largest value recorded by an image sensor or shown on a display Mantiuk (2015). In case of human visual system (HVS), dynamic range is the scene visible brightness which is between its smallest and highest value. For a camera sensor, it lies between the highest observable brightness before the sensor saturates and the lowest detectable luminance above the noise floor Mantiuk (2015). While in displays, Dynamic range lies between the smallest and largest pixel luminance values that can be rendered concurrently on the screen Mantiuk (2015).

The dynamic range is 1,000,000:1, or $6\log_{10}$ units, for instance, if the lowest and highest values are 0.001 and 1,000 cd/m^2, respectively. The dynamic range is frequently expressed in stops/f-stops, which are $\log_2$ units, in case of camera systems. Alternatively, the signal-to-noise ratio (SNR), which is often expressed in decibels and has the formula SNR = 20 $\log_{10}$ (I saturatedpoint/I noisefloor)dB, can be used to specify the dynamic range of a camera sensor respectively. The definition of high dynamic range is not entirely clear from the literature on HDR imaging, and it may change depending on the application. The terminology is often used to describe something that has a higher dynamic range than typical cameras and screens. The dynamic range of an HDR image may actually be very limited in some circumstances, therefore this can be misleading in some cases. The term (LDR) is usually used to describe describe images that are not HDR.

2.1.2 HVS dynamic range

To compare the capabilities of the HVS to other capturing and display technologies, Figure 2.1 (a) displays typical dynamic ranges. The HVS can observe a very wide range of luminances, with a total dynamic range of around 14 $\log_{10}$ units, ranging from about 10^6 cd/m^2 up to 10^8 cd/m^2 Ferwerda et al. (1996). However, the eye must adapt to the various lighting conditions in order to observe this wide dynamic range. The photoreceptors' bleaching and regeneration processes is a vital component in this process, as well as pupils of the eye Eilertsen (2018). The

processes can take quite some time, particularly when regenerating photopigment while adjusting to a dark environment. For instance, when an individual arrives a dark room after being in a bright outdoor environment; it takes several minutes before details can be perceived, and it can take up to half an hour for complete adaptation to the dark environment Mantiuk (2015). The retina have two different kinds of photoreceptors that are responsive to various luminance ranges. In contrary to the cones, which becomes activated in brighter environments and provide color vision and higher resolution, the rods are more sensitive but offer worse acuity and no color vision. Figure 2.1 (b) shows the working capabilities of the different photoreceptors. Scotopic vision refers to the portion of the HVS vision where only rods are active, while photopic vision refers to the portion of the HVS vision where only cones are active. The mesopic vision, is a region where both rods and cones contribute to the vision, due to the overlap of the working ranges of these cones. Given the complexity of how the HVS functions, it is challenging to determine the simultaneous dynamic range of the eye, which is also depicted in Figure 2.1 (b).

2.1.3 Camera and display dynamic range

Compact digital cameras capture a dynamic range of $2\log_{10}$ units, high-end digital single-lens reflex (DSLR) cameras have a capability of capturing a dynamic range of over $4\log_{10}$ units, While a professional HDR-capable cinematographic video cameras can have a dynamic range of up to $5\log_{10}$ units. The dynamic range of a typical consumer-level camera sensor is shown in Figure 2.1 (c). Due to Sensor saturation, maximum luminance for the current exposure time cannot be recorded. Due to noise and quantization, the information below the lowest detectable value is lost. The ability to handle noise is mainly attributable to differences in a dynamic range of sensors. For example, a big sensor with low resolution can lower the noise level by integrating over the larger pixel regions. There are several techniques to assess a sensor's noise floor, and the results provided by the manufacturers are frequently overly optimistic. This indicates that it may be challenging to attain in reality the dynamic ranges mentioned above, with up to $5\log_{10}$ units. The process of HDR reconstruction may be used to merge a number of different exposures into a final image in order to produce an HDR image or video. This method of extending the dynamic range is shown in Figure 2.1 (d). Figure 2.1 (e) provides an example of a learning based reconstruction method for increasing dynamic range. which typically involve HDR image reconstruction using deep learning techniques to recover lost information in the dark and bright image regions. Our study is based on learning-based technique it is briefly discussed in Chapter 3. Finally, Figure 2.1(f) shows the typical dynamic range of multiple display devices.

Figure 2.2c shows that the dynamic range of a consumer-grade camera sensor

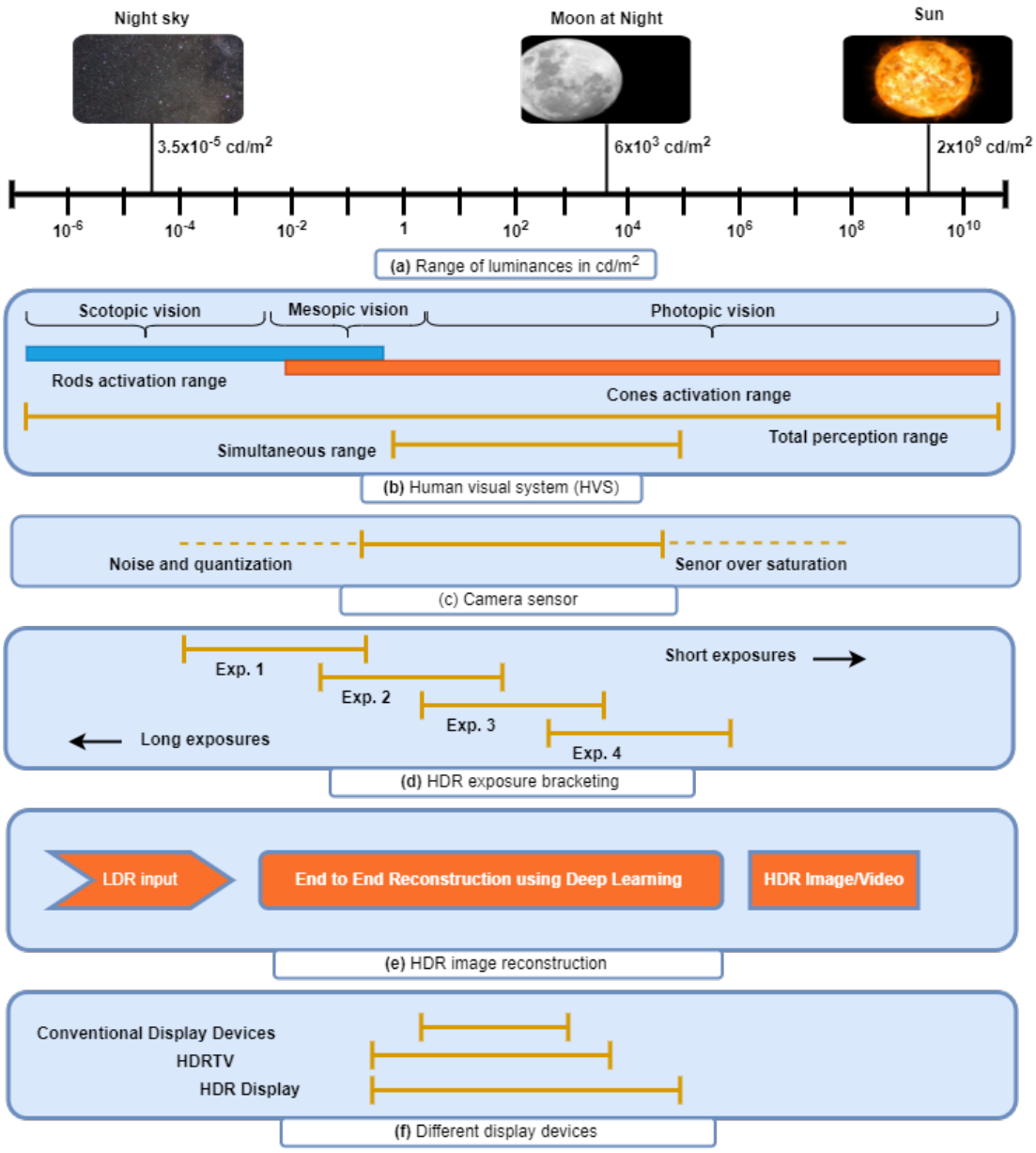
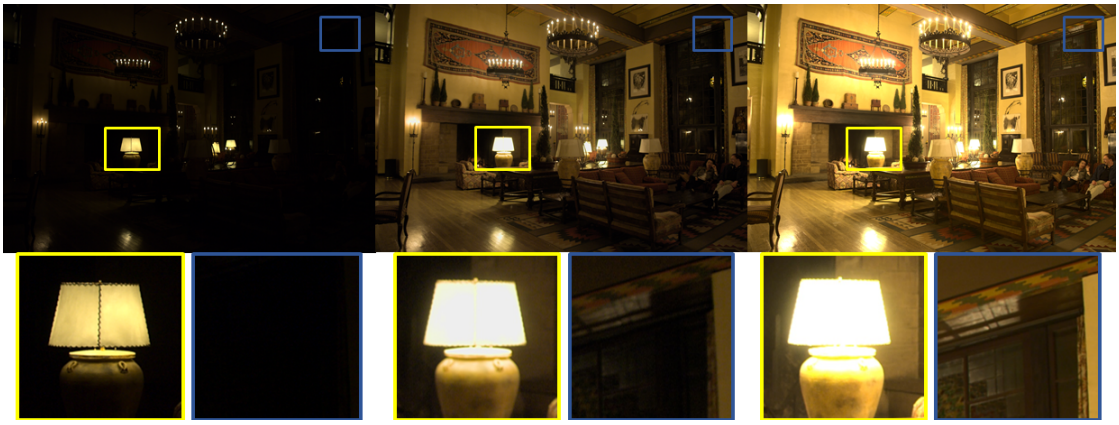


Figure 2.1: *(**a**): Illustrates luminances for a few objects provided as a reference for the range of observable values. (**b**): shows the working ranges of the different HVS photoreceptors. (**c**): shows conventional camera sensor capabilities. (**d**): represents HDR reconstruction by merging multiple frames with different exposures. (**e**): represents learning based strategy. (**f**): represents the comparison of different Display capabilities in luminance range.*

which capabilities is almost equivalent to that of a traditional liquid-crystal display (LCD), which has a range of 2.3–2.7$\log_{10}$ units. Image content are lost in shadows or highlights when the image's dynamic range is significantly higher than the display device. The dynamic range of the image can be compressed while maintaining the majority of the features by employing tone-mapping techniques. Figure 2.2 provides an illustration of the differences between representing an HDR image directly or displaying it after applying a TMO. Tone-mapping may be used for more than just adapting an HDR image to a standard display. It may also be used to take into account more subtle variations in the dynamic range and color characteristics of cameras and displays.



(a) *(a) Exp.: 1/60s, F-stop: f/2.8* **(b)** *(b) Exp.: 1/8s, F-stop: f/2.8*

(c) *(c) Exp.: 1s, F-stop: f/2.8*

Figure 2.2: *An HDR image represents a full range of luminance in the scene. The top row represents three exposure bracketed images. An HDR image can be produced by using a technique similar to Reinhard et al. (2002). The enlarged bright and dark image regions are displayed in the bottom row. The numbers specify both the absolute exposure times and the relative exposures in relation to (b). The example illustrates the need for a relatively wide exposure difference to capture both highlights (a) and details of dark image regions (c), and it also shows that saturated pixels may still be found in the brightest highlights of the darkest image.*

2.2 HDR imaging Pipeline

An HDR imaging device, either with a sensor that can cover a large dynamic range or with a multi-exposure system, can also be used to directly capture an HDR

image (Section 2.3). After acquisition, it is then necessary to effectively compress and encode HDR images or videos using a variety of HDR capable formats either for transmission or storage (Section 2.5). For faithful representation of HDR images or videos on display devices requires tone mapping (Section 2.6) transforming the HDR content to a display-referred format by compressing the dynamic range to the capabilities of the display device while retaining all the important visual image information. To show low dynamic range content or tone-mapped images or videos on HDR capable devices, LDR content must be upscaled using inverse Tone-mapping methods (Section 2.4). Professional HDR display technologies along with commercial displays are discussed in (Section 2.7) Finally, Image quality metrics capable of assessing the quality of HDR images and videos are discussed. (Section 2.8) Overall the acquisition, distribution, tone-mapping, reconstruction, display and the quality assessment components of a general HDR image pipeline are covered in this chapter (Figure 2.3). Background information on research and advancement in HDR imaging is covered in this chapter.

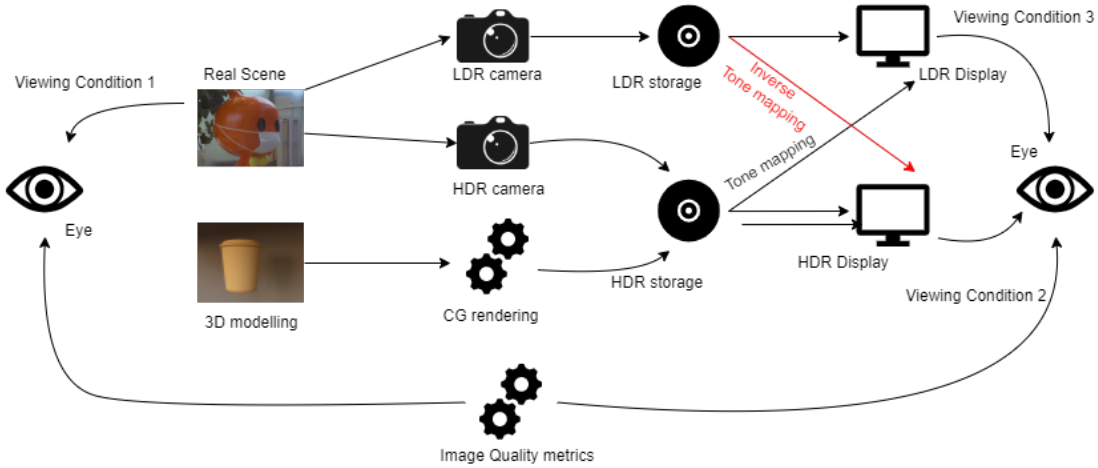


Figure 2.3: *Represents different components of the imaging pipeline along with HDR methodologies.*

2.3 HDR image and video acquisition

HDR content can either be generated through abstract scene modeling using computer graphics tools or acquisition of real world scene using HDR imaging techniques. In case of computer graphics approach, realistic and visual appealing HDR content are generated by means of image synthesis and global illumination. Abstract scene modeling with computer graphics tools and real-world scenes acquisition through

camera devices are the two major sources of HDR content. In the first case, photometrically calibrated pixel values are used to produce compelling results through realistic image synthesis and global illumination computation. Traditional cameras with (LDR) sensors can be used to capture multiple exposures of a static scene, which are then combined using computational methods to create an HDR image. Camera misalignment in the case of hand-held photography, as well as ghosting due to dynamic aspects in a scene can be eliminated or reduced using specific techniques. When capturing multi-exposure HDR video, The requirement for image alignment along with accounting for temporal coherency between frames of a video become necessary. Specialized HDR sensors and cameras, which can capture a scene in a single shot, can avoid ghosting and alignment issues. Following a computational approach, the legacy LDR images and videos captured using conventional sensor can be transformed into HDR content by increasing the dynamic range and recovering the lost details in dark and over-saturated region and performing proper image alignment and color correction. As the inverse Tone-mapping is an is an ill-posed problem where the goal is to recover as much detail as possible in final HDR image

2.3.1 Computer graphics based HDR acquisition

In the mid-eighties researchers began to integrate realistic image synthesis with physically based lighting simulation in computer graphics Kajiya (1986); Pharr et al. (2016). Physical-based illumination modelling requires data to be in radiometric or photometric scale. Such type of data can taken from the manufacturer of lighting equipment because they usually give access to emissive properties of their luminaries Mantiuk (2015). In case of BRDF which shows the material reflectance characteristics can be estimated by utilizing a data for similar materials or through analytical reflectance models with appropriate parameter configurations. With the use of physically-based data, a decent approximation of illumination distribution in real-world scenes may be achieved by physically-based lighting modeling Mantiuk (2015). Apart from that, the rendered images have pixel values in radiance or luminance which is defining property of HDR imaging. Real-time HDR image rendering has been made possible because of recent improvements in graphics processing units (GPUs) mostly used in gaming consoles, which perform operations and rendering in floating point precision. In conclusion, computer graphics is an important source of high dynamic range (HDR) content that features virtually arbitrary contrast ranges and negligible quantization errors. This is something that is significantly difficult to achieve using sensors based methods, mainly due to the limitations of optical systems.

2.3.2 Camera RAW and JPEG images

Due to display hardware limitations, many inexpensive camera devices generate compressed JPEGs images or video. The reason is that the uncompressed video require higher bandwidth due to that cheap webcams transmit video as JPEG images. These conventional camera devices perform Tone-mapping to convert CCD or CMOS sensor's linear responses into gamma-corrected pixel values. This tone-mapping and JPEG compression distort and limit the image dynamic range. More costly cameras, especially DSLRs, provides an option of acquiring images in RAW format, which stores sensor's values providing higher dynamic range than JPEG format. These RAW format images can be then tone-mapped. Additionally, The dynamic range increase is especially significant for bigger sensor sizes, which have more photon capacity and so capture more luminance range.

2.3.3 HDR sensors and cameras

The more accurate results can be achieved from specialized single-shot HDR cameras because deghosting and alignment techniques might not be effective in some circumstances. These devices are quite expensive, Which limits the commercial usage of these imaging devices. One approach is to make use of varying pixel sensitivities on a sensor, with the need of unique sensor design Mantiuk (2015). With this method, sensor sensitivity and usually spatial resolution are traded for higher dynamic range. Conversely, multiple conventional cameras with various exposure settings can be interconnected together by an optical component that split light onto their sensors Froehlich et al. (2014). Lastly, HDR sensors can be specifically build by directly capturing the incoming illumination from a scene as a logarithmic response Seger et al. (1999); Kavadias et al. (2000).

2.3.3.1 Spatial exposure change

Typically, a mask with a per-pixel changeable optical density is used to vary the spatial exposure. It is flexible to choose the number of various optical densities, and they may produce a regular or irregular pattern. Nayar and Mitsunaga (2000) suggest using a mask that is put right in front of the sensor chip with a regular pattern for capturing the scene at four separate exposures. A sensor with the capturing capabilities of scenes in an 8-bit can produce a dynamic range of around 85dB as a result of combining those four exposures.

To modulate the spatial exposure, a mask with adjustable optical density per pixel is typically utilized. There is flexibility in the number of different optical densities that may be used, and they can result in either a regular or irregular

pattern. Nayar and Mitsunaga (2000) suggested that taking four different exposures of the scene by placing a mask with a regular pattern directly in front of the sensor chip and combining all those four exposures, a sensor with the ability to capture scenes in 8-bit may provide a dynamic range of about 85dB. Instead of using a fixed pattern element, Adaptive Dynamic Range Imaging (ADRI) employs spatial exposure variation using an adaptive optical density mask Nayar et al. (2004). The typical mask modifies its optical density per pixel in response to input from the image sensor. Saturated pixels have a higher density of related pixels in the mask than noisy pixels. The feedback, however, results in a delay that might show up as temporal aberrations in moving high contrast edges that are over or under-exposed.

2.3.3.2 Multi-camera methods

A comparatively less expensive alternative to spatial exposure multiplexing is the combination of pictures from many independent cameras. For instance, some techniques make use of stereo camera capturing setups, in which the two cameras are programmed to record images at various exposures. To align the different views of the cameras, the pictures need accurate stereo matching. The camera views can be aligned using an external beam-splitter to solve this issue. Using two Arri Alexa cameras in this configuration, Froehlich et al. (2014) were able to successfully capture a wider range of HDR video with a dynamic range of up to 18 stops.

2.3.3.3 Multiple sensors with beam splitters

Using beam splitters, which directs the light to multiple sensors, one may capture multiple exposures per video frame simultaneously in order to increase dynamic range Tocci et al. (2011) and Kronander et al. (2013). This totally eliminates the motion issue, but it needs extremely controlled and precise optics design for proper alignment of images taken by sensors. The focal length and aperture adjustment are readily streamlined when a single lens system is employed. The effective dynamic range is a function of number of utilized sensors, which is generally 3–4. Any extra sensor decreases the amount of light per sensor while also raising the cost of the camera and complicating the light splitting optics.

2.3.3.4 Solid state sensors

Currently, there are two main methods for improving image sensor's dynamic range capabilities. One type of sensor accumulate charge produced from photocurrent while at the same time exposure duration is varied per pixel like in Lulé et al. (1999), the quantity of charge collected per unit of time is linearly related to their radiance on the chip like in conventional CCD chip Janesick et al. (1987). A second

kind of sensor calculates the logarithm of the irradiance in the analog domain using the logarithmic response of a sensor Seger et al. (1999) and Kavadias et al. (2000). Both the methods needs an appropriate analog-to-digital conversion and generate non-linearly sampled signals that are typically represented by using 8–16 bits per pixel value. These sensors have already been used in a number of HDR video cameras that are readily accessible in market. These cameras don't need to have their exposure times controlled, thus they may capture dynamic images with significant lighting changes. Although their pixel resolution is often poor and, for the logarithmic sensors, the visual noise in dark regions of the scene might be a concern, they typically offer a significantly wider dynamic range than multi-exposure video solutions Mantiuk (2015).

2.4 HDR reconstruction from conventional sensors

2.4.1 Single exposure HDR acquisition methods

The goal of single-exposure approaches is to increase the dynamic range without the aid of multiple exposure data, specialized equipment, or high end imaging methods. As a result, techniques in this category may be used with the great majority of LDR images and videos already exist, making it easier to have them in HDR applications. Decontouring, tone expansion, and restoration of underexposed and overexposed image regions are three unique subproblems that may be found in single-exposure reconstruction. Additionally, noise is a very important issue that degrades the information in the image's dark parts. Denoising, however, is a traditional and well-researched image processing work and is not particular to the single-exposure HDR image or video challenge.

2.4.2 Decontouring LDR images

The most common encoding for LDR pixels is 8 bits per color channel. When the dynamic range is increased, quantization may possibly produce banding artifacts that can be seen when Tone-mapping or viewing the content on an HDR display. Use of a dithering-based approach that adds noise to mask the artifacts is one way to solve the issue. Dithering can be done either before as done by Daly and Feng (2003) or after quantization Mukherjee et al. (2018). These techniques are designed to mask false contours at the same bit-depth as the input image. There are several filtering-based techniques that can enhance the bit-depth Daly and Feng (2003);

Song et al. (2016); Luzardo et al. (2017) as well. As an illustration, the method suggested by Daly and Feng (2003) quantizes the image at the input bit-depth after filtering the image. The difference between the filtered and quantized pictures, which is subtracted from the input image, represents false contours. Although they have certain drawbacks, bit-depth extension techniques can increase accuracy by 1-2 bits.

2.4.3 Tone expansion

The camera response function must be inverted in order to expand the dynamic range and transform the image tones to the linear domain when converting an LDR image to an HDR image. As the most typical objective for single exposure HDR techniques is to display LDR images or videos on HDR displays. Given that the result of the tone expansion TE is evaluated on an HDR display, it defines a composite mapping $TE = TMf^{1}$ where TM represents a Tone-mapping operation for the specific HDR display and f denotes the CRF. Additionally, as it is challenging to accurately recreate highlights, the ideal mapping TE can differ from what it would be if this information were accessible. Furthermore, since it is difficult to reconstruct highlights convincingly, the optimal mapping TE may be different than it would be if this information was available. A second common goal is to use the LDR images in image base lighting IBL. If highlight information is missing, a global boost in brightness generally yields an IBL rendering that is preferred over the otherwise too dark result. Consequently, tone expansion is, in general, a different matter than the inversion of a CRF, and the optimal end result may be very different from the true underlying HDR image Eilertsen (2018).

According to Banterle et al. (2006, 2007), an inverse tone-mapping operator (iTMO) is a technique usually used for increasing the dynamic range of LDR images. Numerous perceptual studies have indicated that a global mapping, either utilizing a gamma function Masia et al. (2017) or a linear scaling De Simone et al. (2014) may be preferred for the display of LDR images on HDR displays.

2.4.4 Recovering the details in under and over-exposed regions

Recovering a lost information in under-exposed and over-exposed regions is a real challenge when reconstructing an HDR image from a single-exposed image. Since the majority of HDR applications need bright image information but not the dark, over-exposure is typically the biggest challenge in HDR field. Many iTMOs make an effort to solve the issue by applying separate expansion to pixels that are classified

as saturated. For instance, Meylan et al. (2006) used several linear functions in saturated and un-saturated image regions. To build an expand map for enhancing highlights, Banterle et al. (2006) employed the median cut algorithm for recovering details. Additionally, the expansion map's cross-bilateral filtering was included, and the approach was expanded for use with video processing by Banterle et al. (2008). Rempel et al. (2007) come up with new expand map solution that optimizes estimation for real-time performance by utilizing a Gaussian filter. In comparison to Kovaleski and Oliveira (2009) earlier approach, Kovaleski and Oliveira (2014) employ a cross-bilateral expand map and come up with an iTMOs, that operates in a wider range of exposures. Didyk et al. (2008) proposed an alternative technique, in which the video frame components are decomposed into diffuse, reflections, and light sources using a semi-manual classifier. The components other than diffuse part is expanded with a wider dynamic range. These highlight boosting techniques are supposed to generate results that are more closely related to the real HDR image than the global iTMOs, which increases the dynamic range without explicitly taking saturated regions into account. The boosting, however, struggles to reconstruct fine details and colors in saturated image regions since it is just a mere rough estimate of luminance.

A second group of over-exposure correction techniques focuses on using statistics of adjacent non-saturated pixels for reconstructing colors and details in the over-exposed regions. Using information from the non-saturated channels of the same pixel, Zhang and Brainard (2004) applied Bayesian estimation to estimate the values of 1-2 saturated color channels of a pixel. Additionally, Xu et al. (2011) examined reconstructing pixels that had saturated color in all channels, the proposed techniques can handle greater portions of missing data. All of these exposure correction techniques have the limitation of just marginally increasing the dynamic range. High-intensity highlights, which are crucial for HDR reconstruction, are not taken into account.

2.4.5 Learning based single-exposure HDR image reconstruction

More recently, a number of methods employ deep learning strategies for single-exposure HDR image reconstruction. Eilertsen et al. (2017) approach the problem of HDR reconstruction from single exposure image using deep learning techniques. The CNN based encoder and decoder architecture reconstruct colors, intensities and details in saturated regions in a complete automatic way which were trained on a large HDR dataset. Through merging bracketed LDR images, Endo et al. (2017) indirectly recreate an HDR image from a single LDR input. The proposed model is composed of two processes: Learning and inference. The bracketed LDR

images are generated using HDR dataset during the learning phase by simulating different CRFs. Next, the up and down exposure networks are trained to learn changes in the exposures of the bracketed images. The learned models produce LDR images with various exposures from a single input LDR image during the inference phase. The two networks models generate brighter and dimmer bracketed images, which are then combined to create the final HDR image. The above method however, shows minimum performance in case videos and in more challenging over-exposed cases. The generic end to end network of LDR-to-HDR mapping using deep learning is decomposed in three subtasks by Liu et al. (2020) and developed three deep network for dequantization,linearization, and hallucination of missing details in the over-exposed regions. However, their approach cannot handle the noise in the dark regions and can only hallucinate smaller saturated regions. Furthermore, the aforementioned methods yield outputs with flickering artifacts and are not intended to handle videos. Eilertsen et al. (2017) proposed method, which shows how HDR videos may be produced from input LDR videos using only one exposure.But the performance of these methods regarding recovering details in over and under-exposed regions is still limited.

2.4.6 Time sequential multi-exposure techniques

Taking a sequence of images, each with a different exposure setting, is the least complicated way to capture high dynamic range (HDR) images. Even while an LDR sensor may only record a small portion of the whole luminance range of a scene at any given time, its functional range has the potential to include the entire luminance range if the exposure settings are adjusted appropriately. Because of this, the exposure of each image in a series is adjusted such that a varied range of luminance is recorded. After that, the images are merged into a single HDR image by using a weighted average of the pixel values over the exposures after taking into consideration a camera response and normalizing the exposure change Mitsunaga and Nayar (1999) and Robertson et al. (2003).

Reinhard et al. (2010) examine many methods that may be used to derive the camera response function. The inverted version of this function makes it possible to recover the irradiance values directly from the pixel values that correspond to in each input image. Gallo et al. (2012) conducts an analysis of the image histogram and then adaptively chooses a minimal number of exposures in order to take an image of the scene that has the best possible signal-to-noise ratio. The multi-exposure method makes use of a camera's full resolution and capture quality capabilities, able to capture scenes with an arbitrary dynamic range (see Fig. 2.4). This method requires a sufficient number of exposures to be taken for each frame. This methodology can be found in a wide variety of consumer devices, including

mobile phones.

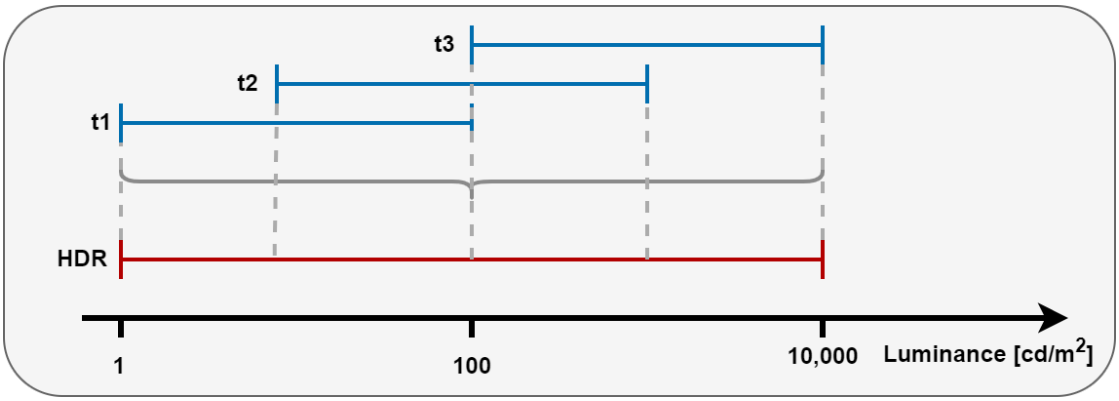


Figure 2.4: *Three consecutive exposures captured at subsequent time steps t1, t2, t3 register different luminance ranges of a scene. The HDR frame merged from these exposures contains the full range of luminance in this scene*

2.4.6.1 HDR video reconstruction solutions

With the increase of technical capabilities of digital cameras it is possible to alternate exposures between subsequent video frames, which in turn enables the application of multi-exposure techniques for HDR video. The problem of frame alignment to compensate for camera and object motion arises, but then solutions similar to deghosting which can be discussed in the following section, can be readily applied. An additional requirement in video case is temporal coherence between the resulting HDR frames. Real-time HDR video capture at 25 fps is effectively achieved using two alternating levels of exposure Kang et al. (2003) and Mangiat and Gibson (2011). Optical flow is used to unidirectionally warp the previous/next frames to a specific HDR frame in order to achieve frame alignment. In order to impose similarities between neighboring frames and improve temporal coherence ,the particular benefits of Kang et al. (2003), Kalantari and Ramamoorthi (2019) and Chen et al. (2021a) optical flow method in HDR image and video synthesis can be utilized along with recent deep learning based image reconstruction methods for more improved results. Furthermore, it is possible to improve the texture and motion synthesis in areas with constant motion. A different approach has been suggested by Unger and Gustavson (2007) that covers a significantly greater dynamic range of roughly 140dB but does not account for motion artifacts. A 200Hz camera with eight exposures per HDR frame was used to achieve this high dynamic range. More recent approaches, detailed in Guthier et al. (2012), likewise use high FPS cameras but account for camera motion. On the other hand, a shorter per-frame exposure time puts more constraints on sensor sensitivity, which often

causes noise in HDR video in terms of low light conditions Mantiuk (2015).

2.4.6.2 Deghosting for camera correction and object motion

The movement of the camera or other objects in the frame may cause certain image parts to be misaligned when merging multiple exposures taken at various times. The problem of alignment is resolved using global homography derived from applying RANSAC on SIFT or SURF features for proper alingnment of frames. However such homography-based alignment might not work in cases where there are large differences in frames due to the motion of objects and camera movement. Other methods rely on local motion detection and the weighting of each exposure contribution in accordance with the likelihood of such motion Khan et al. (2006). A general-purpose method suggested by Granados et al. (2013) for scenarios with large object displacement. Prior to reconstructing irradiance from pixels that are likely to belong to the same static scene object, they estimate the likelihood that a pair of colors in several images are observations of the same irradiance. This allows them to employ a Markov random field. A category of approaches reconstruct HDR videos from input sequences that are captured by alternating the exposure of each frame and use optical flow estimation to correct for object and camera motion. Following image alignment, some kind of color averaging is then carried out. Kang et al. (2003) propose the first HDR video reconstruction algorithm for sequences with alternating exposures by using optical flow to align neighboring frames to the reference frame. They then combine the aligned images with the reference frame using a weighting strategy to avoid ghosting. However, in cases with large motion, their approach typically introduces optical flow artifacts in the final results. Mangiat and Gibson (2010) improve Kang et al. (2003) approach using a block-based motion estimation method coupled with a refinement stage. In a follow up work, Mangiat and Gibson (2011) propose to filter the regions with large motion to reduce the blocking artifacts. However, their approach still shows blocking artifacts in cases with large motion Moreover, their method is limited to handling sequences with only two alternating exposures. Kalantari et al. (2013) propose a patch-based optimization system to synthesize the missing exposures at each frame. These images are then combined to produce the final HDR frame. To increase the temporal coherency, they estimate an initial motion between the neighboring and reference frames. They then constrain the patch search to a small window around the predicted motion, where the size of the window is obtained by a greedy approach. This method produces results that are generally significantly better than the other approaches. However, it usually takes several minutes to solve the complex patch-based optimization and produce a single HDR frame. Moreover, this approach is often not able to properly constrain the patch search and over/under-estimates the search window size. In these cases, it produces results

with ghosting artifacts or wobbly and unnatural motion.

Gryaditskaya et al. (2015) improve the method of Kalantari et al. (2013) by adaptively adjusting the exposures. However, the idea of adaptive exposures can also be used to improve our system. Finally, the recent method of Li et al. (2016) poses the HDR video reconstruction problem as maximum a posteriori estimation. Specifically, they separate the problem of HDR frame reconstruction by finding the foreground and background in each frame. They propose to find the background using rank minimization and compute the foreground using a multiscale adaptive regression technique. This approach is computationally expensive, their method produces results with noise, ghosting, and discoloration in more difficult cases. Kalantari and Ramamoorthi (2019) address the drawbacks of his previous approaches by proposing to use convolutional neural networks (CNN) to learn the HDR video reconstruction process from a set of training scenes. Specifically, their approach was builds upon their recent HDR image reconstruction method of Kalantari et al. (2017), which breaks down the process into alignment and HDR merge stages and uses a CNN to model the merge process. both the networks are trained in an end-to-and fashion by minimizing the error using L1 loss function between the reconstructed and ground truth HDR frames on a set of training scenes. Their optical flow improved the performance in term of temporal consistency than traditional optical flow methods, and learning-based flow estimation approaches. However, Their proposed method introduce decolorization and ghosting artifacts in high over-exposed cases.

Yan et al. (2019) propose an attention-guided deep neural network (AHDRNet) for HDR imaging. The proposed method does not apply optical flow similar to Kalantari et al. (2017); **?** and Chen et al. (2021a) as an intial step for image alignment instead they applied attention mechanism for alignment of content in neighbouring frames to reference frame. The learnable attention modules generate attention guided feature to guide the merging process for obtaining the required HDR images. The attention modules give importance to only those features which are similar to the reference image and exclude regions with motion and severe saturation. The LDR image features with attention guidance are then fed to the merging network with dilated residual dense blocks (DRDBs) The dilated convolutions enlarge the receptive fields, helping to recover the details contaminated by saturation and moving objects. Following the architecture of Yan et al. (2019), Liu et al. (2021) proposed attention guided deformable convolution network with a dual branch pipeline for multi frame HDR imaging of dynamic scenes with out the use of optical flow module. The LDR and gamma corrected images are process separately with dual branches. Specifically, a spatial attention module is used for extracting features for the LDR images through attention for better fusion, and a Pyramid, Cascading and Deformable (PCD) alignment module is adopted to align

the gamma-corrected images in the feature level. Such design is motivated by the intuition that the images in the LDR domain help detect the noisy or saturated regions while the gamma corrected counterparts help to detect misalignment's. The proposed methods struggle in case of video reconstruction due the temporal aspects since the alignment of frames through single attention and pyramid cascading deformable alignment (PCD) module is challenging.

Recently, Chen et al. (2021a) come up with a two-stage coarse-to-fine framework for HDR video problem. Their first stage aligns images using optical flow in the image space and blends the aligned images to reconstruct the coarse HDR video. recovering/removing a large part of missing details/noise from the input LDR images, For further improving the results from their first model which still left some artifacts due large motion, They passed the inputs from their first model to second model performing more sphisticated alignment fusion in the feature space of the coarse HDR video using deformable convolution Dai et al. (2017) and temporal attention with addition of using L_1 and perceptual loss for more robust feature extraction. Apart from that, Chen et al. (2021a) create a real-world video dataset containing both static and dynamic scenes captured with alternating exposures as a benchmark to enable quantitative and qualitative evaluation for this problem.

2.5 Storage and compression techniques

HDR image representation in their floating point format, entail significant storage costs. For instance, the common JPEG format needs between 0.7 MB and 3 MB to store a 15 mega-pixel image. However, storing the same resolution image in "RAW" HDR format would require around 200 MB Mantiuk (2015). This illustrates the significance of improving HDR image and video encoding and compression. The majority of the suggested HDR compression algorithms rely on the LDR video and image compression standards that are currently in use. It is necessary to convert the floating point HDR pixel values into a more effective representation utilizing the fewest possible bits in order to employ those compression standards successfully. Such HDR pixel representations are addressed in Section 2.5.1, and Section 2.5.2 presents the resultant HDR file formats. The discussion of various strategies for encoding HDR photos and video using current compression standards are mentioned in Sec. 2.5.3, and the discussion of backward-compatible solutions that also support regular 8-bit JPEG and MPEG formats comes up in Sec. 2.5.4.

2.5.1 Pixel formats and color spaces for HDR content

The color space and pixel encoding used to compress images or videos have a significant impact on the efficiency of compression and the encoding format's capabilities. The described encoding approaches attempt to reduce the number of bits needed while yet offering adequate precision and the capacity to represent a higher dynamic range. Banding (quantization) artifacts become apparent if the bit-depth precision is too insufficient. The most common HDR pixel encodings are described in the following section.

2.5.1.1 16-bit floating point numbers

Half-precision floating point numbers, often known as fp16 or S5E10, is a compact format used in graphics cards from nVidia and AMD. The floating point number is described by the code-name S5E10 as having a one-bit sign, a 5 exponent, and a 10 mantissa Mantiuk (2015), as seen in Fig 2.5.

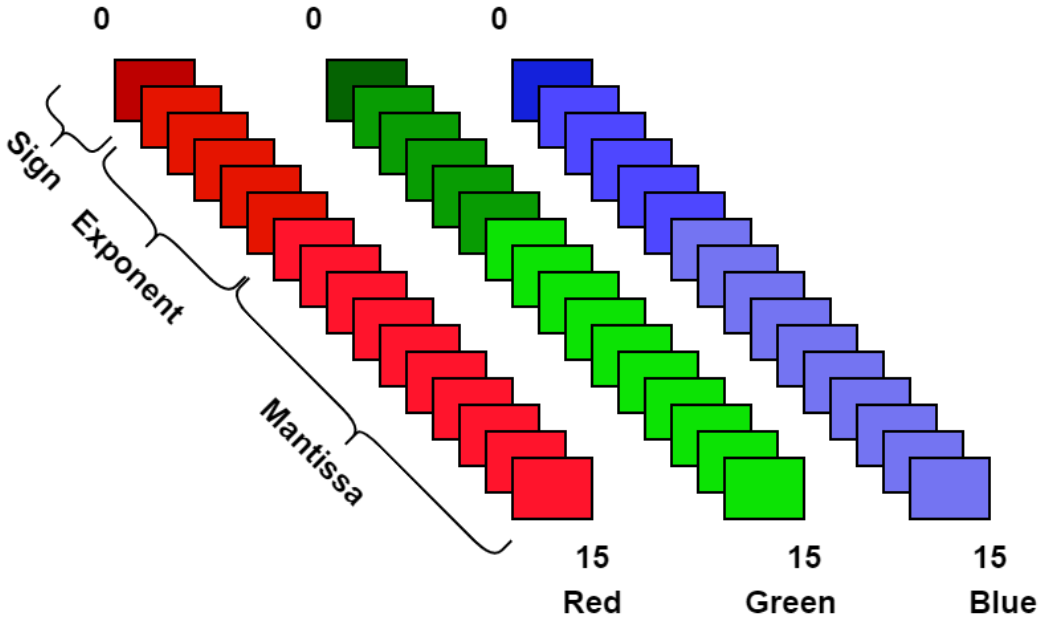


Figure 2.5: *Encoding of Red-green-blue image channels using half-precision floating point numbers.*

The OpenEXR picture format employs such 16-bit floating point encoding. As opposed to the standard 32-bit floating point format, the half-precision float delivers the flexibility of floating point numbers at a reduced cost for storage. Floating point

numbers are ideal for encoding linear luminance and radiance values because they can readily accommodate larger dynamic ranges. The greatest number that the half-precision float format can represent is 65,504, which is less than, for example, the luminance of bright light sources. Because of this, it is sometimes necessary to scale down HDR images with absolute luminance or radiance units by a fixed amount before saving them in the half-precision float format Mantiuk (2015).

2.5.1.2 RGBE: common exponent

Radiance file format uses RGBE pixel encoding. Red, green, and blue color channels are represented by the first three bytes of the RGBE pixel encoding, while the final byte is a standard exponent for all channels (see Fig. 2.6). The RGBE format, which employs 8 bits for the exponent and another 8 bits for the mantissa to express pixel values in floating point, is essentially a method of displaying images in this manner (8E8). The RGB color spaces' high correlation between all color channels and the fact that their values are almost the same order of magnitude are exploited by RGBE encoding Mantiuk (2015).

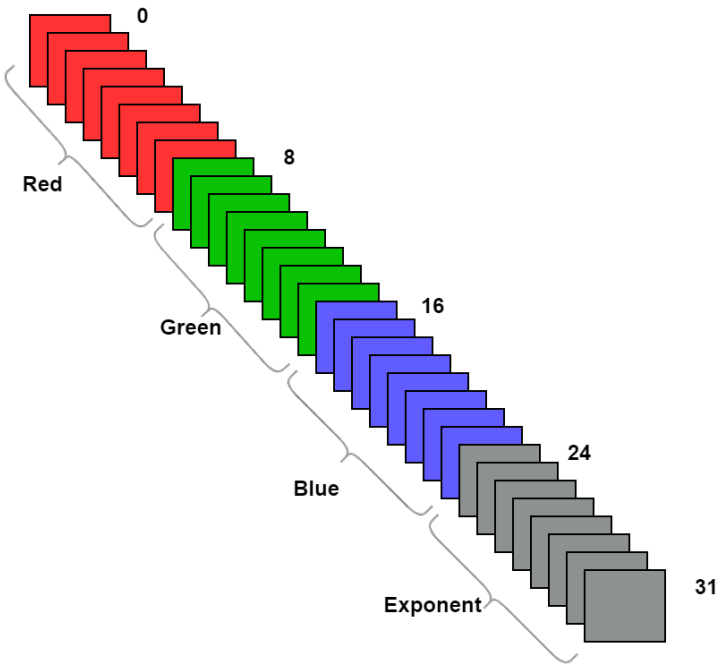


Figure 2.6: *RGBE 32-bit per pixel encoding.*

As a result, it is unnecessary to keep a different exponent for each color channel. The RGBE encoding has the drawback of not being able to express very saturated

colors outside of the Rec. 709 (sRGB) color space. When such intense colors are translated to the RGB color space, one or more of its color components become negative. since negative values cannot be expressed using the RGBE standard, significant color information is lost Mantiuk (2015). To address this issue, the Radiance format additionally supports XYZE encoding for pixels in the CIE XYZ color space.

2.5.1.3 LogLuv: Logarithmic pixel encoding

Floating point numbers have the drawback of not being the best for image compression techniques. This is partially due to the fact that more bits are needed to encode the mantissa and exponent separately instead of integer value. Color data does not require such a flexible format. Furthermore, unlike our visual system's "precision," floating point numbers' precision error fluctuates across the entire range of possible values. Integer numbers can therefore be utilized to encode HDR pixels for greater compression. In order to capture the entire luminance and color gamut that is discernible to the human eye, Larson (1998) introduced a new approach for encoding high dynamic range digital pictures utilizing log luminance and uv chromaticity.

The suggested format offers enormous benefits to the users of digital photography while requiring little extra storage per pixel which is supported by TIFF library as an optional encoding. The human eye is not equally sensitive to all luminance levels, and the proposed logarithmic encoding by Larson (1998) took advantage of this human vision capabilities as our perception of difference varies in the dark and in the sunlight and this phenomenon is known as luminance masking. The detectable threshold values, however, do not fluctuate as much when the logarithm of luminance is taken into account instead of direct luminance, and a constant value can be a reasonable approximation of the visible threshold . Luminance and chromaticity are each represented by two bytes in the 32-bit LogLuv encoding (see Fig. 2.7). The CIE 1976 Uniform Chromacity Scales (u, v) are used to encode chrominance. There is a LogLuv encoding variation that employs just 24 bits per pixel while maintaining enough accuracy. Due to discontinuities brought on by storing two chrominance channels with a single lookup value, this format may not be useful for arithmetic coding compression.

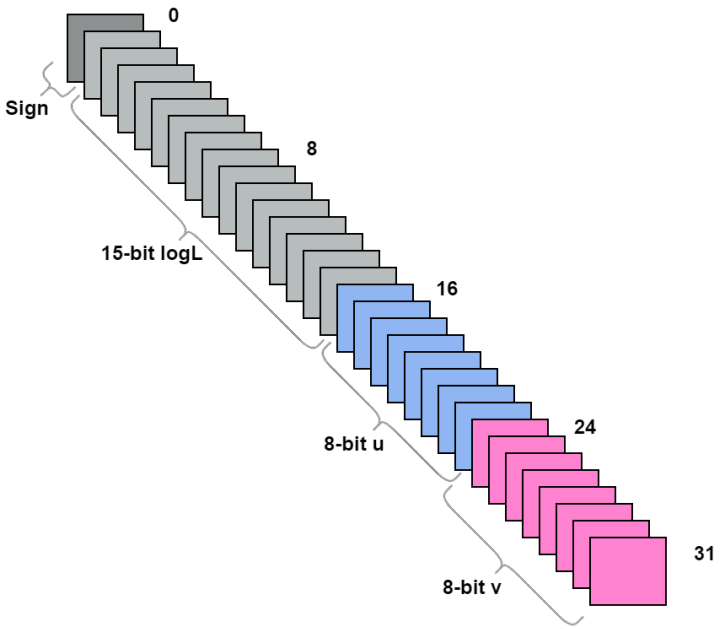


Figure 2.7: *LogLuv 32-bit per pixel encoding.*

2.5.1.4 Perceptually uniform encoding

The ideal characteristic of LDR pixels is that their values are approximately linearly related to the perceived brightness of those pixels. Due to the fact that the distortions produced on by image compression have the same visual effect over the whole range of signal values, LDR pixel values are thus well suited for picture encoding. As a result, when the same amount of distortion is introduced in low-luminance and high-luminance picture parts, the artifacts are more obvious in the low-luminance regions because HDR pixel values lack this feature. The issue can be solved by encoding logarithmic luminance rather than just luminance, similar to the LogLuv encoding explained above. However, logarithmic luminance is not a precise representation of the human visual sensitivity to light, the logarithmic encoding does not totally resolve the issue. Due to this, more precise encodings were proposed by Mantiuk et al. (2004); Miller et al. (2013) to model the human visual sensitivity to light changes. Such perceptually uniform encoding (see Fig. 2.8) is derived in a manner similar to that of the visual system's response to light, The derived function converts real luminance values in cd/m^2 into units related to just noticeable differences (JNDs).

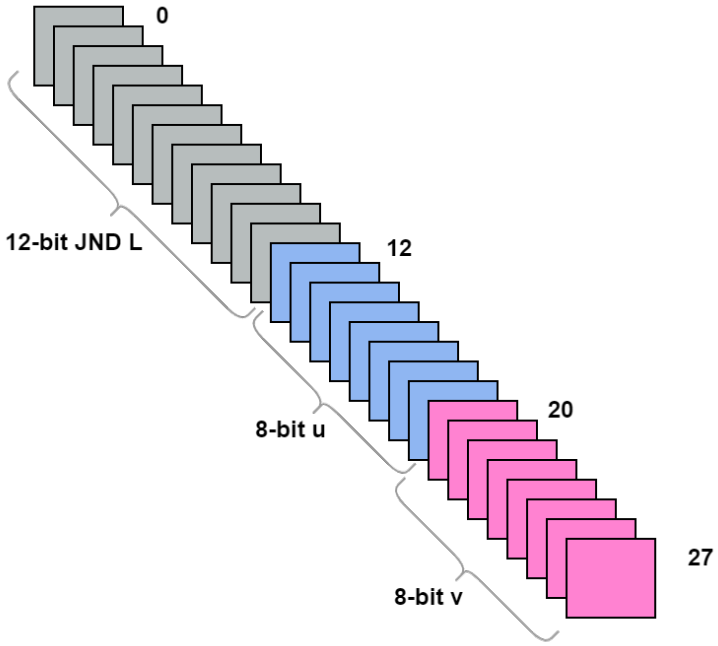


Figure 2.8: *JND 28-bit per pixel encoding.*

Each encoding method is an effort to strike a compromise between the storage to encode a larger dynamic range and the accuracy at which such a range is encoded. The encoding process produces quantization errors due to the insufficient precision, and these flaws manifest in images as banding (contouring), especially in regions with smooth gradients.

2.5.2 HDR image file formats

A variety of file formats are available that are specifically made for HDR images. An HDR image format's primary intention is to store the linear pixel values with floating point precision, for example in the RGB color space. However, if 32-bit floating-point integers are utilized, this implies that 96 bits per pixel (bpp) must be used to represent colors. Without any compression this will produce a very large HDR file. Just because of that, smaller pixel representations are used in floating point HDR image formats.

The two most popular HDR image formats are Radiance HDR and OpenEXR, which will be discussed in the below section.

2.5.2.1 Radiance's HDR format

In 1989, the Radiance rendering software included one of the earliest HDR image formats referred to as the Radiance image format with the .hdr or.pic file extension Mantiuk (2015). The radiance image format includes a short text header which is followed by run-length encoded pixels. The HDR pixel values are encoded using XYZE or RGBE pixel formats, The RGBE employs red, green, and blue primary whereas the XYZE uses the CIE 1931 XYZ primaries, Because of this, the XYZE format is able to encode the whole visible color gamut, in contrast to the RGBE format, which is only able to encode chromaticities that fall inside the triangle produced by the red, green, and blue color primaries of the Rec. 709 color gamut.

2.5.2.2 OpenEXR

Industrial Light and Magic developed and released an open source C++ library in 2002 that included the OpenEXR format, also known as (the EXtended Range format) and known by the file extension.exr). Since then, the format has been adopted by a wide range of Open Source and commercial applications, and it has evolved into a de facto standard for HDR images, especially in the special effects field. This format has several properties like:

- High Dynamic range and color precision.
- supports 16, 32-bit floating-point, and 32-bit integer pixels values.
- Several lossless and lossy picture compression techniques are supported. On images with film grain, several of the available codecs may achieve 2:1 lossless compression ratios. The visual quality and decoding efficiency of the lossy codecs have been improved.
- Flexible in customization and expandability. By modifying the C++ classes supplied in the OpenEXR software release, new compression codecs and new image formats may be simply added. Without compromising backward compatibility with already-existing OpenEXR applications, new image features can be added to OpenEXR image headers.

Although the OpenEXR file format supports a variety of data types for channel encoding, color data is often encoded with 16-bit floating point numbers, sometimes known as half-precision floating point.

2.5.3 High bit-depth encoding for HDR

While the floating point formats can distribute high-quality HDR pixels, the file size is still large compared to common LDR formats. This is especially, problematic

for video sequences, as these HDR formats do not explore inter-frame correlations. While this can be accepted in the industry, where quality is a high priority, it is not feasible e.g. for HDR TV streaming. Any common compression format that supports higher bit-depths can be used to store HDR images and video in addition to the unique file formats covered in previous sections. It is simple to add HDR content to existing image and video compression standards which are mainly built for LDR content but they all include an additional option for higher bit-depths. For instance, the JPEG2000 standard enables up to 16 bits, whereas the high-quality content profiles provided in the MPEG4-AVC/H.264 video coding standard permit encoding up to 14 bits per color channel Sullivan et al. (2007). The most current JPEG XR image compression standard also supports higher bit-depths up to 16 or 32 bits. For HDR applications, such bit-depths are more than enough. Fig.2.9 provides an illustration of how the current standards have been expanded to support HDR. Pixels must first be encoded using one of the pixel encodings in order to use the current HDR compression. By doing this, the perceptual uniformity of the introduced distortions is improved while simultaneously lowering the number of bits needed. The highest performance is achieved using perceptually-uniform encodings proposed by Mantiuk et al. (2004); Miller et al. (2013). logarithmic and floating point coding can also be adopted.

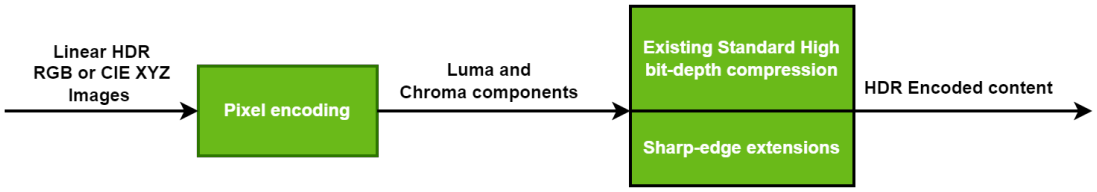


Figure 2.9: *Encoding HDR image or video content flow using standard high-bit-depth codecs, such as JPEG2000, JPEG XR or selected profiles of H.264. The HDR pixels need to be encoded into one luma and two chroma channels to ensure good decorrelation of color channels and perceptual uniformity of the encoded values. The standard compression can be optionally extended to provide better coding for sharp-contrast edges.*

There are several benefits to this strategy. First, LDR codecs have evolved for a long time and are today very efficient. Second, by employing an LDR codec it is easy to enable support of HDR material in existing software, and also to allow for backward-compatibility. Moreover, LDR codecs rely on integer pixel representations, which allow for better compression properties as compared to floating points.

2.5.4 Backward-compatible compression techniques

Since almost all software and hardware associated with digital photography supports the common low-dynamic range (LDR) file formats for images and video, such as JPEG or MPEG, it cannot be expected that these formats will be instantly replaced by their HDR equivalents. Backward compatible HDR formats that are fully compatible with existing LDR formats and offer extended dynamic range and color gamut are required to make transition from conventional to HDR imaging in a straightforward manner. Furthermore, the cost of HDR information must be minimal for such a format to be viable and widely used.

2.6 Tone mapping

The process of rendering scenes with high contrast and possibly large color gamut on a media with restricted contrast and color reproduction is known as tone mapping. In most cases, it involves transforming high dynamic range images (or animation frames) that represent scene radiance or luminance into pixel values that can be shown on a computer display. Tone-mapping algorithms, however, focus on a wide range of objectives, strategies, and applications.

2.6.1 Categorization

The fundamental mechanics and assumptions of many of these operators are, nevertheless, very similar but overall, in terms of processing, tone mapping operators are divided in to two parts.

2.6.1.1 Global tone mapping operators

The "tone reproduction curves" or global tone mapping methods are spatially invariant, meaning that the processing is identical for all the pixel values with in a frame despite the variations in values of its neighbouring pixels Mantiuk et al. (2008) and Van Hateren (2006). The tonemap pipeline first acquires the luminance of image before computing global statistics. Imaging systems, capturing content at 30 or 60 frames per second, some algorithms additionally calculate these statistics from the previous frame under the assumption that there aren't many changes between them. After getting the global statistics from luminance of the image the tone-mapped image is computed, by using a logarithmic or exponential-like function in the pipeline. These functions are difficult to implement on hardware, thus a frequent solution is to approximate them while allowing for a tolerable degree

of error in the actual implementation. After tone mapping, the output image's color must be restored in order to be displayed.

2.6.1.2 Local Tone Mapping Algorithms

Local tone mapping methods, which are often referred to as "tone reproduction operators," are spatially variant and the processing may vary based on the region around the input pixel Reinhard et al. (2012) and Boitard et al. (2012). In comparison to global tone mapping methods, local tone mapping techniques are computationally more expensive and time-consuming . Apart from processing point of view it is important to distinguish the TMOs which are build for single images and those which are explicitly designed to work with video. Since TMOs for static images do not ensure temporal coherence of pixel values, which for example can result in severe flickering artifacts, It is necessary for tone mapping method to include the temporal aspect of the video for getting rid of artifacts which is specific to temporal features of video.

2.6.2 Intents of tone mapping

Depending on the application and context, tone mapping's might vary greatly. It is crucial to explicitly define these aims since the variety of them is the cause of a lot of tone-mapping misconceptions. According to their intended use, Eilertsen (2018) classified tone-mapping operators roughly into the following categories.

2.6.2.1 Visual system simulators

Tone mapping operators main goal is to mimic the HVS capabilities. Since the HVS has a higher dynamic range than a standard camera, which mean a large of visual information is collected when compared to a typical LDR images.The image that minimizes the perceived difference between the tone-mapped image and the original captured scene is the ideal outcome of a Visual system simulators. Apart from that it should also consider and mimic the limitations of the HVS, such as decreased color saturation, glare and loss of acuity in low light situations. When comparing the tone-mapped image to the original captured scene, the image with the least noticeable variation is the best outcome of a Visual system simulators. Ferwerda et al. (1996) introduced one of the earliest Visual system simulators tone-mapping methods. Based on a series of psychophysical experiments, HVS's adaptation mechanism is modelled. One of the most detailed perceptual models for tone-mapping was brought out by Pattanaik et al. (1998). It takes both threshold and supra-threshold perception into consideration and employs a

multiscale representation of luminance, detail, and color processing of the HVS. The response of the captured HDR scene was simulated by Pattanaik et al. (2000) in subsequent work by combining adaption and appearance models. An LDR display device can be used to display the HDR simulated response by inverting the models. Although real quantitative measurements have also been employed in some cases, Apart from the majority of Visual system simulators approaches which are based on data from psychophysical experiments.For instance, Van Hateren (2006) used a model that was created using data from measurements made on macaques retinas.

2.6.2.2 Scene reproduction operators

When displaying an image on a device with a limited color gamut, contrast, and peak brightness, Scene reproduction operators make an effort to maintain the scene's original look, including contrast, sharpness, and colors. These operators will not attempt to replicate visual changes based on by perceptual phenomena, including night color vision deficiency and acuity loss. Instead, they concentrate on overcoming the output medium's limitations and work to produce the best match possible given the medium's constrained gamut and dynamic range.

A technique for maintaining the apparent, or perceived, brightness of the HDR image was put forth by Jack and Holly (1993) when tone-mapping was first introduced to the computer graphics community. Using a global scaling factor, Ward (1994) attempted to maintain the contrasts from the HDR image. However, because this technique essentially adjusts exposure automatically, a lot of the visual information is lost in dark and saturated regions of images. Given that the dynamic range is constrained to a particular display technology, another strategy is to work toward reducing contrast changes by Mantiuk et al. (2008).

2.6.2.3 Best subjective quality Tone-Mapping operators

To generate the image that is most desirable upon visual inspection is a main objective for tone-mapping. Specifically, it refers to the image that, when not compared to the reference HDR image, has the highest subjective quality.In terms of personal preference or artistic purposes, tone mapping operators are built to generate the most desired images or videos. These operators frequently have a set of adjustable parameters that may be adjusted in accordance with artistic goals. Resulting in tone-mapping which extremely varied depending on the specific application and the person who is assessing the result.Photo editing programs like Adobe Photoshop Lightroom are a fantastic illustration of such Tone mapping operators. TMOs that meet the criteria for beingBest subjective quality operators span a wide range, some of the most often mentioned operators in the literature are Reinhard et al. (2002); Fattal et al. (2002); Drago et al. (2003)

There are applications that do not fit easily into any of the aforementioned categories, and the applications described above may not fully cover all potential aspects of tone-mapping. However, the intentions describing the variations in the underlying expectations and assumptions for tone-mapping was that, it helps to partially explain why there isn't a single "optimal" tone-mapping as it is crucial in investigations comparing operators that achieve two quite distinct goals.

2.7 HDR displays

Due to the physical limitations of the existing display devices, simulating and reproducing the real world appearance is challenging. The discretization of information in pixel and images does not reflect the continuous nature of spatial and temporal information. But at the same time Display devices are not required to to meet all the requirements since human visual systems (HVS) has its own physical limitations. For instance, limited amount of photoreceptors in retina and flaws in human eye optic system and with the ability of observing details in 60-70 cycles per visual degree and with addition, the ability to differentiate temporal signals is restricted by temporal critical flickering frequency over 60Hz Wandell (1995). These limitation of human visual system (HVS) is taken in to account when it comes to designing display devices.

Recent HDR display devices have specifications which offers contrast and brightness close to the HVS limitations and support HDR encoded material. But still display devices requires Tonemapping in order to map the HDR content to the particular display because there is still limitations in display capabilities in terms of brightness, black level, and color gamut.

2.7.1 Professional HDR display devices

Professional HDR display usually follow two approaches to be able to show content with high contrast, brightness and color gamut.

- Modulating each individual pixel directly over a wide range of luminance.
- Combining two or more modulators sequentially to produce the same effect.

The first method needs its pixel to be handle in 12 to 16 bit depth precision which is a technological challenge. Considering that, zero luminance and a high luminance value (ideally, 3,000–6,000 cd/m^2 Seetzen et al. (2006)) should be easily accessible without causing significant light leaks between adjacent pixels. Scanning Laser

Display Technology created by JENOPTIK GmbH Deter and Biehlig (2004), meet the above criteria's by modulating the amplitude of the RGB laser beams which directly reproduced bright and dark pixels. The use of laser technology not only increased the color gamut that result from the wavelength of lasers providing more saturated and sharp color primaries but also making a seamless transition between adjacent pixels without pixel discernible boundaries. The fact that there is no light in black pixels, which causes the contrast ratio to be higher than 100,000:1, However, such systems require incredibly costly high power laser diodes, they are relatively uncommon Mantiuk (2015). The uses of organic light emitting diodes (OLED) in HDR displays are also promising. Although the maximum luminance level is still a limiting element and no OLED display with a driver capable of 12-16 bit depth has yet been demonstrated, it is relatively easy to obtain the zero luminance value by turning off each diode.

The second approach considering two modulators is more practical. The optical multiplication of two independently modulated representations of the same image using dual modulation already lead to high quality HDR image displays. While only standard 8-bit drivers are used to control the pixel values in each modulator, the contrast of the final image is effectively a product of the contrast attained for each component image. The so-called backlight device acts as the first modulator directly by actively generating a quantity of light that can be controlled spatially, similar to what would happen with a grid of light emitting diodes (LEDs). The second modulator, a passive transmissive LCD screen (liquid crystal display), which regulates the quantity of transmitted light per pixel, is illuminated by the backlight device. A projector may also be utilized as the backlighting device, however this requires a powerful light source because two passive layers will modulate the transmitted light. Such projectors might use a variety of light modulation technologies, including transmitive LCDs, reflective LCDs , and Texas Instruments' digital micro-mirror devices (DMDs), also known as digital light processing, or DLP. Backlight devices in HDR displays are based on both passive and active light modulation theories. Seetzen et al. (2003) and Seetzen et al. (2004) have for the first time investigated both display design possibilities.

The first method, uses a DLP projector to create a modulated backlight that travels via a Fresnel lens and collimates before striking the LCD display. The Fresnel lens's diffuser prevents moiré patterns from forming. This design produces a contrast of 54,000:1 and a peak brightness of 2,700 cd/m^2. By focusing the projected image on the rear of the LCD panel, Wanat et al. (2012) created an updated version of this concept with a projector that has a contrast that is five times better and significantly reduces blur. This makes it possible to reproduce high luminance contrasts across a wide range of spatial frequencies, which is critical for accurately representing sophisticated luminaires and highly specular materials.

Ferwerda and Luka (2009) employed a tiled array of cheap DLP projectors that were geometrically and colorimetrically calibrated to match the front LCD panel's high resolution (2,560 x 1,600). In the ultimate case, two identical, precisely aligned, high resolution LCD screens may be directly stacked to eliminate the projectors method proposed by Guarnieri et al. (2008). Using the same image to drive both panels eliminates the requirement for any image processing stage, resulting in sharp images with a stunning 50,000:1 contrast and a peak brightness of 500 cd/m^2. Although color filters were removed from both LCD panels, Guarnieri et al. (2008) designed their display having grayscale images for medical purposes, therefore it is unclear if the alignment precision is adequate for registering RGB components as well. The backlight for the full HD (1,920 x 1,080) LCD panel is generated by a hexagonal close-packing matrix of 1,200 independently modulated light emitting diodes (IMLEDs) designed by Seetzen et al. (2004). This design has an impressive 200,000:1 global contrast, but the black and white checkerboard pattern's calculated ANSI contrast only reaches 25,000:1 with a peak luminance of 3,000 cd/m^2. Although white LEDs were employed in the initial concept, subsequent expansions have demonstrated a considerable increase of the color gamut for integrated RGB LED packages. It's interesting to note that a backlight device based on LEDs is 3-5 times more energy-efficient than a standard LCD display using uniform light at a same brightness Mantiuk (2015). One of the main elements driving the usage of IMLED technology in commercial LCD TV sets, is the power efficiency, aside from enhancement in contrast and black level

Today's consumer TV industry is dominated by ultra HD, with the majority of TVs having a 4K resolution specification. Now, some new modern displays have 8K resolution. The present emphasis is on maturing in the dynamic range domain, continuing the prior trend of enhancing spatial information. The TV business, which has undergone significant growth in recent years, now includes a new class called TVs with HDR capability. In order to attain a higher dynamic range, research and development now focuses on increasing peak brightness and enhancing local dimming strategies.Additionally, there is now work being done to standardize the HDR format (see Section 2.4). Similar to professional high performance HDR displays, most HDR TVs employ back-light modulation for local dimming. but the utilized back-lighting system suffers in terms of brightness and precision. Recent commercial display panels use LCD modulation, and mounted LEDs are specifically used for back-lighting allowing a more compact and cheaper display solutions

There is no doubt that HDR content and HDR displays are becoming more and more common on the consumer market. Improved backlighting and local dimming methods as well as single-modulation solutions will be possible in the future. As a result, certain HDR TVs with higher resolutions may soon surpass in performance

regarding the dynamic range and brightness capabilities of existing professional devices.

2.8 HDR image quality

Estimating human perception of image and video quality is the primary objective of image and video quality assessment (IQA, VQA). Practical data demonstrates that numerical distortion measurements, such as root mean squared error (RMSE) or (MSE), are frequently insufficient for comparing images because they inadequately predict the variations between the images as experienced by a human observer Wang and Bovik (2006) and WU HR (2005) . Different image and video quality metrics (IQM, VQM) have been developed to effectively handle this problem. A human observer may quickly determine which of the two video clips is more appealing, but it is sometimes impossible to do extensive subjective tests on a variety of video clips and algorithm parameter variations. In order to substitute time-consuming trials, there is a need for computational metrics that might forecast the quality of visually significant differences between a test image and its reference. Assessment of image and video quality is useful in many contexts. For instance, The major use case of IQA is to monitor the image quality in lossy image/video compression, imaging application benchmarking, and algorithm optimization through parameter adjustment. Additionally, image and video quality metrics have been effectively used to evaluate the perceptual impacts of multiple computer graphics and vision algorithms as well as image database retrievals. The majority of image quality metrics take into account evaluating the quality of a single media, like an LCD display or a print. The results of physically realistic computer graphics techniques, however, are not connected to any specific device. In contrast to a display device's gamma-corrected RGB values, they create pictures in which pixels contain linear radiometric values . Additionally, real-world scene radiance values can have a very wide dynamic range , which is larger than the contrast range of a common display device. As a result, the issue of comparing the quality of such images or videos, which show real scenes rather than their tone-mapped replicas, arises.

2.8.1 Display-referred and luminance independent metrics

Whether the images are provided in relative or absolute luminance units, quality metrics for HDR images and videos differ between the two. The display-referred metrics anticipate that the values in the displayed images will match the absolute luminance throughput from an HDR or LDR display. They explain why the

distortions in darker areas of the picture are less obvious. Perceptually uniform encoding, HDR−VDP and HDR−VQM, are some HDR image and video quality techniques which are discussed in th following section. Any relative HDR pixel value can be used with luminance-independent metrics, and multiplying values by a constant yields the same results. They typically convert the values of HDR pixels into the logarithmic domain under the assumption that the observer's sensitivity to light follows the Weber law. One illustration of such a metric is log-PSNR, which uses the formula for normal PSNR but computes for logarithmic values instead:

$$logPSNR = 10.log_{10}\frac{log_{10}(L_{max})}{MSE} \tag{2.1}$$

and

$$MSE = \frac{1}{N}\sum_{i=1}^{N}[log_{10}(L_t(i)) - log_{10}(L_r(i))]^2 \tag{2.2}$$

where

$$L_t(i) = max(L_t(i), L_{min}) and L_r(i) = max(L_r(i), L_{min}) \tag{2.3}$$

here N in equation 2.2 is the total number of pixels in the image and $L_t(i)$ represents test image i_{th} pixel luminance where $L_r(i)$ represents reference image i_{th} pixel luminance. The minimum luminance above noise level is represented by L_{min}. Without such clamping of the lowest values will produce large error for dark and noisy pixels in the image Mantiuk (2015). and L_{max} represents selected peak luminance value which is usually set to 10 000, as some of HDR displays exceed this peak luminance level. It must be noted that the maximum pixel value in an image should not be chosen as L_{max}, as by doing that such metric would become image dependent.

2.8.2 Perceptually-uniform encoding for quality assessment

The use of PSNR and SSIM Wang et al. (2004) metrics with HDR images is made possible by the straightforward luminance encoding provided by Aydın et al. (2008). Physical brightness values (measured in cd/m^2) are converted by the encoding into a roughly perceptually uniform representation. To simulate the sRGB non-linearity, the transformation is further restricted to translate the brightness values generated by a typical CRT display (between 0.1 and 80 cd/m^2) to the range of 0-255. This will allow typical low dynamic range image quality predictions comparable to those

which uses pixel values. However, the metric has the capability to operate in a much wider range of luminance.

2.8.3 Visual difference predictor (VDP) for HDR images

Mantiuk et al. (2011) proposed HDR-VDP-2 quality metric which is able to identify variations in achromatic images with a wide range of absolute luminance values. The visual models utilized in this measure are substantially different from those used in previous metrics, even though they have a common origin with the classical Visual Difference Predictor Daly (1992) and its extension, HDR-VDP Mantiuk et al. (2005). The metric is also an attempt to create an extensive model of the contrast visibility for a very broad variety of illumination conditions.

The dispersion of light in the eye's optics and on the retina is one of the main factors limiting how much contrast may be seen in high contrast (HDR) scenes McCann (2008). It is modeled by the HDR-VDP-2 as a frequency-space filter and was fitted to the relevant data set specifically (inter-ocular light scatter block). Lower luminance levels, when rod photoreceptors, which are primarily used for night vision, mediate vision, lead to deterioration in contrast perception. Small contrasts that are near to the detection threshold are particularly affected by this, according to research. This phenomenon is treated as a hypothetical steady-state response of the photoreceptor to light. According to the measurements of the contrast detection algorithm, such response decreases the magnitude of the image difference for low luminance. To predict the threshold elevation caused by contrast masking, the masking model operates on the image split into several orientation- and frequency-selective bands. Such masking is caused by contrast in both the same band (intra-channel masking) and nearby bands (inter-channel masking). The same masking model also takes into account the effect of neural CSF, which is the contrast sensitivity function without the sensitivity reduction caused by interocular light scatter. To account for contrast constancy, which causes the CSF to "flatten" at the super-threshold contrast levels, one must combine neuronal CSF with a masking model. Overall, The visual difference predictor is made up of two visual models that are exactly the same and are used to process the test and reference images, respectively. Most often, a feature that has to be detected is present in a test image while it is absent in a reference image.

2.8.4 HDR-VQM

HDR−VQM is a full-reference HDR video quality metric proposed by Narwaria et al. (2015). The method is based on signal pre-processing, transformation, frequency based decomposition and subsequent spatio-temporal pooling. However, the main difference resides in the application area. HDR−VQM targets the signal processing, video transmission and related fields, where the distortion of the signal is often considerable and the information about the overall video quality is thus an expected and sufficient measure. Accordingly, HDR−VQM aims to predict human perception of the supra-threshold video distortions, which are then pooled to a single number, a measure of an overall video quality.

Transformation into emitted luminance:
The input videos are first converted into the luminance values that the display device produces. This is a challenging issue as the HDR values recorded in the HDR video are often not calibrated (i.e. relative), and as a result, they are only proportional to the input luminance. In addition, the accurate display processing model is typically unknown. Instead, the developers of HDR−VQM use the following basic approximation with a scaling factor: The highest 5 percent of the mean of all the HDR values in the video sequence are used to normalize the input HDR videos. To imitate the display's physical constraints, a clipping function is finally applied.

From emitted to perceived luminance
The second stage simulates the human perception of the emitting brightness , which is known to be approximately logarithmic and nonlinear. The perceptually uniform (PU) encoding suggested by Aydın et al. (2008) was utilized to simulate this behavior in HDR-VQM. Making differentials of the curve proportionate to the luminance detection thresholds is the main goal of PU encoding. Even while it is believed that the PU encoding would represent the HVS more accurately than a straightforward logarithmic function, it is still merely a rough estimate of the HVS luminance response.

Decomposition into visual channels:
Following that, visual channels are divided up according to perceived brightness. The decomposition that has been put in place only distinguishes between spatial frequencies and orientations, leaving temporal processing until the last stage of pooling, in order to maximize efficiency. As a result, the spatio-temporal contrast sensitivity (CSF) of the human visual system cannot be predicted. More specifically, the implemented frequency domain log-Gabor filters Field (1987) provide the foundation of the decomposition that is being used.

Pooling:
The last step of HDRVQM is error pooling which is achieved via spatio-temporal processing of the sub band errors. This comprises of short term temporal pooling, spatial pooling and finally, a long term pooling. After performing spatial and long

term temporal pooling the final the global video quality score is achieved Narwaria et al. (2015).

3 Methodology

3.1 Overview of the study

We approach to the problem of HDR video reconstruction by using the Yan et al. (2019) work as a baseline architecture which is a learning-based technique. The baseline architecture of Yan et al. (2019) is improved by integrating multiple attentions modules for image alignment in relation to the problem of ghosting artifacts. For further refining the aligned features of attention modules, PCD alignment module is utilized. Robust merge network composed of DSKFRDBs is used to recover the missing content in the saturated regions. We trained our proposed method along with Yan et al. (2019) on synthetic dataset composed of multiple publicly available HDR datasets with a simulated limitations of conventional camera systems. We use three and five 6 channel frames as input to the models and the models estimate either 3 channel HDR image or produce a 15 channel output which is then used for pixel blending for getting final 3 channel HDR image in different experiment and checked the performance of our proposed attention modules compared to Yan et al. (2019). Our approach to the problem of HDR video reconstruction is as follows:

- We train and evaluate our Multi-Attention SKFHDRNet without optical flow and pixel blending module to check whether our attention modules are more robust than Yan et al. (2019) single attention. We also train the Yan et al. (2019) work in a similar way on a synthetic training dataset discussed in (section 3.2).
- The proposed Multi-Attention SKFHDRNet and Yan et al. (2019) are then trained with optical flow module with out pixel blending.
- Multi-Attention SKFHDRNet and Yan et al. (2019) architecture are trained with optical flow module with addition of pixel blending strategy.
- Multi-Attention SKFHDRNet with PCD alignment module are trained along with Yan et al. (2019) architecture with optical flow network with addition

of pixel blending strategy.

- Multi-Attention SKFHDRNet with PCD alignment module using a combined L_1MS−SSIM loss are trained with optical flow network and used a pixel blending strategy.

The overall flow of our approach to the HDR video reconstruction can be seen in (Fig. 3.2)

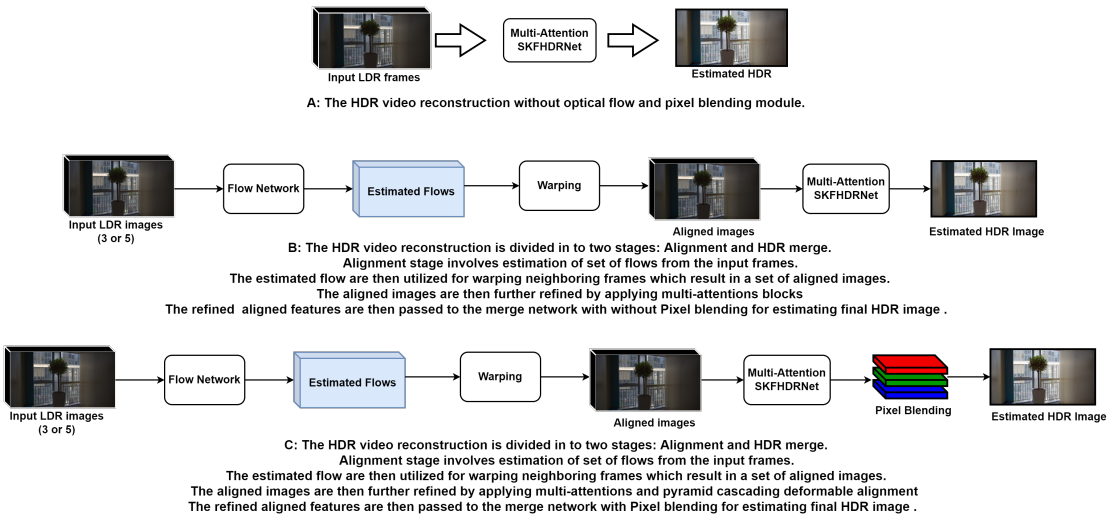


Figure 3.1: *A: Reconstruction of HDR video with out optical flow using Multi-Attention SKFHDRNet B: Reconstruction of HDR video with optical flow using Multi-Attention SKFHDRNet C: Reconstruction of HDR video with optical flow using Multi-Attention SKFHDRNet and utilizing the blending weight strategy used by Kalantari and Ramamoorthi (2019); Chen et al. (2021a)*

3.2 Dataset overview

3.2.1 Synthetic dataset for training

For training our HDR deep learning model, A large number of training samples is required which consist of three or multiple consecutive LDR frames where a center reference frame along with two neighbouring frames with alternating exposures and their corresponding ground truth HDR. Following the work of Kalantari et al. (2017); Kalantari and Ramamoorthi (2019); Chen et al. (2021a), We also used publicly available 13 HDR video scenes from Froehlich et al. (2014) please refer to Table A.1 in appendix for specification of the used Froehlich et al. (2014) dataset

scenes. Eight downsampled video scenes of resolution 1280x720 from Kronander et al. (2014) are used for training purpose. Both the datasets are captured in different environments considering different scene illuminations. Each scene in these datasets is represented by hundreds of HDR frames that were taken with cameras that had certain optical configurations that included either external Froehlich et al. (2014) or internal Kronander et al. (2014) beam-splitters. Additionally, Chen et al. (2021a) also used high-quality Xue et al. (2019) Vimeo-90K dataset as training samples due to the limited size of the training HDR video dataset. Xue et al. (2019) Vimeo-90K dataset is preprocessed 91,701 clips and each clip is composed of 7-frames with an image resolution of 448x256. Following the methodology of Eilertsen et al. (2017), Chen et al. (2021a) The LDR frames of this dataset is converted to linear radiance domain by applying randomly sampled camera parametric curve in the form of

$$F(x) = \frac{(1+\sigma)x^n}{(x^n+\sigma)}, \tag{3.1}$$

where $n \sim N(0.65, 0.1), \sigma \sim N(0.6, 0.1)$. Some of the scene of training dataset is represented in Fig. 3.1.

Figure 3.2: *The first two row represents some training samples frames from Froehlich et al. (2014) dataset. Specifically, Show girl 02 and FirePlace scene is presented and highlighted in orange. The last row represent sample frames from the Bridge scene in Kronander et al. (2014) dataset.*

3.2.2 Datasets for evaluation

We took two HDR video scenes from Froehlich et al. (2014) dataset Specifically, CAROUSER FIRWORKS and POKER FULLSHOT for testing and evaluating our

proposed method. To facilitate a more comprehensive evaluation on real data, Chen et al. (2021a) captured a real-world dataset with a reliable ground-truth HDR for evaluation. They capturing videos with alternating exposures (i.e., two and three exposures) in a variety of scenes, including indoor, outdoor, daytime, and nighttime scenes. The captured HDR videos have a frame rate of 26 fps and a resolution of 4096×2168. Three different types of video data are captured, namely, static scenes with ground truth augmented with random global motions, which is a random of pixel value in range [0,5] for each frame. Chen et al. (2021a) did not pre-align the input frames to investigate for finding models robustness against input with inaccurate global alignment. Dynamic scenes with ground truth dataset contains more challenging local motions specifically built for checking model robustness when counter regions with large motions. Dynamic scenes without GT contains uncontrolled dynamic scenes for qualitative evaluation. In future, this dataset can be utilized for semi-supervised or unsupervised learning-based techniques. Table A.2 appendix represents specification of the real world dataset with two and three alternating exposures.

3.2.3 Data preparation

To utilize the above mentioned datasets for training. Similar to the work of Kalantari et al. (2017); Kalantari and Ramamoorthi (2019); Chen et al. (2021a) similarity transforms including rotation, translation, and isometric scaling is applied to globally align the adjacent frames to the reference (center) frame in order to simplify the learning process of our proposed model. This is done by applying RANSAC for finding dominant similarity model from the estimated correspondences by finding corresponding corner features in reference and each neighbouring frame. Additionally, A gamma curve was applied on the input images instead of using of the original camera response function (CRF). Specifically, inverse CRF is first applied to translate all the frames into the linear HDR domain, i.e., $L_i = f^{-1}\frac{(Z_i)}{t_i}$, where f is the CRF and t_i is the exposure time of frame i. We then use a gamma curve with $\gamma = 2.2$ to transfer the images from HDR to LDR domain $lin_i(L_i)$:

$$F_i = lin_i(L_i) = clip[(L_i t_i)^{1/\gamma}] \tag{3.2}$$

where clip is a function that keeps the output in the range [0, 1] and lin_i is a function that transfers the image L_i from the linear HDR domain into LDR domain at exposure i.

Overall, the preprocessing step involves globally aligning $F_i - 1$ and $F_i + 1$ to the reference(center) image, F_i, and then replacing the original CRF with a gamma curve to produce neighbouring frames $F_i - 1$, $F_i + 1$, and a reference frame F_i. Even CRF replacement step is skipped, the system will then need to estimate the original

CRF in order to transform the images from the LDR to the HDR domain in the merging phase. Almost all prior techniques Mangiat and Gibson (2010); Kang et al. (2003); Kalantari et al. (2013); Li et al. (2016); Kalantari and Ramamoorthi (2019); Chen et al. (2021a) done this process as an initial requirement but this is not a big constraint because the CRF can be readily computed from a sequence of images with varied exposures using Debevec and Malik (2008).

Figure 3.3: *Representation of three consecutive frames with two alternate exposures of the carousel firework scene in Froehlich et al. (2014) HDR dataset. Each frame in three consecutive frame input contain some missing contents with the Presence of noise in frame $F_i - 1$ and $F_i + 1$ in the darker region due to acquisition with low exposure whereas F_i, which was taken with high exposure, lacks details in over-saturated and bright regions. The missing content of a final HDR image has to be reconstructed from neighboring frames with alternating exposures.*

The three frames were then used as an input to the model with the middle frame to be used as ground truth HDR frame as shown in Fig. 3.3. The datasets are generated with exposures separated by one, two, and three stops, where the low exposure time is randomly selected around a base exposure. Real world cameras often produce noisy images and are difficult to calibrate. It is necessary for the training dataset to represent those limitation of conventional camera systems in order for the learning-based model to perform and generalize effectively on scenes captured with conventional consumer cameras. Kalantari et al. (2017); Kalantari and Ramamoorthi (2019); Chen et al. (2021a) imitate the flaws of common consumer cameras by introducing noise and altering the tone of the synthetic images in their synthetic training dataset for ensuring the generalizability of their proposed network during the inference time. Image acquisition through conventional digital cameras usually contains noisy pixels in the dark regions. Then the information from those darker regions of the image should be taken from the high exposure image which has more details in that region. The input LDR synthetic training dataset usually have same amount of noise for both of the exposures. Directly using the dataset without modification, the content of the high exposure image in the dark regions will not be able to utilize, which eventually produce noisy results in real scenes

Kalantari and Ramamoorthi (2019). Similar to Kalantari and Ramamoorthi (2019); Chen et al. (2021a), Zero-mean Gaussian noise was added to the input LDR images with low exposure making the models to use the information in the dark regions of a clean high exposure image. The zero-mean Gaussian noise was specifically applied to the images in linear domain. The intention was to magnify the noise in the dark regions after transforming the image to the LDR domain. In order to account for noise variation similar to Kalantari and Ramamoorthi (2019); Chen et al. (2021a), random Gaussian noise range using standard deviation between $10 \times^{-3}$ and 3×10^{3} is used.

It is practically challenging to perform accurate camera calibration and determine the precise CRF. This even become more difficult after exposure adjustment, where the color and brightness of neighbouring frames is different than the reference frame. The synthetic training dataset does not represent the mentioned behaviour thus limiting model performance to generalize well on videos captured with conventional RGB cameras. Kalantari and Ramamoorthi (2019); Chen et al. (2021a) apply a gamma function with $\gamma = exp(d)$, where d is randomly selected from the range [-0.7, 0.7] and applied to different color channels of the reference image for a slight tone perturbation making the models to manage the inconsistencies of the reference and nearby frames while estimating the flows and the blending weights by using this perturbed reference image as its input. To create HDR images that reflect the ground truth, they employ the original reference image (prior to tone perturbation) combined with the nearby images throughout the blending process. The proposed network take cropped patches of size 256x256 as input along with random horizontal/vertical flipping and rotation.

3.3 Deep HDR video reconstruction

Given an input LDR video/sequential frames with alternating exposures the Multi-Attention SKFHDRNet reconstructs a high-quality HDR video. Kalantari et al. (2017); Kalantari and Ramamoorthi (2019); Chen et al. (2021a) followed a similar strategy of stacking input images both in Linear and LDR domain in the early stage of network for merging. The proposed Multi-Attention SKFHDRNet similar to Yan et al. (2019) architecture first retrieve attention guided features from the neighbouring frames in relation to a center reference frame using multiple attention mechanisms. The attention guided feature maps are then feed to the merge model in case out initial studies. Later attention guided feature are passed to PCD module for further improving the alignment of frames. After that, the aligned frames are then passed to the merge network for final HDR video reconstruction.

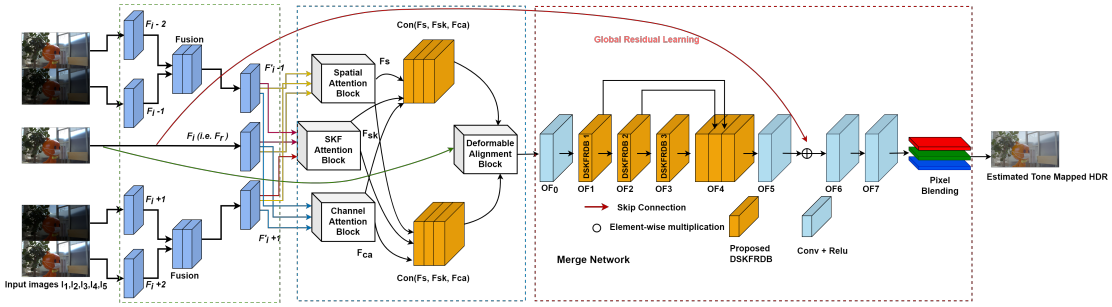


Figure 3.4: *The proposed Multi-Attention SKFHDRNet network consists of a multiple attention blocks along with PCD alignment module for extracting improved and relevant features and a merging network for estimating the final HDR image. The Multi-Attention blocks is used to give less importance to useless information caused by misalignment and saturation and highlighting the useful details using a three way attention. The PCD module further align the feature at multiple scales. The merging network perform final reconstruction based on a series of (DSKFRDBs) along with global residual skip connection on the aligned features. The final 3-channel HDR image is reconstructed by blending the pixel values of five frames (15 channel weights estimated by model).*

As shown in Fig. 3.4, the Multi-Attention SKFHDRNet consists of two major sub networks, i.e. the attention modules having (Spatial, Channel, Attention through adaptive kernel selection and fusion) for aligning and recovering missing content in the reference (center) frame. The Multi-Attention blocks give importance to only those features which are relevant to the center frame. This is done by first fusing the neighbouring frames features to the reference frame. The fused features are then passed to our Multi-Attention blocks for extracting missing content from neighbouring frames $F_i - 1$, $F_i - 2$ and $F_i + 1$, $F_i + 2$ in relation to the reference (center) frame F_i, For further refining the alignment of neighbouring frames to the reference frame. The aligned features are introduced to the PCD alignment module to further improve the temporal coherency and alignment. The refined feature are then introduced to the merge network which is composed series of DSKFRDBs for estimating high quality HDR video. The DSKFRDBs with dilation convolution help to recover the details affected by over-exposure and moving objects by enlarging the receptive field.

The proposed Multi-attention blocks and (PCD) module is discussed in great detail in Section (3.4.1, 3.4.2, 3.4.3) and Section 3.5 respectively. Overall, The attention module first applies convolution operation individually to each of the LDR images for extracting features. Then for getting relevant features from the non-reference image features multiple attention modules (Spatial, Channel-wise and attention guidance through adaptive receptive field consideration the scales of the information in the input) are applied. Attention guided feature maps are retrieved

from the multiple attention blocks. The spatial attention mechanism usually follows concatenation, convolution operation on extracted features of reference (center) and non reference (alternative) frames. While in case of channel attention, important information across channel of the feature are extracted. Spatial information is extracted through max and average pooling. These grouped features are then forwarded to a shared light-weight MLP network which assigns attention guided weights per channel according to how much relevant information is available in a particular channel of a feature map. In order to give importance to features considering the scale of the information in incoming feature, selective kernel fusion based attention through adaptive receptive field sizes is applied, giving importance to information considering it scales in the incoming input.

Since the main goal our proposed method is to reconstruct an HDR image which has to be consistent with the reference image, the intention of applying attention and PCD alignment module on the non reference (alternative) frames in relation to reference (center) frame prior to merge network is to get rid of ghosting artifacts by identifying misaligned information.The merging network then combines the LDR image features by hallucinating the details in the affected regions due to large over-saturation and misalignment from moving objects. By using a set of (DSKFRDBs) along with global residual learning strategy, the merging network become able to reconstruct high quality HDR video using the attention guided features from attention and PCD blocks as input. Multiple-Attentions and PCD modules make our learning-based merge network to focus more on the important information, rather than learning non-critical background information.

3.4 Multi-Attention Guided Image alignment

This section will discuss the attentions we introduce to our learning based method for reconstruction HDR video separately.

3.4.1 Spatial attention

The spatial attention guided module were given three or five 6-channels input LDR images F_i, where $i = 1, 2, 3$ or $i = 1, 2, 3, 4, 5$. In case of 5 frame input, First neighbouring input frames $F_i - 2$, $F_i - 1$ and $F_i + 1$, $F_i + 2$ were concatenated and fused. Convolution operation and LeakyReLU activation are then applied on the fused features (see Fig 3.5). Similarly, Convolution with LeakyReLU operation are applied on the individual 6 channel reference F_i to extract 64 channel features. The fused $F'_i - 2$, $F'_i - 1$ and $F'_i + 1$, $F'_i + 2$ are then concatenated with the reference (center) frame F_i and passed it to the attention blocks A_1 and A_3. In case of three

frame input the initial fusion part is omitted and a direct convolution along with LeakyRelU is applied similar to F_i as shown in Figure 3.5.

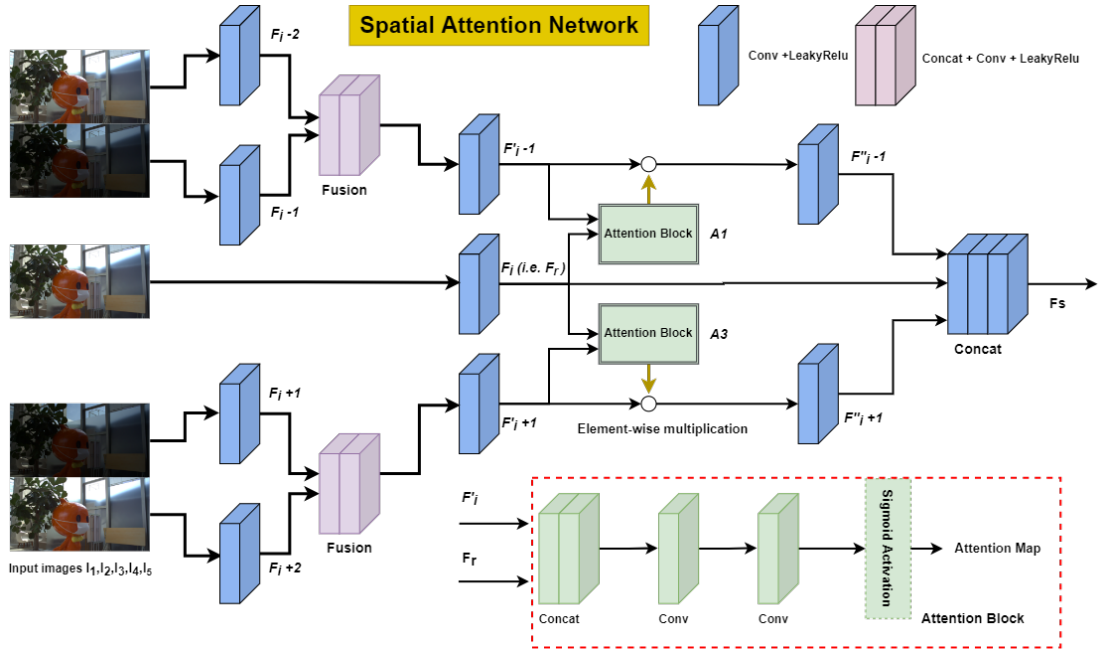


Figure 3.5: *The attention block initially concatenates the two inputs before obtaining attention maps through two Conv layers following sigmoid activation function to limit the output in [0,1].*

To obtain the attention maps for the non-reference images, we feed the fused features F_i' , $i = 1, 3$ of the non-reference images to the convolutional attention module $a_i(\cdot)$, $i = 1, 3$ along with the reference image feature map F_r, and then obtain the attention maps A_i , $i = 1, 3$ for the non-reference frames using Equation. 3.3.

$$A_i = a_i(F_i', F_r), (i = 1, 3). \tag{3.3}$$

A_i attention maps has the same size as F_i' where the values in A_i are in the range [0, 1]. Details of the attention modules are provided below. The predicted attention maps are used to attend the features of the non-reference images via Equation 3.4:

$$F_i'' = A_i \circ F_i', (i = 1, 3), \tag{3.4}$$

where $\circ$ denotes the point-wise multiplication between A_i and F_i' , $(i = 1, 3)$. The F_i'' denotes the feature maps with attention guidance. The reference feature map F_r (i.e. F_i) and the attention guided features of the non-reference images $F_i'' - 1$

and $F_i'' + 1$ are stacked and fused to get final 64 channel attention guided feature map F_s with the guidance F_i see(Equation 3.5).

$$F_s = Concat(F_i'' - 1, F_r, F_i'' + 1), \tag{3.5}$$

where Concat$(\cdot)$ denotes the concatenation operation. F_s will be used as the input along with inputs from other attention modules to the (PCD) module for further refining the alignment process. Since the HDR imaging process centers on the reference image, the attention maps are predicted and applied according to the reference frame.

3.4.2 Channel attention

By taking advantage and exploiting the dependencies between features across channels, we employ channel attention proposed by Woo et al. (2018). The architecture of the channel attention network is shown in Fig. 3.6.

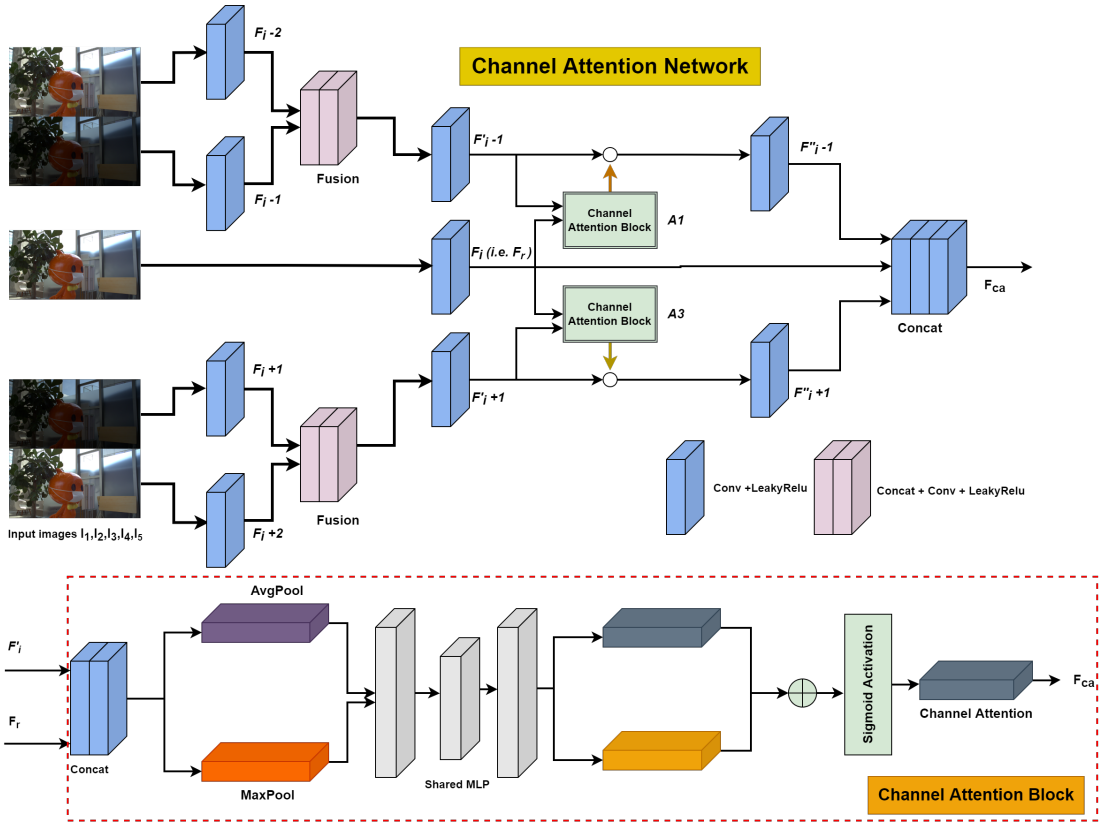


Figure 3.6: *Represents attention sub-module. As shown, the channel sub-module use both max and average pooling with a shared MLP network.*

Similar to spatial attention block, The fused $F_i' - 1$ and $F_i' + 1$ are concatenated with the reference (center) frame F_i and passed it to the channel attention blocks. Channel attention blocks first apply both average and max-pooling operations for grouping spatial information from the feature maps resulting in averaged-pooled and max-pooled features represented as F_{avg} and F_{max} respectively (see Equation 3.6). After that, a shared network which is a Multi-layer Perceptron (MLP) with one hidden layer receives both descriptors to produce attention guided weights per channel $W \in R^{C\times1\times1}$. The MLP hidden activation layer size was controlled and set to $R^{C/r\times1\times1}$, with r (reduction ratio) used for reducing the parameters of the hidden activation layer. Finally the output features vectors of the shared network (MLP) on the F_{avg} and F_{max} features is merged using element-wise summation. The overall channel attention computation follow as in Equation 3.6:

$$\begin{aligned} A_i &= \sigma(MLP(F_{avg}(F(_i)(_r))) + MLP(F_{max}(F(_i)(_r)))) \\ &= \sigma(W_1(W_0((F_i, F_r)F_{avg})) + W_1(W_0((F_i, F_r)F_{max}))), \end{aligned} \tag{3.6}$$

where σ denotes the sigmoid function, $W_0 \in R^{C/r\times C}$, and $W_1 \in R^{C\times C/r}$ represent MLP layer weights. Note that the MLP weights, W_0 and W_1, are shared for both inputs and the ReLU activation function is followed by W_0. The estimated attention maps are point-wise multiplied to attend the features of the non-reference images via Equation 3.7:

$$F_i'' = A_i \circ F_i', (i = 1, 3), \tag{3.7}$$

where $\circ$ denotes the point-wise multiplication between A_i and F_i' , $(i = 1, 3)$. The F_i'' denotes the feature maps with attention guidance. The channel attention modules $A_i, (i = 1, 3)$ pick the stack of features with the guidance of the reference as F_i. Lastly, Attention guided features $F_i'' - 1$ and $F_i'' + 1$ are concatenated with the reference frame F_r to get final stack of channel attention guided features F_{ca} see(Equation 3.8).

$$F_{ca} = Concat(F_i'' - 1, F_r, F_i'' + 1), \tag{3.8}$$

where Concat$(\cdot)$ denotes the concatenation operation. which concatenate the reference frame features with the attention guided features of the non-reference images. The channel attention guided features F_{ca} are then passed to the PCD alignment module for further refinement.

3.4.3 Soft attention using selective kernel fusion

To enable neurons to adaptively adjust their RF sizes according to the information in the input, we apply the work of Li et al. (2019) which select multiple kernels

having different receptive field sizes through automatic selection process. Overall, selective kernel fusion involve three main operations specifically, split, fuse and select as illustrated in Fig. 3.7.

Split: The split operation first perform two transformation on the incoming fused feature maps F_i', F_r of size $H' \times W' \times C'$. Incoming Features F_i', F_r are transformed to $U_3 \in R^{H \times W \times C}$ and $U_5 \in R^{H \times W \times C}$ features based on the receptive field sizes. We specifically choose receptive field of sizes 3 and 5 in our proposed architecture. The split operation extract features from U_3 and U_5 by applying efficient depthwise convolutions, followed by ReLU activation function. Additionally, dilated convolution of 3x3 receptive field with dilation size of 2 was used instead of conventional 5x5 convolution operation.

Fuse: Adaptively adjusting the receptive field sizes of neurons considering the scale of the content is done through the fuse module. Fuse module controls the information flow of different scales from the two branches having different receptive fields into the activation functions in the next layer. The information from the two branches are integrated via element wise-summation illustrated in Equation 3.9.

$$U' = U_3 + U_5, \tag{3.9}$$

The global average pooling is then performed with the aim of adding global information to generate channel wise statistics as $S \in R^C$.

$$S = F_{gp}(U') = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U'(i,j), \tag{3.10}$$

Furthermore, the global averaged pooled feature vector are then passed to a simple fully connected (fc) layer for more precise adaptive selection of features $Z \in R^{d \times 1}$ with addition of dimentionality reduction parameter for better efficiency.

$$Z = F_{fc}(S) = \delta((WS)), \tag{3.11}$$

where δ is the ReLU function and $W \in R^d \times^C$ represent fully connected layer (fc) parameters. where d represent reduction ratio controlled by parameter r for improving the efficiency of the model by controlling the parameters of the fully connected layer.

$$d = max(C/r, L), \tag{3.12}$$

We use $L = 32$ which represent the minimal value of variable d. **Select:** The final operation of the selective kernel fusion block is to adaptively select different scales of informative content from a guided feature descriptor Z by applying a channel-wise softmax operator using Equation 3.13.

$$a = softmax(Z), b = softmax(Z) \tag{3.13}$$

The softmax based attention guided feature maps are multiplied with U_3 and U_5 and then summed to get final attention guided feature map using an Equation. 3.14.

$$A_i = a \cdot U_3 + b \cdot U_5, \tag{3.14}$$

Where A_i is a soft attention guided features which are then point wise multiplied with non reference features F_i' using equation 3.15.

$$F_i'' = A_i \circ F_i', (i = 1, 3), \tag{3.15}$$

where $\circ$ denotes the point-wise multiplication and F_i'' denotes the feature maps with attention guidance. The selective kernel fusion attention maps $A_i, (i = 1, 3)$ with the guidance of the reference as F_i point-wise multiplied to get attention guided features for non reference frames. Lastly, Attention guided features $F_i'' - 1$ and $F_i'' + 1$ are concatenated with the reference frame F_r to get final stack of selective kernel fusion based soft attention guided features F_{sk} by using equation 3.16 which is then feed to the PCD alignment module.

$$F_{sk} = Concat(F_i'' - 1, F_r, F_i'' + 1), \tag{3.16}$$

The overall structure of the selective kernel fusion based soft attention is represented in Fig. 3.7.

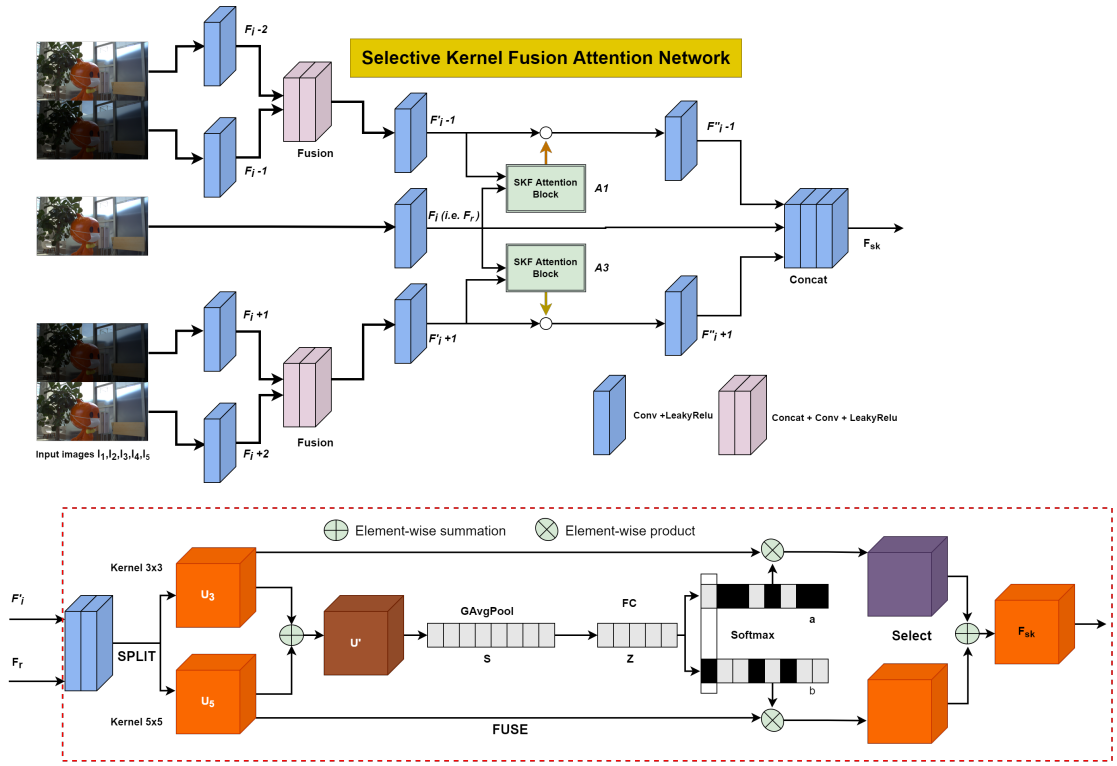


Figure 3.7: *Represents selective kernel fusion attention block which consist of three stages where convolution with 3x3 and 5x5 kernel are applied first. Fuse stage maps features using two receptive field and then summed those features. For more precise adaptive selection of features global average pooling and fully connected layer are utilized. Finally in select stage, adaptive selection of different scales of informative content from a guided feature descriptor using softmax operator.*

3.5 Refined deformable feature alignment

Recently, Deformable feature alignment for the problem of video super-resolution by using Deformable convolution which was proposed by Dai et al. (2017) has been successfully applied by Wang et al. (2019) and Tian et al. (2020)). The core concept behind deformable alignment is that, An offset is predicted using an offset prediction module using equation 3.17, which uses general convolutional layers, given two features as input (for example in our case, fused features F_s, F_{ca} and F_{sk} and a reference frame feature map F_i).

$$\Delta p_i - 1 = func([fused(F_s, F_{ca}, F_{sk})), F_i]) \tag{3.17}$$

With the learned offset, the fused multi-attention guided features F_s, F_{ca} and F_{sk} can be sampled and aligned to the reference frame F_i using Dai et al. (2017) deformable convolution using equation 3.18:

$$\tilde{F}_i = DFConv(fused(F_s, F_{ca}, F_{sk}), \Delta p_i - 1). \tag{3.18}$$

The overall structure of PCD alignment module is represented in Fig. 3.8 where the alignment is performed at multiple scales between fused refined features and reference frame.

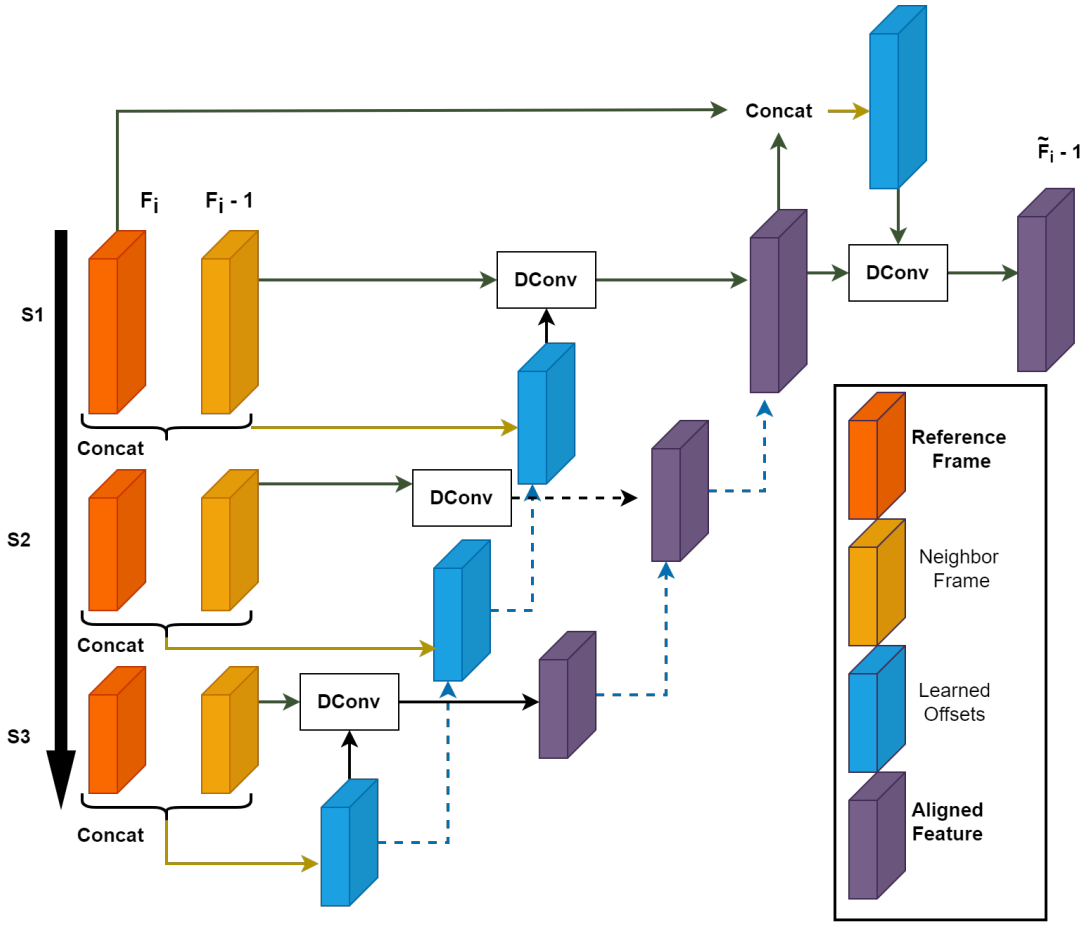


Figure 3.8: *Structure of the deformable alignment module.*

We adopted the pyramid, cascading and deformable (PCD) alignment module proposed by Wang et al. (2019), as our feature alignment module. The final HDR video reconstruction is optimized by implicit learning capabilities of deformable convolution offsets for this alignment process.

3.6 Image alignment using optical flow

We also adopted the optical flow methodology adopted by Kalantari and Ramamoorthi (2019) and Chen et al. (2021a) for efficient frame alignment. It is necessary to align the frames in the initial phase of learning-based techniques with the reference frame K_i. This is done by estimating the flows of neighbouring frames K_{i-1} and K_{i+1}, to the reference frame, K_i. warping is then performed on nearby images K_{i-1} and K_{i+1}, using the estimated flows for setting a series of aligned images Ki_{-1}, i and K_{i+1}, i in relation to the refernece frame K_i for efficient treatment of non-rigid motion and the inaccuracies introduced by global alignment. Using the work of Kalantari and Ramamoorthi (2019); Chen et al. (2021a), which estimates the flow using CNN based method for more accurate image alignment of neighbouring frame to the reference frame is represented in Fig. 3.9.

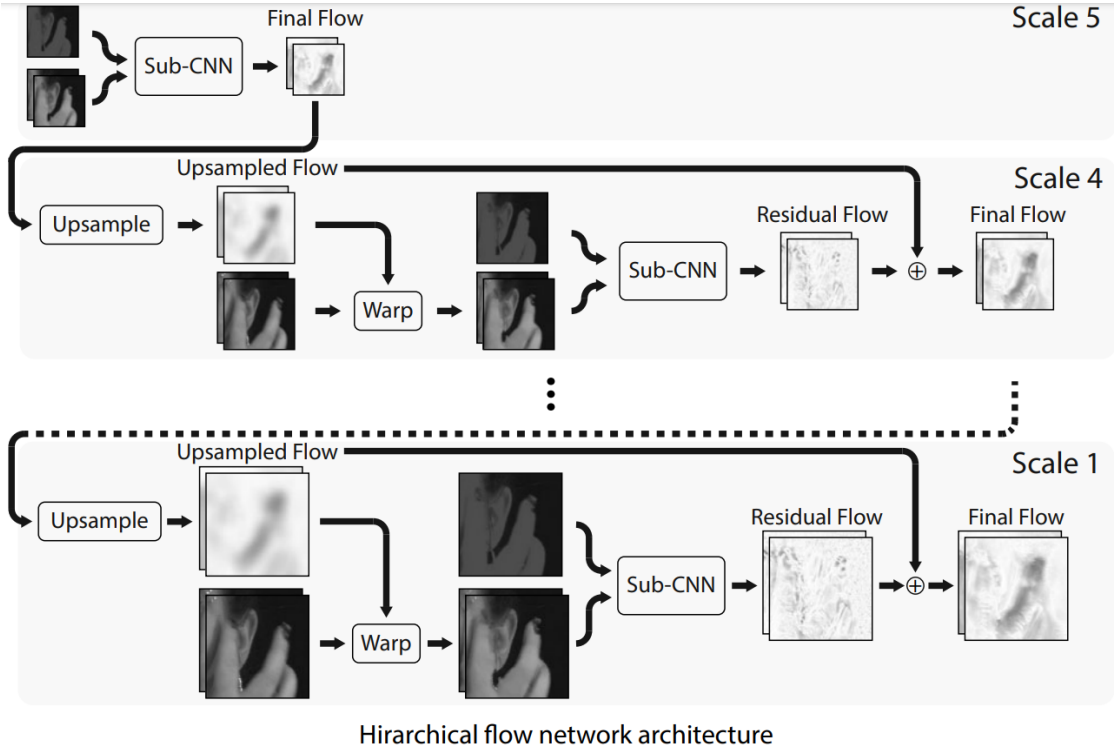


Figure 3.9: *Represents the hierarchical multi-scale Flow network used by Kalantari and Ramamoorthi (2019) and Chen et al. (2021a). Figure reproduced from Kalantari and Ramamoorthi (2019).*

Additionally, in comparison to optimizing optical flows based on non-learning techniques, CNNs-based methods provide a number of benefits. As an example,

computing is done on Graphical Processing Units (GPU), which executes faster and operate more effectively than traditional approaches. Second, the proposed CNNs-based approach for flow estimation is trained in a end-to-end fashion for high quality HDR video reconstruction and making the overall architecture of the HDR video reconstruction more compact.

For predicting optical flow, a number of Deep Learning-based methods have been put forth by Ilg et al. (2017); Ranjan and Black (2017) and Teed and Deng (2020). The two input images used in the aforementioned optical flow estimation methods are utilized to estimate the flow between them. However, utilizing two images to estimate flow is not a good solution for HDR video reconstruction approach. It is challenging to estimate the precise flow between the reference frame and neighboring frames in HDR since the reference image typically has missing content.

To overcome this problem, Kalantari and Ramamoorthi (2019); Chen et al. (2021a) used two neighbouring frames K_{i-1} and K_{i+1} along with the reference frame K_i as an input to the flow network. By using three consecutive input with the center as a reference more appropriate flows is estimated by using the neighbouring images where reference image has regions with missing information. The issue of frames having different exposures arises as the input frames are captured with alternative exposures creating differences in brightness between frames. Two neighboring frames K_{i-1} and K_{i+1} together with the reference (center) frame K_i were employed as inputs to the flow network by Kalantari and Ramamoorthi (2019) and Chen et al. (2021a). More suitable flows are estimated by utilizing the neighboring frames when the reference image contains regions with missing data by using three successive inputs with the center as a reference. As the input frames are taken with different exposures, the problem of frames with varied exposures occurs which ultimately change the the brightness of the frames. This limitation is resolved by Kalantari and Ramamoorthi (2019); Chen et al. (2021a) by adjusting the reference frame exposure to that of neighbouring frames $g_{i+1}(K_i)$ using the equation 3.19:

$$g_{i+1}(K_i) = l_{i+1}(E(K_i)), \tag{3.19}$$

where $E(K_i)$ is a function that takes the image K_i from the LDR domain to the linear HDR domain by using equation:

$$E(K_i) = K^{\gamma i}/t_i \tag{3.20}$$

The input is then obtained by concatenating the exposure adjusted reference image as well as the two neighboring frames, i.e., $\{g_{i+}1(K_i), K_{i-1}, K_{i+1}\}$. The network generates an output with four channels that consists of two sets of flows in the x and y directions from the reference frame, K_i to the frames immediately before K_{i-1} and after , $K_i + 1$. The adjacent frames are then warped using these flows to create a set of aligned images. The proposed flow network by Kalantari and Ramamoorthi

(2019); Chen et al. (2021a), have a hierarchical coarse-to-fine architecture, with three input images, as shown in Fig. 3.9. Utilizing factors of 16, 8, 4, and 2, three input images are downsampled to create a multi level pyramid. Each pyramid level estimates two sets of flows. The whole flow is created by adding the upsampled flow from the previous scale with the estimated flows at current scale. This method is continued until the finest scale is obtained in order to generate the final flows. Using the estimated flows, the nearby frames are then warped to produce the aligned images K_{i+1} and K_{i-1}. These aligned images are then put forwarded to the attention blocks for further improved alignment in features domain.

3.7 Merge network for HDR image reconstruction

The merge network main objective is to reconstruct a high-quality HDR frame from the attention guided aligned features. This network should basically locate and remove the remaining alignment artifacts exist in the registered images and recover missing content in the over-exposed regions in the final HDR image. Considering the challenging problem of HDR image reconstruction, we apply selective kernel fusion blocks as a residual dense network similar to Zhang et al. (2018). Our merge network is composed of convolution layers, (DSKFRDBs) with addition of skip of connections (see Fig. 3.10).

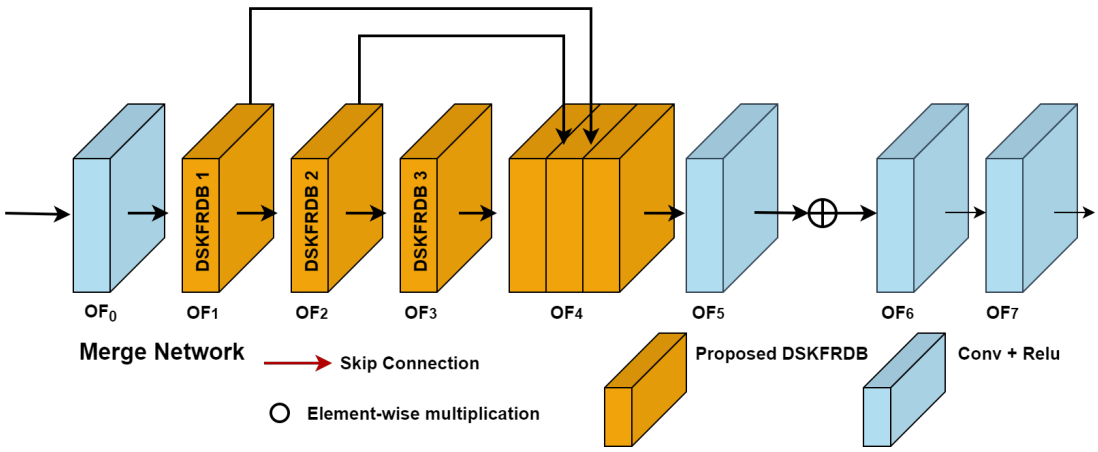


Figure 3.10: *Represent merge network composed of series of dilated selective kernel fusion residual dense blocks with skip connections.*

The merging network takes the stacked features from the PCD alignment module and passed to the merge network. The merge network first apply a Conv layer to

produce a 64-channel feature maps. These feature maps are then passed to three DSKFRDBs having two branches representing the usage of two receptive fields with dilation $M = 2$ outputting three corresponding feature maps OF_1, OF_2 and OF_3. All the three feature maps are then concatenated to get OF_4. Then convolution operation are applied for extracting more relevant information from all the three merged feature maps produced from DSKFRDBs to get OF_5.

Global residual learning with the reference features:
Motivated from the work of Ledig et al. (2017); Yan et al. (2019), global residual learning strategy was adopted by adding the shallow reference frame feature F_r to OF_5 where the representation from the original reference information is integrated before reconstructing the final HDR image from OF_5 for optimizing the model accuracy.

$$F_6 = F_5 + F_r, \tag{3.21}$$

The final feature map $OF6$ contains almost all the ingredients for reconstructing the final HDR image without ghosting artifacts with details recovered in over and under-exposed regions with large motion. The final HDR image is estimated in the HDR domain after two convolution layers followed by activation funtion.

Dilated selective kernel fusion residual dense block:
The merging network requires a larger receptive field for hallucinating details since the reconstruction of some local regions of the HDR images cannot receive enough information from the LDR images due to the occlusion of moving objects and saturation. Thus, we used dilated residual selective kernel fusion block with two branches with dilation. The proposed (DSKFRDBs), which is represented in Fig. 3.11, is perform final HDR video reconstruction by adaptive feature selection using two different receptive fields using the Split, Fuse, and Select strategy with dense concatenation based skip-connections where the input for each layers is the concatenation of all the feature maps from preceding layers.

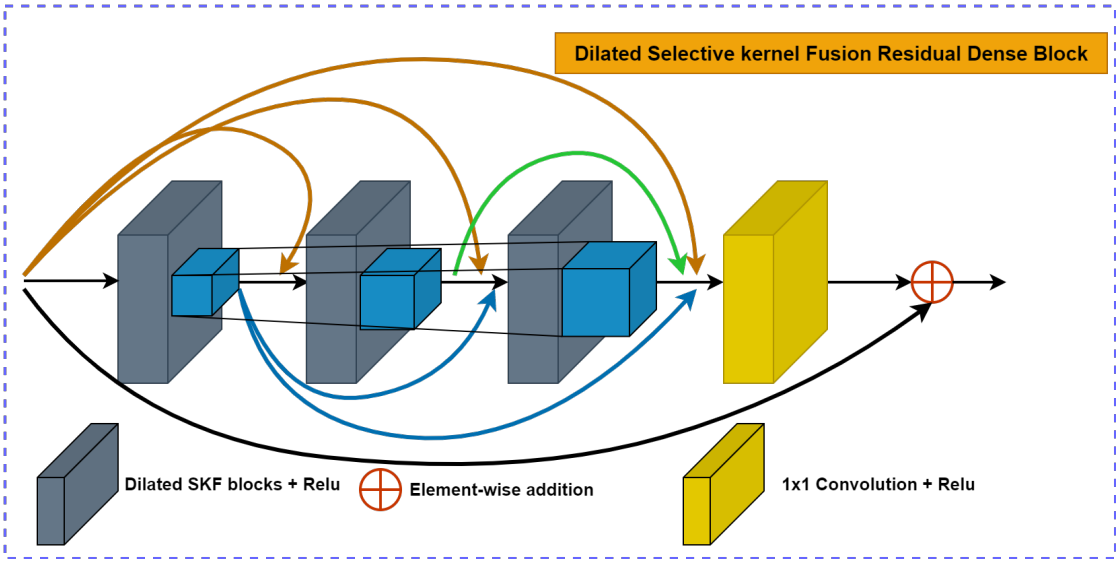


Figure 3.11: *Illustration of a three-layer convolution dilated selective kernel fusion residual dense block structure following the residual dense block strategy of Zhang et al. (2018) as a framework.*

3.8 Pixel blending

To our full multi-Attention SKFHDRNet we provided five 6-channel input images in both LDR and linear domains making a 30 channel input. Then, for these five images, our network predicts the blending weights and produced a 15 channels output. To effectively utilize the information in each color channel, we estimate a blending weights for each color channel in a manner similar to the methods already proposed by Debevec and Malik (2008); Kalantari et al. (2013). The five input images are averaged using their blending weights to get the final HDR image HDR_i at frame i by using the following Equation.

$$HDR_i = \frac{w_1 I_i + w_2 I_i - 1 + w_3 I_i + 1 + w_4 I_i - 1 + w_5 I_i + 1}{\sum_{K=1}^{5} wk} \tag{3.22}$$

Here, wk is the estimated blending weight for each image.

3.9 Training

Our learning-based has two primary phases, training and testing, which are typical to most machine learning techniques. The optimal network parameter weights are determined using an optimization approach by training the networks, which is an

offline operation. To perform this task, we need a good amount of training data set and suitable metric to compare estimated scene with the ground truth HDR. Once the model is trained, we can evaluate and apply our trained network to new test scenes.

3.9.1 Loss function

Updating model parameter by directly applying loss function on the images in the linear HDR domain will produce inaccuracies by underestimating the error in pixel values of the dark regions. For boosting the pixel values in the dark regions of the image, Tone mapping operation is usually applied. Following the work of Eilertsen et al. (2017); Kalantari and Ramamoorthi (2019); Chen et al. (2021a) the linear HDR images are transformed in to log domain in order to overcome the above mentioned problem. To transform the HDR images into the log domain, we specifically employ the differentiable μ-law function using Equation 3.23:

$$T_i = \frac{log(1 + \mu HDR_i)}{log(1 + \mu)}, \tag{3.23}$$

Where HDR_i represent linear HDR frame with the pixel values in range of [0, 1]. The parameter μ is set to 5000 for controlling the rate of compression range. The model parameter are updated by minimizing the L_1 distance between the estimated, $\hat{T}_i$, and ground truth,T_i, HDR frames in the log domain:

$$E = \|\hat{T}_i - T_i\|_1, \tag{3.24}$$

Our proposed model is trained end to end by directly minimizing L_1 or L_1MS−SSIM loss. The gradients using the chain rule are computed. The parameters or weights of the networks are modified using these computed gradients continuously until convergence.

3.9.2 L_1MS−SSIM loss function

Zhao et al. (2016) investigate the use of three alternative error metrics (L_1, SSIM, and MS−SSIM), and come up with a new metric that combines the advantages of L_1 and MS−SSIM. Their findings states that MS−SSIM preserves the contrast in high-frequency regions better than the other loss functions. On the other hand, L_1 preserves colors and luminance and error is weighted equally regardless of the local structure—but does not produce quite the same contrast as MS−SSIM. To capture the best characteristics of both error functions, Zhao et al. (2016) propose a combined L_1MS−SSIM loss function which is represeted by Equation 3.25:

$$L_{mix} = \alpha L_{MS-SSIM} + (1 - \alpha) G^M_\sigma{}_G . L_1, \tag{3.25}$$

where α is empirically set to 0.84 with a point-wise multiplication between ${G_\sigma^M}_G$ and L_1. ${G_\sigma^M}_G$ represents the computation of mean and standard deviations with a Gaussian filter. We adopted the work of Zhao et al. (2016) for optimizing the training of our model.

3.9.3 Implementation details

PyTorch framework is being used to implement the Multi-Attention SKFHDRNet model architecture. We integrated the flow network implemented by Chen et al. (2021a) using Pytorch into our pipeline for HDR video reconstruction. End-to-end training is done for both the flow and our Multi-Attention SKFHDRNet. The technique used by Glorot and Bengio (2010) is used to initialize the initial weights of the network parameters. Using ADAM with the default settings of $\beta 1 = 0.9$ and $\beta 2 = 0.999$ with a learning rate of 0.0001, we solve the optimization problem. Mantiuk et al. (2008) approach was used for Tone-mapping the results. Given training images, we randomly crop the images of size 256×256 for training. The model was trained for 20 epochs on two NVIDIA Tesla V100 32Gb of NTNU cluster Själander et al. (2019).

3.9.4 HDR video reconstruction quality assessment

HDR−VQM is a full-reference HDR video quality metric proposed by Narwaria et al. (2015). We adopted HDR−VQM for evaluation of learning based models objectively. Overall, HDR−VQM is based on a series of steps specifically, signal pre-processing, transformation, frequency based decomposition and subsequent spatio-temporal pooling. Developed by Narwaria et al. (2015), HDR−VQM is a full-reference HDR video quality metric. Which was utilized for the objective assessment of learning-based models. Overall, HDR-VQM is built on a number of processes, including signal pre-processing, transformation, frequency-based decomposition, and subsequent spatio-temporal pooling Narwaria et al. (2015). The goal of HDR−VQM is to anticipate human perception of supra-threshold video aberrations, which are then aggregated into a single value, a measure of overall video quality. we also evaluated our proposed model on HDR−VDP2 quality metric proposed by Mantiuk et al. (2011). Which is able to identify variations in achromatic images with a wide range of absolute luminance values. The metric is also an effort to build a comprehensive model of contrast visibility for a wide range of illumination conditions. Additionally, We also compute the PSNR values for images after Tone-mapping using μ-law (μ PSNR) using equation 3.23.

Color difference formulas are widely used to evaluate the perceived color difference between two samples. The CIE published the CIEDE2000 formula in 2001

CIE (2001). The formula, which was created by CIE Technical Committee 1-47 members, offers a better way of calculating color differences between two samples. Luo et al. (2001) explained the process used to create the formula using experimental color-difference data. We compute color differences between estimated and ground truth HDR using CIEDE2000. For evaluation the estimated and ground truth HDR samples are converted CIELAB space. Lab values of both estimated and ground truth HDR image is then passed to the color difference formula for predicting the color differences betwee the two samples.

4 Results

4.1 Experiments overview

In this section, we conduct experiments and performed evaluation on synthetic Test HDR scenes and real-world datasets (dynamic and static scenes of Chen et al. (2021a)) to verify the effectiveness of the proposed method.

We perform our initial comparisons with Yan et al. (2019) in the case of no optical flow and no pixel blending where the model estimated a 3-channel final HDR image. In the next part of our experiment, we added the optical flow network where two flows were first estimated which were then utilized for warping the neighbouring frames for alignment process. The aligned frames are then introduced to the Multi-Attention SKFHDRNet which estimates a three-channel HDR image without using pixel blending and compared the model performance with Yan et al. (2019) AHDRNet. In the next phase, We trained our Multi-Attention SKFHDRNet with optical flow and use a pixel blending strategy. By pixel blending we mean our network estimates the blending weights for five images having 15 channels output. The estimation of blending weights for each color channel is done similar to Debevec and Malik (2008); Kalantari et al. (2013) techniques for proper usage of information in each channel. The final HDR image estimation is done through weighted averaging of five input images using their blending weights estimated by the models.

Our full model is composed of multiple modules specifically, optical flow network of Chen et al. (2021a), Multi-Attention modules, Pyramid cascading deformable alignment (PCD) module, with pixel blending strategy, we compared our full model performance with Kalantari et al. (2017), Kalantari and Ramamoorthi (2019), Yan et al. (2019) and Chen et al. (2021a). Additionally, our full model is trained with L_1 and a combined L_1MS$-$SSIM loss function of Zhao et al. (2016). The idea behind using MS$-$SSIM with L_1 loss function is that, apart from estimating absolute errors in case of L_1 loss, it is important to consider the pixel inter-dependencies spatially. Motivation behind using a combined cost function that consider structural dependencies along with absolute error will further refine the training process of

model. Note that we re-implemented Yan et al. (2019) method for alternating-exposure HDR video reconstruction, and trained them by changing the network input using the same synthetic training dataset as our method. We used the already trained Chen et al. (2021a) network parameters for comparison. In case of Kalantari et al. (2017) and Kalantari and Ramamoorthi (2019) work we took the model results from Chen et al. (2021a) paper since we are also using the same datasets for comparisons. All the models are visually compared and the predicted HDR image is evaluated in terms of multiple image quality metrics. We specifically used μ-law Tone-mapped PSNR, HDR−VDP2 Mangiat and Gibson (2011) and HDR-VQM Narwaria et al. (2015). We followed the HDR-VQM designed of Chen et al. (2021a) for evaluating the quality of HDR videos. Additionally, all the model were evaluated based on color difference error between estimated and ground truth HDR using CIEDE2000 Luo et al. (2001). All visual results in the experiment are Tone-mapped using Mantiuk et al. (2008) method.

4.2 Evaluation of baseline models with no optical flow and no pixel blending

This section will discuss our baseline Multi-Attention SKFHDRNet results on three datasets (Synthetic, Dynamic and Static) respectively. We did not include optical flow and pixel blending the motivation was to effectively evaluate the performance of our Multi-Attention modules and selective kernel fusion blocks in the merge network. The following subsections will present the model performance quantitatively and qualitatively.

4.2.1 Evaluation of baseline models on synthetic dataset

Our proposed Multi-Attention SKFHDRNet with re-implemented Yan et al. (2019) AHDRNet is evaluated on a synthetic test dataset which is composed of two HDR videos (i.e., POKER FULLSHOT and CAROUSEL FIREWORKS) of Froehlich et al. (2014) HDR dataset, which are not used for training. Each video has a resolution of 1920x1080 and contains 61 frames with Random Gaussian noise added to the low-exposure images Chen et al. (2021a). The zoomed regions of CAROUSEL FIREWORKS frame represented in Fig. 4.1 shows the poor performance of Yan et al. (2019) AHDRNet. Which somehow struggle in reducing ghosting artifacts due to large motion which ultimately introduce higher color difference error which can be seen in color difference maps of the images. However, our proposed Multi-

Attention SKFHDRNet model performed more better alignment in case of large motions and produced a smaller color difference error in relation to the ground truth HDR frame.

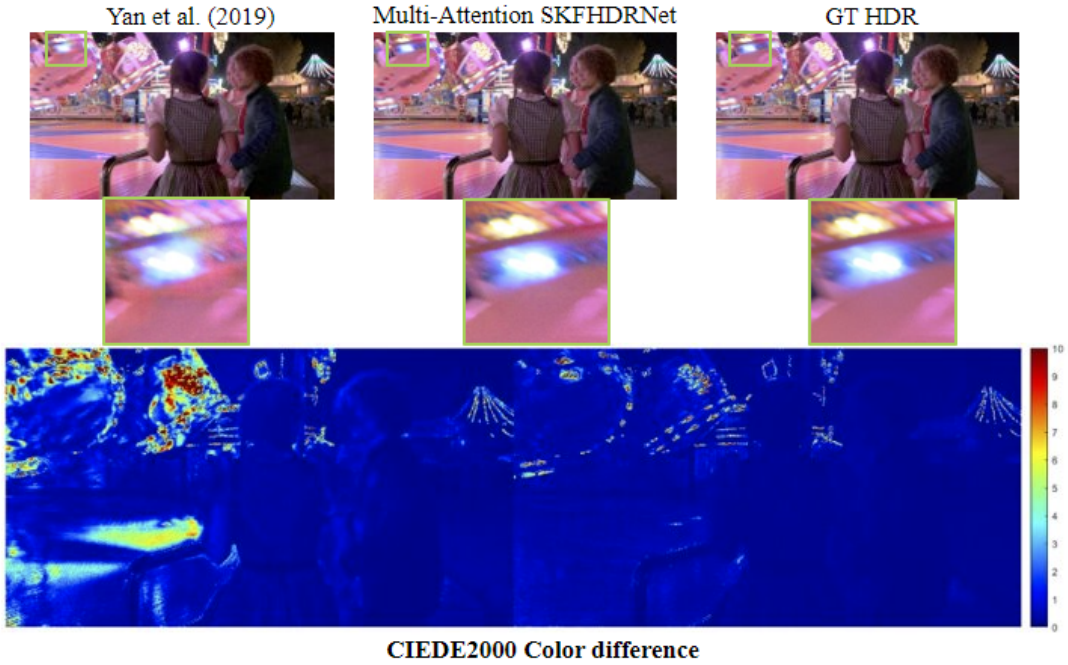


Figure 4.1: *Visual and color difference error results of baseline model on synthetic dataset (CAROUSEL FIREWORKS) scene.*

The quantitative results in terms of μPSNR and HDR−VDP2 are represented in Table 4.1. Our Multi-Attention SKFHDRNet showed better performance on all three image and video quality metrics. This indicates our multi-attention modules efficiency which guides more relevant features from the neighbouring frames in relation to the reference frame.

Table 4.1: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet on synthetic datset is represented. Bold text indicates the best among models.*

Model performance on synthetic dataset		
Models	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	28.78	63.56
Multi-Attention SKFHDRNet	**32.11**	**65.65**

4.2.2 Evaluation of baseline models on static dataset

The proposed model along with the re-implemented Yan et al. (2019) model is evaluated on a static dataset which is augmented with random global motion. Specifically, for each frame random translation was performed in pixel range of [0,5]. Input frames are not pre-aligned in order to evaluate the robustness of the models against input with inaccurate global alignment Chen et al. (2021a). From the visual results, The Yan et al. (2019) model struggles to recover details in over-exposed regions which are illustrated in the zoomed regions of static dataset scene in Fig. 4.2. Multi-Attention SKFHDRNet recover much of the missing information in the over-exposed regions with a small color difference error as shown in Fig. 4.2.

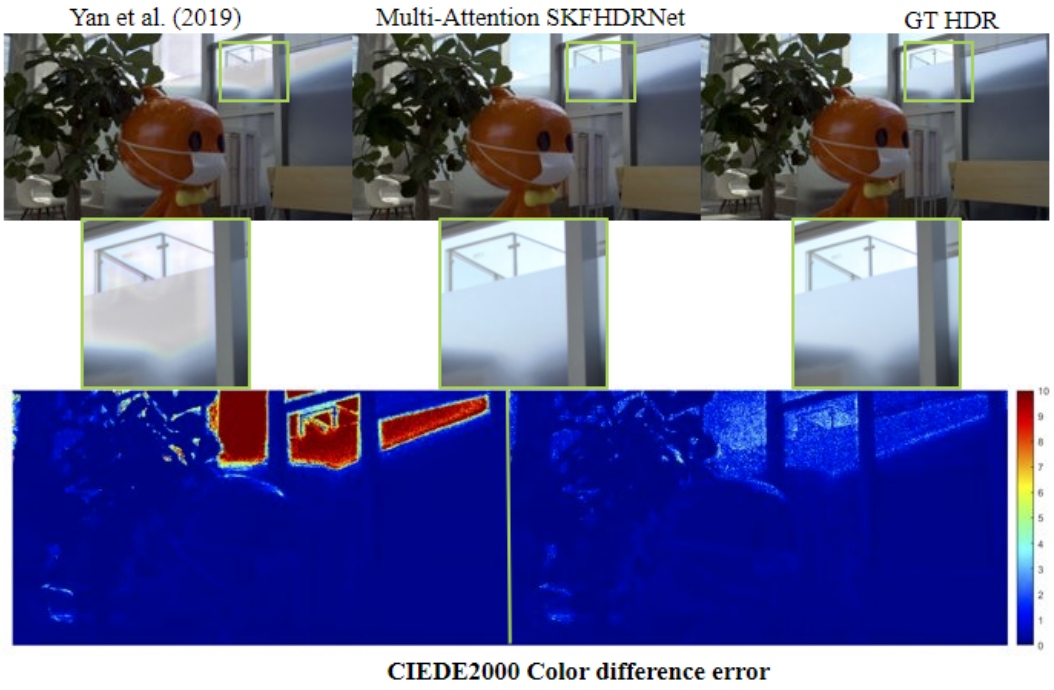


Figure 4.2: *Represents visual and color difference error results on the static dataset scene.*

The Multi-Attention SKFHDRNet performance on static dataset using objective quality metrics, μPSNR and HDR-VDP-2 was higher (see Table 4.2) as compared to Yan et al. (2019) AHDRNet. This proves our model robustness to random translation of pixel values as well as recovering more details in the over-exposed regions.

Table 4.2: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet on static dataset is represented. Bold text indicates the best among models.*

Model performance on static dataset		
Models	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	33.06	69.81
Multi-Attention SKFHDRNet	**36.76**	**71.34**

4.2.3 Evaluation of baseline models on dynamic dataset

Yan et al. (2019) AHDRNet also performed inferior on dynamic dataset which contains frames with more challenging local motions. Yan et al. (2019) AHDRNet single attention based alignment and fusion produced unpleasant artifacts in regions with significant motion. The artifacts can be seen on a person's face see (Fig 4.3) in the estimated HDR frame of Yan et al. (2019) AHDRNet.

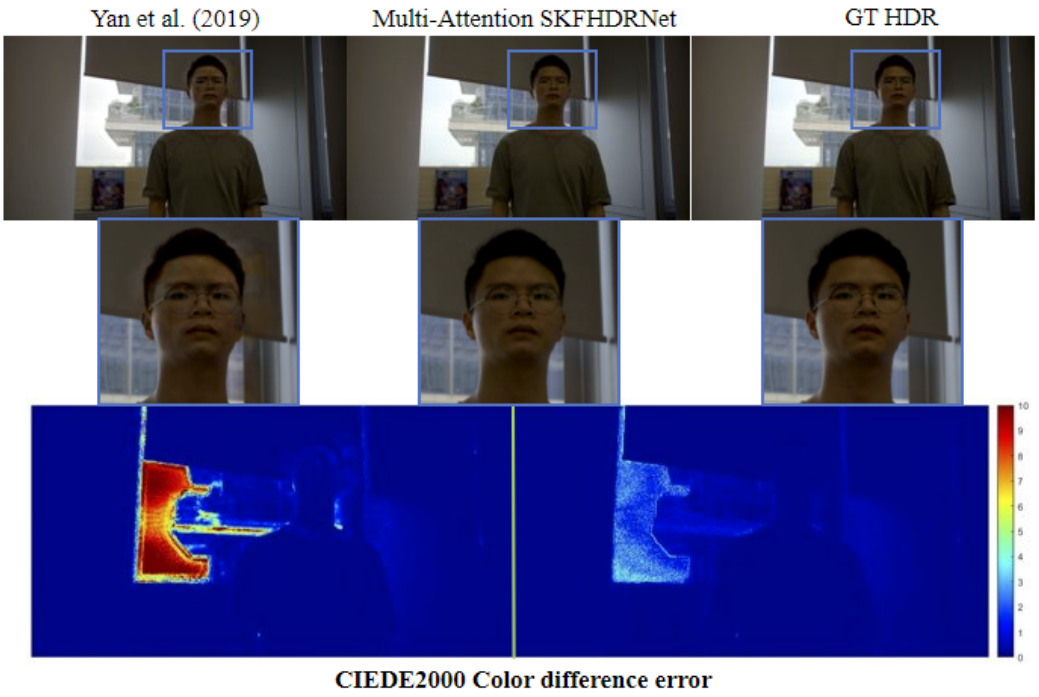


Figure 4.3: *Visualizes dynamic scene test sample and its estimated HDR images along with color difference map.*

Our Multi-Attention guided network was robust to regions with large motion

and produces ghost-free estimated HDR images of scenes with two exposures by performing more accurate alignment and fusion. The improved Multi-Attention SKFHDRNet performance on dynamic dataset using objective quality metrics, μPSNR and HDR-VDP-2 can be seen in Table 4.3. Additionally, from Fig. 4.3. we can see that Yan et al. (2019) AHDRNet does not recover information from the overexposed regions. The CIEDE-2000 indicates large color difference error in the background sky region of the dynamic dataset scene. However, our model recover details in the over-exposed region which can easily be seen in color difference map. This again indicates our model's dilated selective kernel fusion block effectiveness in filling the missing content in large over-exposed regions.

Table 4.3: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet on dynamic dataset are represented. Bold text indicates the superior among models.*

Model performance on dynamic dataset		
Models	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	34.68	68.42
Multi-Attention SKFHDRNet	**40.77**	**73.81**

4.2.4 Per frame objective metric results visualization of our baseline model with out optical flow and pixel blending

We adopted violin plot for per frame objective metric results visualization. Violin plots perform visualization similar to histograms by representing the shape of the data using probability density function. Probability density function's width used in a violin plot depicts the frequent occurrence of data points in dataset. Conversely, probability density functions narrow regions depicts data points that are occurred less frequently. Similar to boxplot, The minimum, first quartile, median, third quartile, and maximum are the five summary values used in a violin plot to describe a data set in addition to displaying its overall shape of the data.

Fig. 4.4 represents our baseline model performance in relation to Yan et al. (2019) AHDRNet on all the three datasets. Blue violin plots represents Yan et al. (2019) model and orange violin plot represent our baseline Multi-Attention SKFHDRNet. The data points represents per frame image quality metrics results specifically, μPSNR and HDR−VDP−2. The median is represented by (the red point), the first and third quartile represented by black bar where the lower region of the bar represent first quartile and the upper region of the black bar represent

third quartile, Our baseline model predicted better per frame quality metrics results considering the median in a violin plot which was higher than Yan et al. (2019) AHDRNet on all three datasets. From the results we clearly see an intersection between data points specially in case of Synthetic and Dynamic dataset. This represent the performance of models on low and high exposures samples. The model showed higher performance in case of samples with low exposure which is represented mostly in the third quartile region of the violin plot above the median. While samples with center frame having high exposure are represented below the median red point in the first quartile region of violin plot. But overall our model performance based on μPSNR and HDR$-$VDP$-$2 was higher than Yan et al. (2019) AHDRNet.

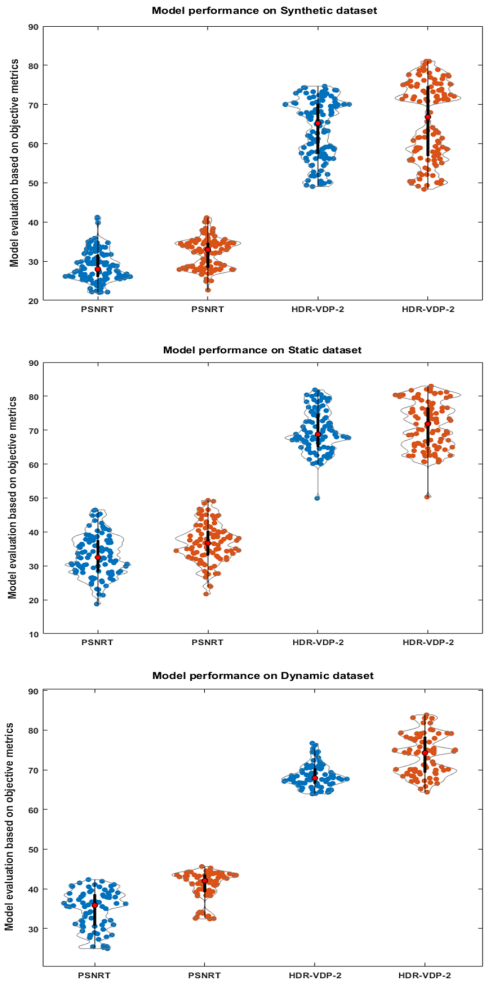


Figure 4.4: *Per frame representation of image quality objective metric results on all three datasets using violin plot of our baseline architecture against Yan et al. (2019) AHDRNet).*

4.3 Evaluation of models with optical flow (no pixel blending)

We also conducted a study by applying optical flow following the work of Kalantari and Ramamoorthi (2019) and Chen et al. (2021a) (see section 3.6). The Multi-

Attention SKFHDRNet and Yan et al. (2019) AHDRNet are trained using two alternating frames. we passed three-6 channel inputs to models. The models first estimated the two optical flows. Warping are then performed on the neighbouring frames using estimated flows for frame alignment. The aligned images are then passed to the Multi-Attention blocks and to the single attention block of Yan et al. (2019) AHDRNet for further alignment. The refined aligned frames are then passed to the merge block for final estimation of HDR frame. The following sections will evaluate the model on a number of HDR test dataset which is discussed in the following section.

4.3.1 Evaluation of baseline models with optical flow on synthetic dataset

The trained models with optical flow are then evaluated on a synthetic dataset. Fig. 4.5 represents visual results of both models in relation the original HDR image. From the visual results, strong decolorization can be seen from the estimated result of Yan et al. (2019) in zoomed region of POKER FULLSHOT scene. Though the result does not show big ghosting artifact due to the addition of optical flow but it introduce new artifacts. While in case of Multi-Attention SKFHDRNet, No decolorization is observed with improved image alignment. The color difference information is captured by CIEDE2000 in estimated HDR scene in relation to the ground truth HDR.

Table 4.4: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet with optical flow on synthetic dataset are represented. Bold text indicates the best among models.*

Model performance on synthetic dataset		
Model	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	28.72	62.90
Multi-Attention SKFHDRNet	**32.56**	**65.98**

The improved Multi-Attention SKFHDRNet performance with optical using three 6-channel frames as input on static dataset using objective quality metrics, μPSNR , and HDR-VDP-2 can be seen in Table 4.4. Our learning-based method produce better image quality metric results on all three datasets as compared to Yan et al. (2019) by estimating high quality HDR images.

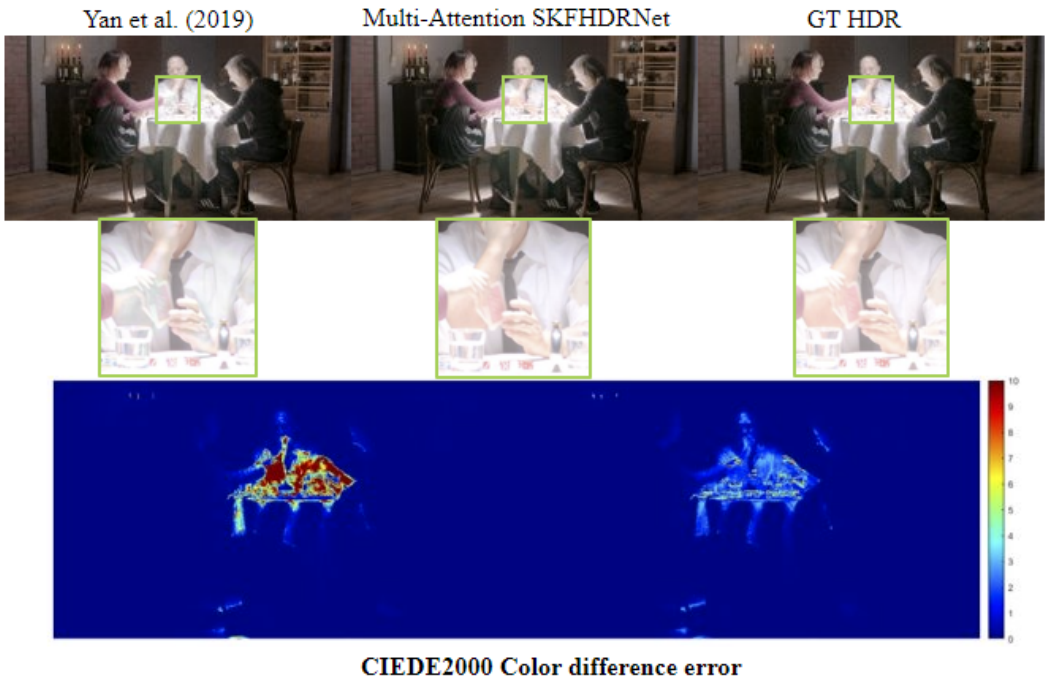


Figure 4.5: *Visual and color difference error results of our baseline model with optical flow on synthetic dataset is represented.*

4.3.2 Evaluation of baseline models with optical flow on dynamic dataset

Using dynamic dataset the baseline models with optical flow using three 6-channel frames as input are evaluated. The optical flow reduce the alignment error in Yan et al. (2019) AHDRNet estimation of HDR scenes. But still it struggle to recover details in challenging cases for instance, the over-exposed sky region in the background of dynamic dataset scene (see Fig. 4.6). Which was not the case in our proposed method estimated HDR image capturing details in the over-exposed regions with improved image alignment which ultimately remove ghosting artifact in the frame.

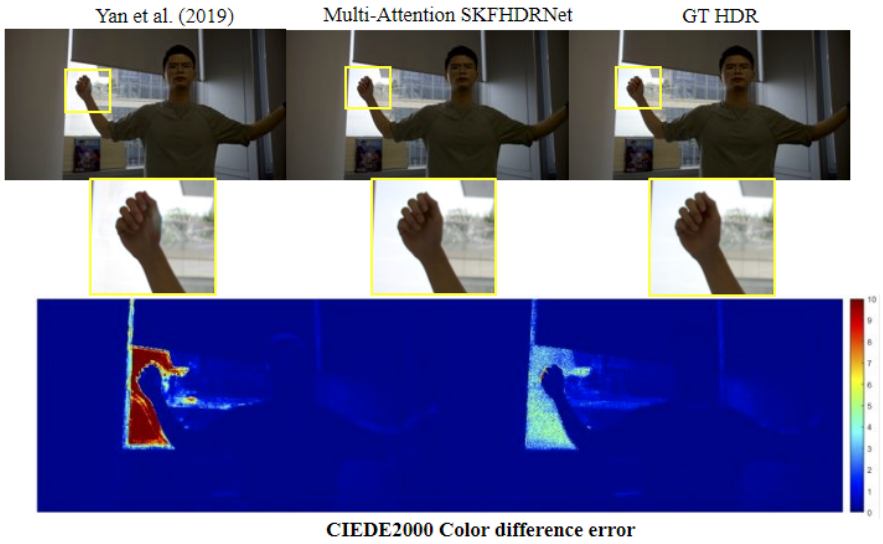


Figure 4.6: *Visual and color difference error results on the dynamic dataset.*

The color difference error map show higher color difference in background window with some error seen in the hand due to ghosting artifact in Yan et al. (2019) AHDRNet result. The Multi-Attention SKFHDRNet performance with optical flow using three 6-channel frames on dynamic dataset using objective quality metrics, μPSNR , HDR-VDP-2 can be seen in Table 4.5. Where our model provide better image quality results on both image quality metrics.

Table 4.5: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet with optical flow on dynamic dataset are represented. Bold text illustrates the best among models.*

Model performance on dynamic dataset		
Model	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	35.08	67.74
Multi-Attention SKFHDRNet	**41.52**	**72.67**

4.3.3 Evaluation of baseline models with optical flow on static dataset

Static dataset are then used to evaluate the baseline models with optical flow using three 6-channel frames as input. Static data set is augmented with random translation of pixel values between range [0, 5] which mean there is less motion

between frames. Both the models estimated HDR have less ghosting artifacts. But Yan et al. (2019) estimated output shows high change in color information related to ground truth HDR. Which produce large colour difference error that are visualize in a zoomed region of estimated HDR frame in Fig. 4.7. On the other hand, our multi-Attention SKFHDRNet produce estimated HDR image with color information similar to the ground truth HDR having less color difference error.

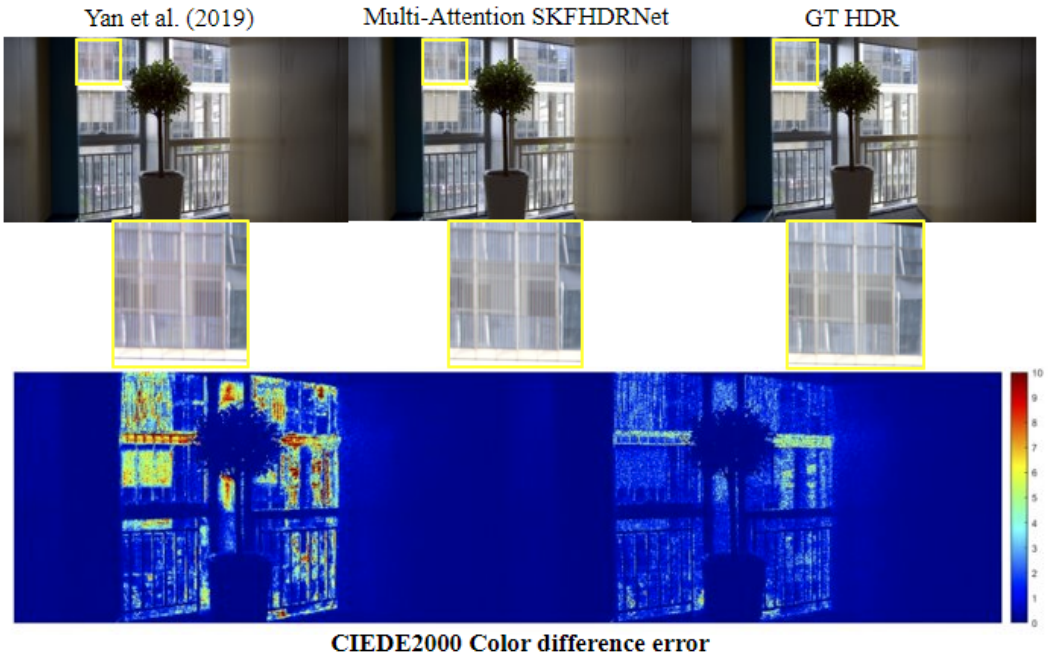


Figure 4.7: *Visual and color difference error results on the static dataset.*

The objective quality metrics, μPSNR , HDR-VDP-2 showed better results regarding our proposed method on all three datasets as compared to Yan et al. (2019) AHDRNet. The quantitative results on static dataset is illustrated in Table 4.6.

Table 4.6: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet with optical flow on static dataset are represented.Bold text indicates the best among models.*

Model performance on static dataset		
Model	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	34.64	71.79
Multi-Attention SKFHDRNet	**38.42**	**73.24**

4.3.4 Per frame objective metric results visualization of baseline model with optical flow

Per frame objective metrics is visualized by using violin plot. Fig. 4.8 represent our baseline model with optical flow with no pixel blending performance in relation to Yan et al. (2019) AHDRNet on all the three datasets using three 6-channels as input. Blue violin plots represents Yan et al. (2019) models and orange violin plot represent our baseline Multi-Attention SKFHDRNet with optical flow results. The data points show the results of per-frame image quality metrics. The first and third quartiles are represented by a black bar in violin plot, with the lower area of the bar being the first quartile and the upper region representing the third quartile. From the median point (in red) in violin plot, our baseline model with optical flow showed more accurate image quality results, As per frame mean results was higher than Yan et al. (2019) AHDRNet on all three datasets (see Fig 4.8). Again from the data points in Fig. 4.8. There is distinct pattern of image quality metrics predictions on HDR test samples. Which is more visible in case of synthetic and dynamic test datasets. This indicates how the models behave on samples with low and high exposures. The violin plot's third quartile area, located above the median, have image quality metrics results for test samples with center frame with low exposure. While the violin plot's first quartile region below the median red point shows the predictions of image quality metrics on test samples with center frame having high exposure. However, in terms of overall model performance and accuracy, our model outperformed Yan et al. (2019) AHDRNet based on μPSNR and HDR-VDP-2 by a large margin.

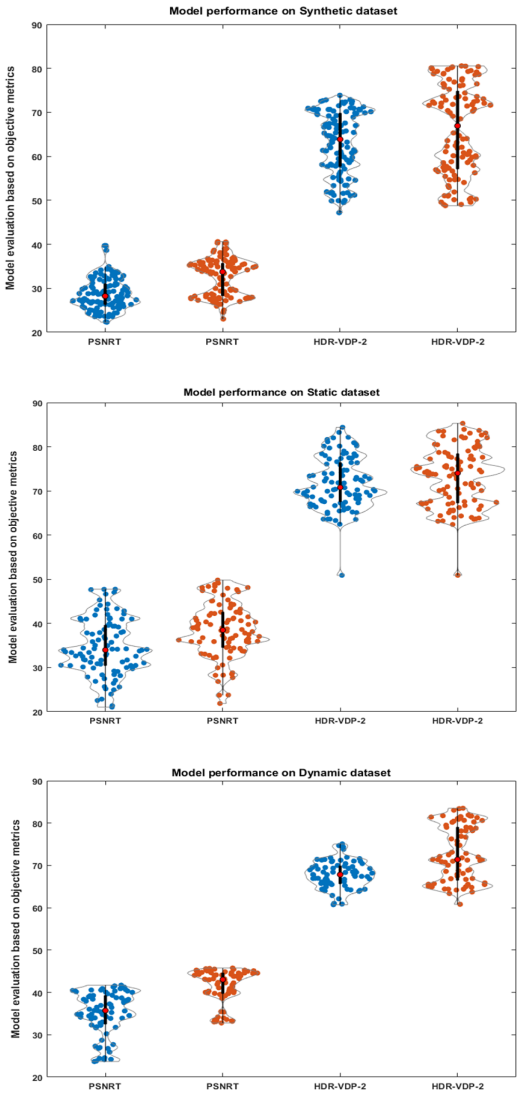


Figure 4.8: *Per frame representation of image quality objective metric results on all three datasets using violin plot of our baseline architecture with optical flow against Yan et al. (2019) AHDRNet).*

4.4 Evaluation of models with optical flow and pixel blending

Similar to the work of Kalantari et al. (2017); Kalantari and Ramamoorthi (2019) and Chen et al. (2021a) for capturing long range dependencies of information, The baseline models are trained this time using five 6 channel frames (30 channels) with two alternating exposures. The optical flow network estimates two flows, these flows are used for warping neighbouring frames for image alignment. The aligned images are then passed to multiple attention blocks for feature alignment. The merge network then estimated five three channel weights from the refined aligned features. The five weights are then merge using a pixel blending strategy to get a final HDR image. The following section will discuss the model performance on synthetic, dynamic and static datasets quantitatively and qualitatively.

4.4.1 Evaluation of baseline models with optical flow and pixel blending on synthetic dataset

Both the models with five 6-channel (30 channel) inputs are evaluated on synthetic dataset. Addition of neighbouring frames along with the usual three consecutive frames increased the performance of the models. Estimated HDR images of both the models showed less ghosting artifacts and details in the over-exposed regions are mostly recovered. But in case of Yan et al. (2019) AHDRNet, The model introduce noisy texture in background in some cases which can be seen in zoomed region of estimated HDR POKER FULLSHOT scene in Fig. 4.9.

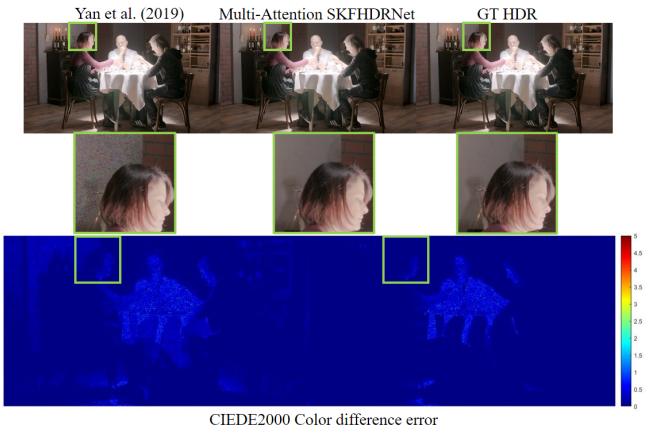


Figure 4.9: *Represents visual and color difference error results on synthetic dataset.*

Multi-Attention SKFHDRNet produce noise free HDR image with smaller color difference error related to ground truth HDR. Which can be visualize in color difference map in Fig 4.9.

Table 4.7 shows the enhanced Multi-Attention SKFHDRNet performance as compared to Yan et al. (2019) employing five 6-channel frames as input with optical flow and pixel blending on synthetic dataset using μPSNR and HDR−VDP−2 and HDR−VQM image and video quality measures.

Table 4.7: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet with optical flow and pixel blending strategy on synthetic dataset are represented. Bold text indicates the best among models.*

Model performance on synthetic dataset		
Model	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	36.49	71.01
Multi-Attention SKFHDRNet	**39.48**	**71.42**

4.4.2 Evaluation of models with optical flow and pixel blending on dynamic dataset

The trained models using five 6-channel frames as input with optical flow and pixel blending module are then evaluated on dynamic dataset which is composed of scenes with large global motion. Previous models of Yan et al. (2019) AHDRNet struggle on dynamic dataset in over-exposed and in large motion regions. The pixel blending strategy improved the performance of the Yan et al. (2019) architecture in recovering details in over-exposed regions but still it introduce ghosting artifacts in samples with large motions (see the zoomed regions of the estimated HDR scenes) in Fig. 4.10. However, our model showed less ghosting artifact in large motion regions with addition of recovering more detail in the over-exposed region.

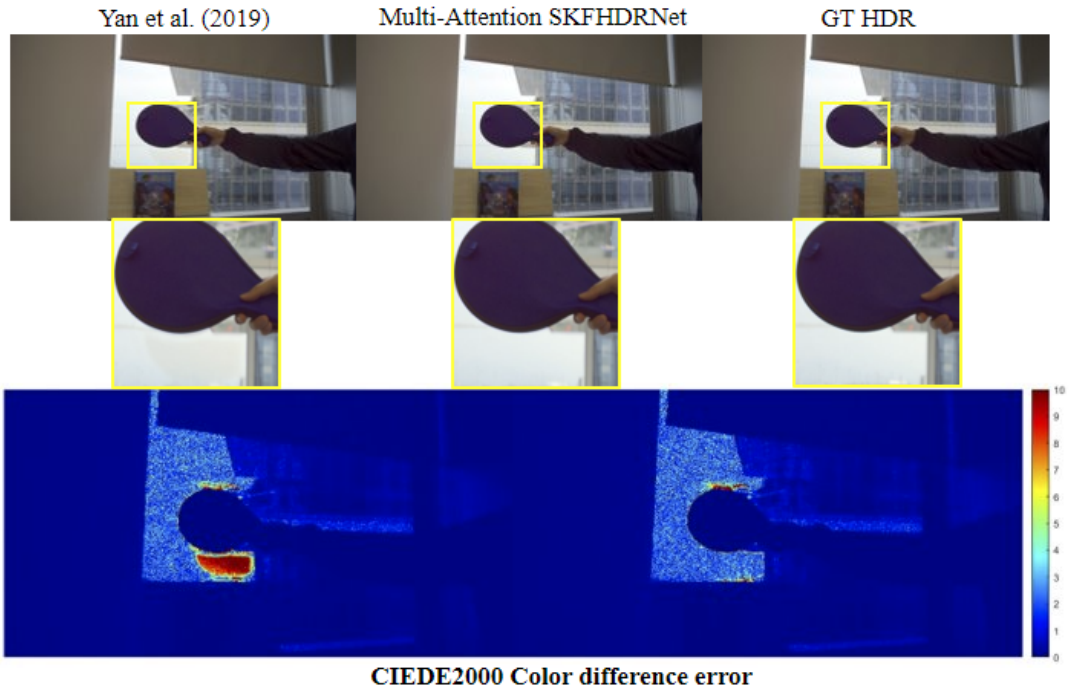


Figure 4.10: *Represents Visual and color difference error results on the dynamic dataset.*

Our model Multi-Attention SKFHDRNet with optical flow and pixel blending showed better results using objective quality metrics, μPSNR and HDR-VDP-2 on dynmaic dataset from Yan et al. (2019) AHDRNet which is illustrated in Table 4.8.

Table 4.8: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet with optical flow and pixel blending strategy on dynamic dataset are represented. Bold text represents the best among models.*

Model performance on dynamic dataset		
Model	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	42.76	78.69
Multi-Attention SKFHDRNet	**45.43**	**79.12**

4.4.3 Evaluation of models with optical flow and pixel blending on static dataset

The trained models are then tested on a static dataset with modification of pixel values by doing random translation in range [0,5], utilizing input of five 6-channel frames with optical flow and pixel blending. Overall, both models were able to restore information in the shadow and overexposed areas. However, In some cases, Yan et al. (2019) AHDRNet estimated HDR image have noisy regions which are visualized in zoomed regions of estimated HDR images in Fig. 4.11. The color difference map also shows higher color error in the highlighted regions because of noise. Multi-Attention SKFHDRNet showed consistent improved results from Yan et al. (2019) AHDRNet by estimating less noisy HDR image which are illustrated in Fig. 4.11 highlighted regions of color difference maps.

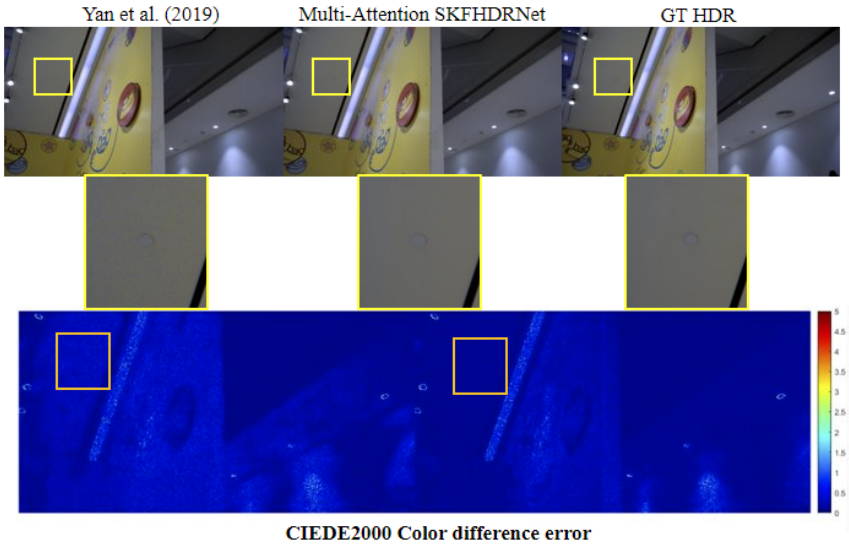


Figure 4.11: *Visual and color difference error results on the static dataset.*

The increased performance of the Multi-Attention SKFHDRNet employing five 6-channel frames with optical flow and pixel blending on a static dataset using the objective quality metrics μPSNR, HDR−VDP−2, and HDR−VQM is shown in Table 4.9.

Table 4.9: *Quantitative results of our baseline Multi-Attention SKFHDRNet and Yan et al. (2019) AHDRNet with optical flow and pixel blending strategy on static dataset are represented. The averaged results for all exposures are shown. Bold text represents the best among models.*

Model performance on static dataset		
Model	$\mu PSNR$	HDR−VDP−2
Yan et al. (2019)	38.05	74.73
Multi-Attention SKFHDRNet	**40.20**	**75.23**

4.4.4 Per frame objective metric result visualization of multi-attention SKFHDRNet with optical flow and pixel blending

Per frame objective metric result visualization is performed using violin plot. Fig. 4.12 represents our Multi-Attention SKFHDRNet with optical flow and pixel blending per frame results of image quality metric predictions on estimated HDR scenes in relation to Yan et al. (2019) AHDRNet on all three datasets with five 6 channels as input. The data points in each violin plot represents the results of per-frame image quality metrics. Overall, our proposed method estimated HDR scenes showed higher image quality results considering the median (red point) in a violin plot, which was higher than Yan et al. (2019) AHDRNet on all three datasets.

Once Again, the results demonstrate a clear separation between the data points, which is more observable in the case of synthetic and dynamic datasets. This shows how both models behave on samples with low and high exposure levels. The third quartile region of the violin plot, which is placed above the median (red point), demonstrates that the model performs well for samples with low exposure as most data points in third quartile represents results for test scene with center frame having low exposure. While the first quartile region of the violin plot displays high exposure samples below the median red point in the violin plot. Which means the data points below median in the first quartile represent test scenes with center frame with high exposure. However, in terms of overall model performance and accuracy, Multi-Attention SKFHDRNet with optical flow and pixel blending module outperformed Yan et al. (2019) AHDRNet based on μPSNR and HDR-VDP-2 by a large margin on all three datasets.

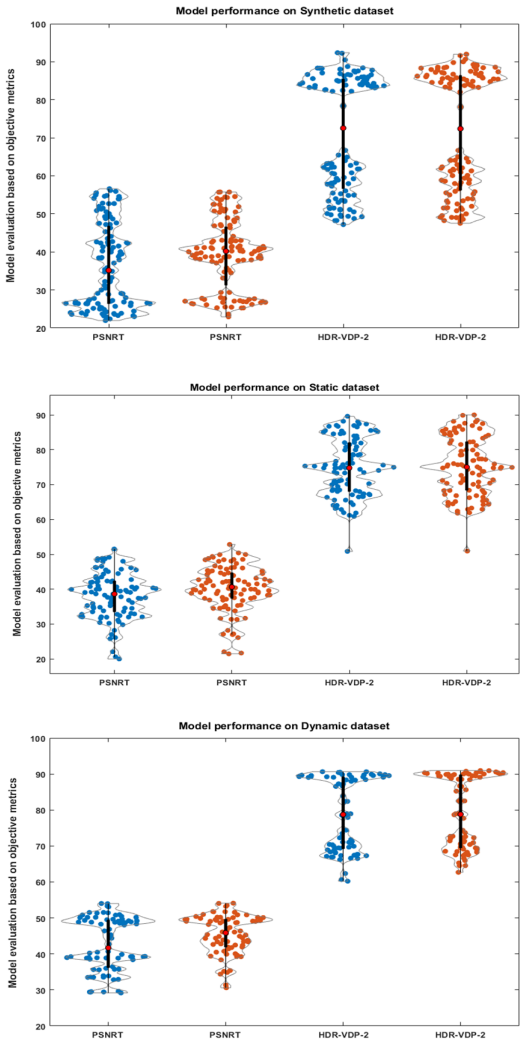


Figure 4.12: *Per frame representation of image quality objective metric results on all three datasets using violin plot on our baseline architecture with optical flow and pixel blending module against Yan et al. (2019) AHDRNet). Blue violin plots represents Yan et al. (2019) models and orange violin plot represent our baseline Multi-Attention SKFHDRNet.*

4.5 Our Full architecture

The final architecture of our model is composed of optical flow, Multi-Attention blocks, Pyramid cascading deformable alignment module and a merge block for final HDR estimation. We also used L_1 and a combined L_1MS−SSIM loss introduce by Zhao et al. (2016) which is considering the properties of L_1 loss with addition of considering spatial pixel dependencies using MS−SSIM. Multi-Attention SKFH-DRNet variants with five 6-channel frames as input having different loss functions were compared and on all the three dataset quantitatively and qualitatively against prior work. We perform evaluation and compare the results with Kalantari et al. (2017); Yan et al. (2019); Kalantari and Ramamoorthi (2019); Chen et al. (2021a) models on all three dataset. The following sections (4.5.1, 4.5.2, 4.5.3) will discuss the performance of our model against the state of the art methods. The visual comparisons of our model variants are only perform against Yan et al. (2019) and Chen et al. (2021a) work. While quantitative comparison of our variant models are perform against Kalantari et al. (2017); Kalantari and Ramamoorthi (2019) along with Yan et al. (2019) and Chen et al. (2021a) networks.

4.5.1 Evaluation on synthetic dataset

Synthetic dataset containing two HDR videos (i.e., POKER FULLSHOT and CAROUSEL FIREWORKS) Froehlich et al. (2014) having 61 frames with the a resolution of 1920x1080, which are not introduce during the training phase are utilized for evaluating the performance of our model. The synthetic dataset is also augmented with random Gaussian noise on the low-exposure images. Fig. 4.13 illustrates the model performance on POKER FULLSHOT HDR scene. From the visual results, The error is more prominent in Yan et al. (2019) AHDRNet estimated HDR image. Our model with L_1 and MS-SSIM loss function produce comparable result similar to Chen et al. (2021a). while in case of our other models with single L_1 loss function, introduce a slight noise in some regions of the estimated HDR scene which are detected by CIEDE-2000 color difference metric. Keep in mind that our model parameters is 50 % less than Chen et al. (2021a) full model while giving almost similar performance in terms of accuracy.

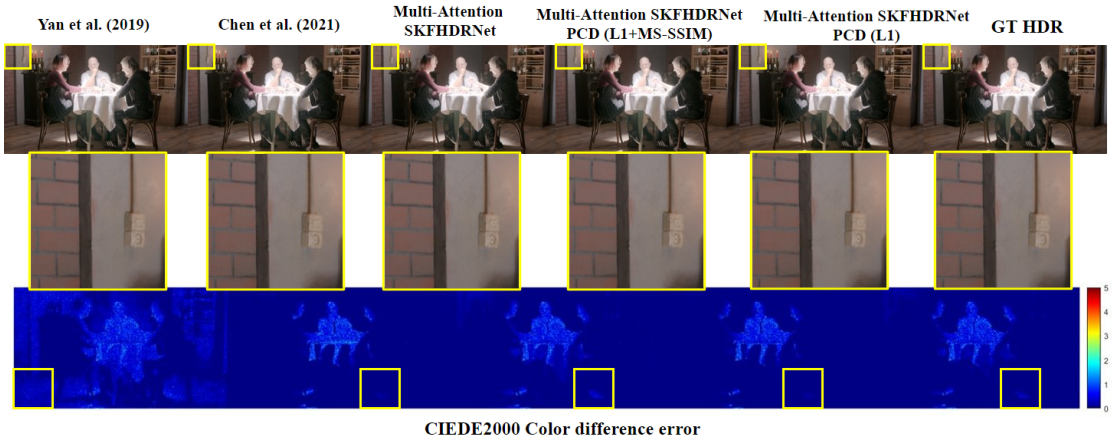


Figure 4.13: *Visual and color difference error results on the synthetic dataset is represented.*

Table. 4.10 shows the overall results of our proposed methodology and the previous methods on synthetic dataset. Overall, Multi-Attention SKFHDRNet variants outperform Kalantari et al. (2017); Yan et al. (2019); Kalantari and Ramamoorthi (2019) models and single (CoarseNet and refineNet) networks of Chen et al. (2021a) on all three objective quality metrics by a large margin.

Table 4.10: *Quantitative results of our Multi-Attention SKFHDRNet variants on synthetic dataset are represented. The averaged results for all exposures are shown. The best model is represented with Red text the second best model is represented by blue text and the third best model is represented by green text, respectively.*

Data set	μPSNR	HDR−VDP2	HDR−VQM
Kalantari et al. (2017)	37.53	59.07	84.51
Yan et al. (2019)	36.49	71.01	69.68
Kalantari and Ramamoorthi (2019)	37.48	70.67	84.57
Chen et al. (2021a) CoarseNet	39.25	70.81	–
Chen et al. (2021a) RefineNet	39.69	70.95	–
Chen et al. (2021a) Full model	40.34	71.79	85.71
Ours (L_1)	39.48	71.42	82.22
Ours PCD ($L_1 + MS - SSIM$)	39.92	71.68	86.04
Ours PCD (L_1)	39.94	71.95	84.05

Our model variant with single L_1 loss function with a PCD alignment module outperform Chen et al. (2021a) full model which is composed of two models

specifically CoarseNet and RefineNet on HDR−VDP−2 image quality metric. Apart from that, our model variant with PCD and $L-1$, MS−SSIM loss function outperform Chen et al. (2021a) full model on HDR−VQM image quality metric. Which clearly indicate our model architecture efficiency having a small number of network parameters.

4.5.2 Evaluation on dynamic dataset

The dynamic dataset contains large local motions which make it challenging for the models to perform well on those cases. Fig.4.14 visualizes the results of our Multi-Attention SKFHDRNet variants along with Yan et al. (2019) and Chen et al. (2021a) models. All of our models clearly shows superior performance in large local motion regions in dynamic dataset scene which can be seen in the zoomed region of dynamic dataset scene in Fig 4.14. The color difference error maps also show large deviation in color information from the original HDR image in the motion regions of the estimated HDR frames of Yan et al. (2019) and Chen et al. (2021a) models.

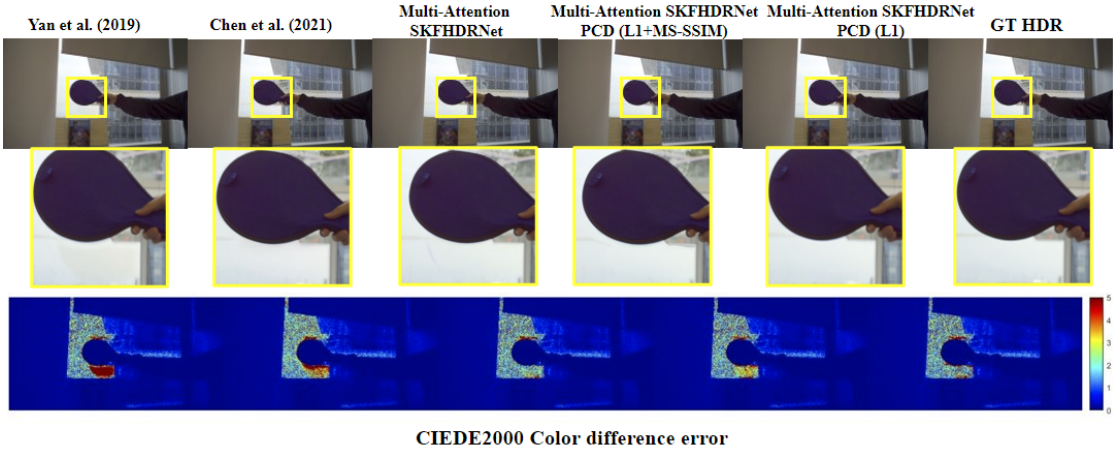


Figure 4.14: *Represents visual and color difference error results on the dynamic dataset.*

Table 4.11 summarizes the results of our Multi-Attention SKFHDRNet variants, where our models showed better results on all the used image quality metrics. The performance of Kalantari et al. (2017); Kalantari and Ramamoorthi (2019) and Yan et al. (2019) significantly drops on dynamic scenes, as this dataset contain scenes with more challenging local motions. Chen et al. (2021a) full model with large number of parameter also struggle in estimating high quality HDR image and produce inferior results as compared to our Multi-Attention SKFHDRNet variants. This give indication of the robustness of our proposed models performance on

Table 4.11: *Quantitative results of our Multi-Attention SKFHDRNet variants on dynamic dataset is represented. All exposures average result is represented here. The best model is represented with Red text the second best model is represented by blue text and the third best model is represented by green text, respectively.*

Models	μPSNR	HDR−VDP−2	HDR−VQM
Kalantari et al. (2017)	41.72	70.36	85.33
Yan et al. (2019)	42.76	78.69	87.27
Kalantari and Ramamoorthi (2019)	44.72	77.91	87.16
Chen et al. (2021a) CoarseNet	44.43	77.74	–
Chen et al. (2021a) RefineNet	43.70	78.97	–
Chen et al. (2021a) Full model	45.46	79.09	87.40
Ours (L_1)	45.43	79.12	88.44
Ours PCD ($L_1 + MS - SSIM$)	44.96	78.49	86.96
Ours PCD (L_1)	45.53	78.89	87.80

samples with large motions using multi attention blocks and the correct usage of PCD alignment block in our architecture.

4.5.3 Evaluation on static dataset

We test our Multi-Attention SKFHDRNet variants on a static dataset that is composed of random global motions. Random translation performed regarding each frame in the range of [0, 5] pixels. For all methods, no pre-alignment is done on input frames similar to Chen et al. (2021a) to evaluate their robustness to input with inaccurate global alignment. Our Multi-Attention SKFHDRNet variants out performed Yan et al. (2019) AHDRNet and Kalantari et al. (2017); Kalantari and Ramamoorthi (2019) learning based methods. Our models also performed better than Chen et al. (2021a) single models (CoarseNet and RefineNet). However, Chen et al. (2021a) full showed slightly better results as compared to our Multi-Attention SKFHDRNet variants. The visual comparison between different models are illustrated in Fig. 4.15. Our proposed model showed comparable results on static scenes in comparison to prior work with half the size of network parameter than Chen et al. (2021a) full model.

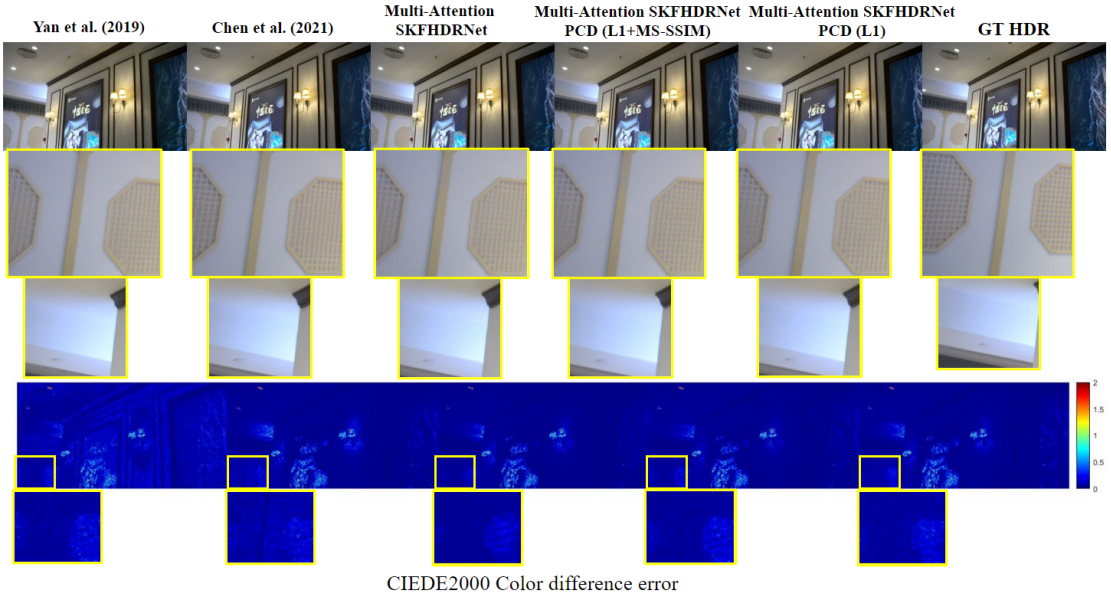


Figure 4.15: *Represents visual and color difference error results on the static dataset.*

Our Multi-Attention SKFHDRNet variant performance in comparison to earlier studies is shown in Table 4.12. Our models showed better results in μPSNR, HDR−VDP−2 and HDR−VQM image and video quality metrics compared Kalantari et al. (2017); Yan et al. (2019); Kalantari and Ramamoorthi (2019) and the single models of Chen et al. (2021a). However, our model performed second best in comparison to Chen et al. (2021a) full model.

Table 4.12: *Quantitative results of our Multi-Attention SKFHDRNet variants on static dataset is represented. All exposures averaged results is represented. The best model is represented with Red text the second best model is represented by blue text and the third best model is represented by green text, respectively.*

Models	μPSNR	HDR-VDP-2	HDR-VQM
Kalantari et al. (2017)	40.02	71.89	76.22
Yan et al. (2019)	38.05	74.73	64.33
Kalantari and Ramamoorthi (2019)	39.88	74.13	73.84
Chen et al. (2021a) CoarseNet	40.62	74.51	–
Chen et al. (2021a) RefineNet	37.61	75.30	–
Chen et al. (2021a) Full model	41.18	76.15	78.84
Ours (L_1)	40.20	75.23	74.05
Ours PCD ($L_1 + MS - SSIM$)	40.47	75.36	75.67
Ours PCD (L_1)	40.62	75.32	75.42

4.5.4 Per frame objective metric results visualization of our full architecture.

Fig. 4.16 represent violin plots of our Multi-Attention SKFHDRNet variants specifically, Multi-Attention SKFHDRNet with L_1 loss, Multi-Attention SKFHDRNet with L_1 loss and PCD alignment module, Multi-Attention SKFHDRNet with L_1MS−SSIM loss along with PCD alignment module. The mentioned model performance are compared with Chen et al. (2021a) network. We did not compare our Multi-Attention SKFHDRNet variants with Yan et al. (2019) model as our baseline architecture was performing superior on all three datasets which was discussed in Section 4.2, 4.3, 4.4.

Figure 4.16 represents violin plot where the blue violin plots represents our model Multi-Attention SKFHDRNet with L_1 loss. The orange violin plot represent Multi-Attention SKFHDRNet with L_1 loss and PCD alignment module. Yellow violin plot represent Multi-Attention SKFHDRNet with L_1MS−SSIM loss function and PCD alignment module. Purple violin plots represent Chen et al. (2021a) model results. Our model variants produce consistent or in some cases showed better results from Chen et al. (2021a) full model considering μPSNR and and HDR-VDP2 per frame image quality results. By looking at the median point in red all the models performance looks almost equivalent. However in some cases like the result of our Multi-Attention SKFHDRNet with PCD and L_1 loss in terms of HDR−VDP−2 image quality metric produce better result by consider the median(red point) point of a violin plot. Overall, the behaviour of all the models were similar, Where all the models performed well in case of HDR test scenes with a center frame under-exposed. While produce inferior results in case of scenes with center frame highly over-exposed. This mean that its easier for the model to recover details in under-exposed samples as compared to over-exposed samples. Most of the data point represent over the median (red point) in violin plot shows the image quality metrics results on under-expose samples while data point below median (red point) are the results of image quality metrics on over-exposed samples.

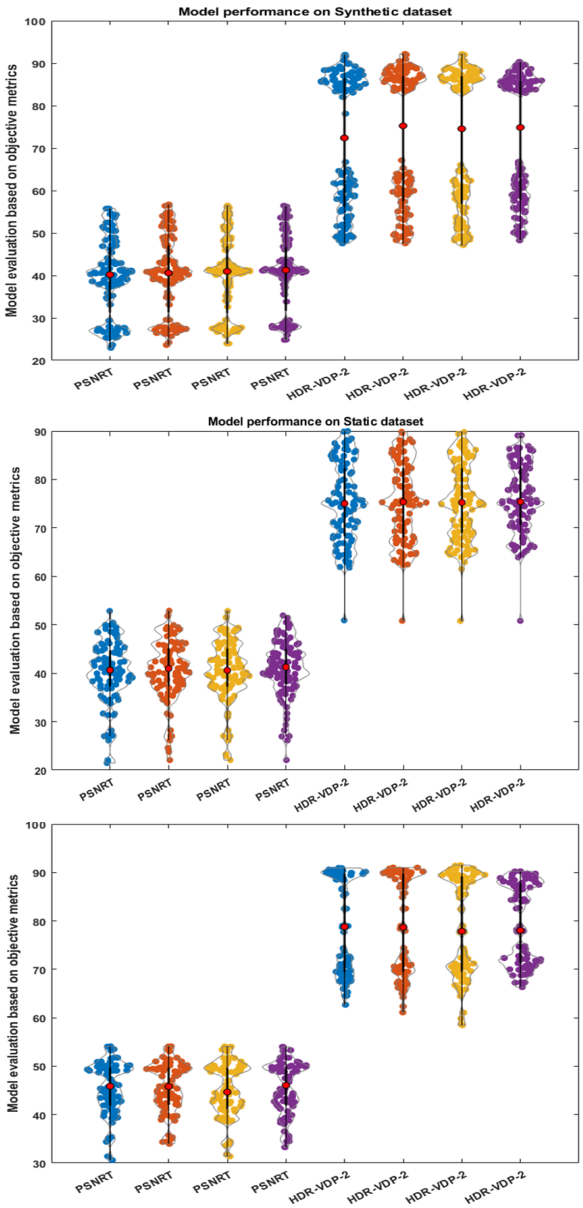


Figure 4.16: *Per frame representation of image quality objective metric results on all three datasets using violin plot of our Multi-Attention SKFHDRNet variants against Chen et al. (2021a)).*

4.6 Network parameters

This section will discuss the comparison of our proposed Multi-Attention SKFH-DRNet based on network parameters with Yan et al. (2019); Kalantari and Ramamoorthi (2019) and Chen et al. (2021a) models. Chen et al. (2021a) full model is composed of 6.1 million parameters, with 3.1M parameters for CoarseNet and 3.0M for RefineNet. While Yan et al. (2019) model contains 1.9M parameters and Kalantari and Ramamoorthi (2019) model have 9.0M parameters mention by Chen et al. (2021a). However, our full model have 2.9M parameters providing almost similar or even surpassing in performance from Chen et al. (2021a) model which have network parameters more than half the size of our model. Our model have higher number of network parameters than Yan et al. (2019) model but our model is performing far superior on all the the three metrics which we used for evaluation. Table 4.7 shows the network parameters of our model with the previous learning-based techniques.

Table 4.13: *Represents our model network parameters with the prior work.*

Models	network Parameters
Yan et al. (2019)	**1.9M**
Kalantari and Ramamoorthi (2019)	9.0M
Chen et al. (2021a) CoarseNet	3.1M
Chen et al. (2021a) RefineNet	3.01
Chen et al. (2021a) Full model	6.1M
Ours full model	**2.9M**

5 Discussion

5.1 Subjective Evaluation

We evaluated our proposed method against prior techniques using HDR image and video quality metrics. Though those metrics are built considering the HVS perceptual capabilities and working in a large luminance ranges and providing an inexpensive way of HDR content quality assessment. However, these image quality assessment (IQA) methods has limitations if we compare it with HVS capabilities. For example, we utilized HDR−VDP2 Mantiuk et al. (2005) IQA method for assessment of our learning-based method. Recently, Narwaria et al. (2014) conducted a study on optimizing HDR−VDP−2 by introducing new pooling strategy to the original HDR−VDP2 IQA metric. Previously, HDR−VDP2 quality prediction performance was tested only on a set of Ponomarenko et al. (2009) TID2008 LDR images. For validating the performance of original, and their improved HDR−VDP2 quality prediction, Narwaria et al. (2014) used HDR images with JPEG 2000 compression errors and tone-mapping distortions instead of LDR images. The proposed optimization introduce by Narwaria et al. (2014) improves the overall performance of original HDR−VDP−2. This indicates limitations of the objective image quality metrics. So for evaluating HDR video reconstruction techniques properly, one should perform psychophysical study by doing visual judgements directly by observers on estimated HDR images. Because in the end, people's quality of experience is the main target of many HDR applications.

5.2 Initial Ablation Study

We conducted an initial ablation study on our proposed Multi-Attention SKFHDR-Net. The idea was to replace the merge network which was composed of multiple DSKFRDBs for final estimation of HDR scene. Transformers are often used in the majority of current research on high-level vision issues like image denoising and super-resolution because they have a defining characteristic called self-attention

that is particularly good at handling long-range pixel inter-dependencies. However, because its complexity increases quadratically with spatial resolution, it is usually impractical to make it usable with high-resolution images for instance in case of image restoration and enhancement. Few attempts have recently been made by Liang et al. (2021) and Wang et al. (2022) to adapt Transformers for tasks that involve image restoration. According to Liang et al. (2021) and Wang et al. (2022), these methods either apply Self-Attention (SA) on small spatial windows of size 8x8 around each pixel or divide the input image into non-overlapping patches of size 48x48 and compute SA on each patch separately in in order to reduce the computational loads which was done Chen et al. (2021b). For our ablation study, we apply the recent work of Zamir et al. (2022) Transformer architecture as our merge network, we follow their same architecture but with reduce no of Transformer blocks as the original model is too big for our purpose. Specifically, the distinctive property of Zamir et al. (2022) is (MDTA) multi-Dconv head 'transposed' attention block which as introduced with the main goal of reducing the limitations of the usual self-attention Transformer block. For the initial work, we trained our Multi-Attention model by replacing the merge network to Zamir et al. (2022) model. Table 5.1 shows the efficiency of Transformer in case of HDR reconstruction as merge network. where it clearly surpass the convolution based architectures in term visual judgment and using HDR image quality metrics.

Table 5.1: *Quantitative results of Transformer architecture as merge network on synthetic test dataset. The proposed methods result are highlighted in bold*

Models	μPSNR	HDR-VDP-2
Yan et al. (2019)	28.78	63.56
Ours Multi-Attention SKFHDRNet)	**32.11**	**65.65**
Ours Multi-Attention with Zamir et al. (2022) Transformer	**34.66**	**66.81**

The visual result in Fig.5.1 also show the superior performance of the Transformer architecture. our initial baseline models without optical flow are compared using the same CAROUSEL FIREWORKS scene. In case of Multi-Attention SKFHDRNet, the estimated HDR scene have regular noise, while in case of Transformer architecture as merge network the estimated HDR scene has a reduced noise.

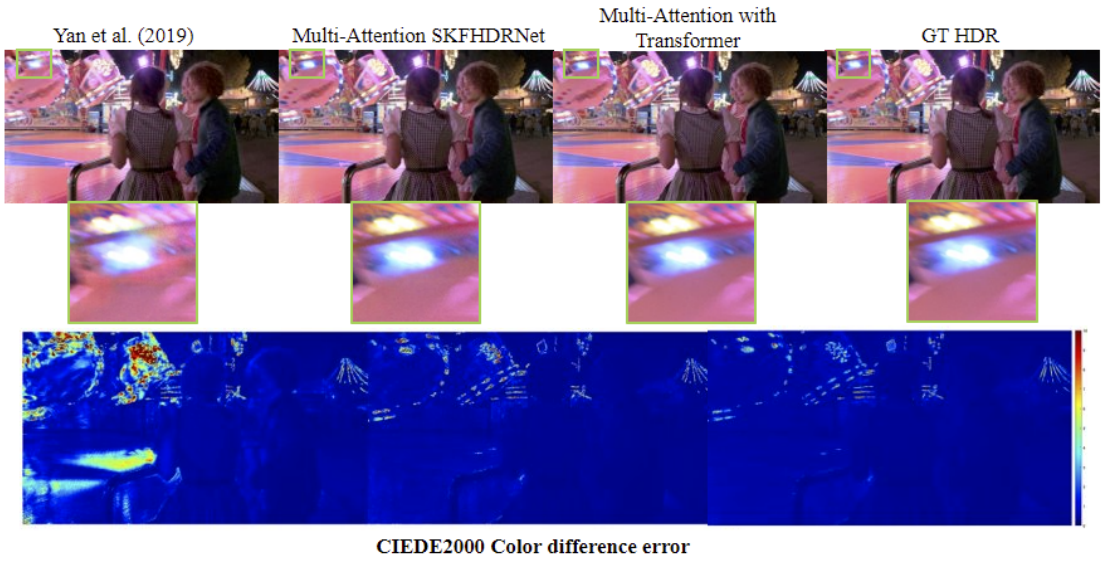


Figure 5.1: *represents the results of our proposed method with Yan et al. (2019). The merge network based on Transformer architecture produce noise free estimated HDR scene while Multi-Attention SKFHDRNet estimated HDR scene have noise which can be seen in the zoomed region in the second row. The models are trained with out optical flow.*

5.3 Limitations of our proposed methodology

Using sequences with alternating exposure for estimating HDR video is a very challenging problem. Although from the results, our approach perform better and produce high quality HDR video. However, some sample use cases were harder and the model struggled to produce satisfactory HDR video reconstruction. One typical example of our model poor performance is observed in cases where the center(reference) frame has highly over-exposed regions and there is apparently large movement of objects during consecutive frames with large occlusion. Which can be seen in Fig. 5.2, our method results in ghosting and other distortions. The optical flow and the Multi-Attention block unable to successfully extract relevant information from neighboring frames in relation to the reference frame those challenging cases. Other methods however, also encounter difficulties in these regions and provided estimated HDR with a similar types of artifacts.

Moreover, in cases where the center (reference) image has low exposure and the neighbouring frames with high exposure contains darker pixels in the same region, This scenario make harder for the models to recover detail in darker regions

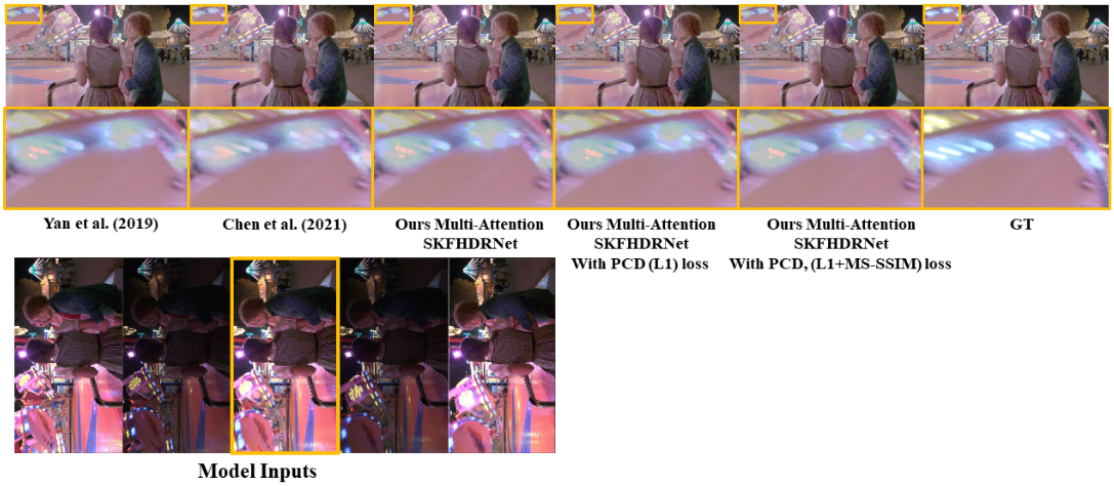


Figure 5.2: *The top row represent estimated HDR scenes for CAROUSEL FIREWORKS scene using two alternating exposures. The bottom row shows the zoomed region where all the models introduced decolorized pixels. By looking at the model inputs, where the center(reference) frame F_i is over-exposed in the highlighted region and the missing content should be recovered from the neighboring frames with low exposure, $F_i - 2$,$F_i - 1$ and $F_i + 1$,$F_i + 2$. Because of significant displacement of objects due to large motions along with high exposure in that region none of the methods are able to properly register and reconstruct details in that region of the image, producing ghosting artifacts which can be seen from the bottom row. Therefore, our method similar to other approaches contains artifacts in this region.*

because the information is very limited in all the frames which produce noise in those regions. Which is illustrated in the zoomed region of static dataset scene in Fig. 5.3. However, our full model results is still considerably better than the other learning-based techniques.

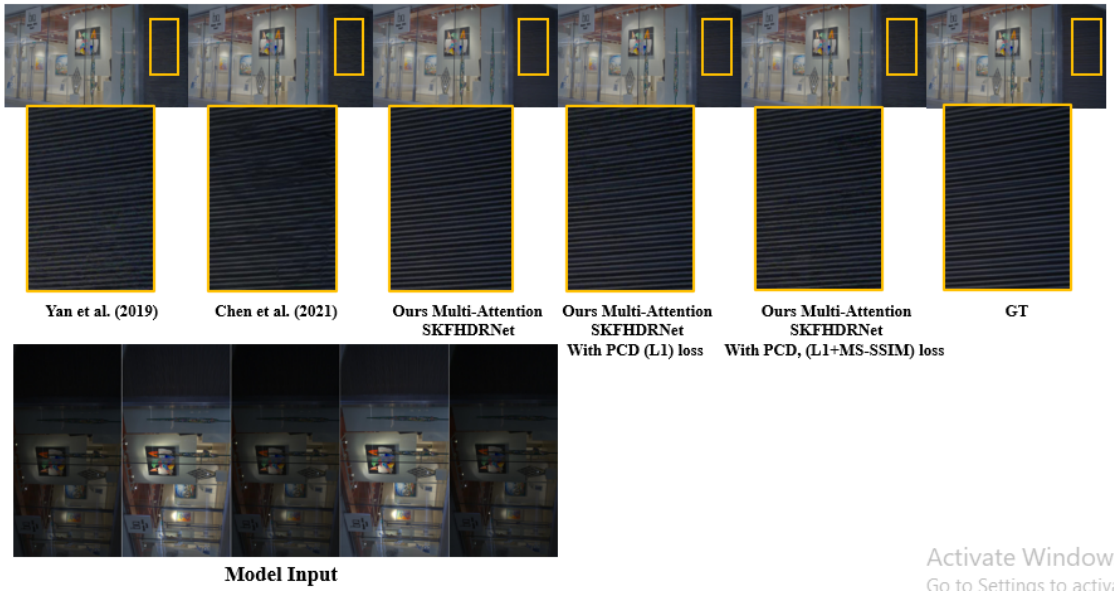


Figure 5.3: *The top row represent estimated HDR scenes for static scene using two alternating exposures. The bottom row shows the zoomed region where all the models introduced noise in the dark region. By looking at the model inputs, where the center(reference) frame F_i is under-exposed and the highlighted region have very dark pixels. Upon that the neighboring frames with high exposure, F_i-2,F_i-1 and F_i+1,F_i+2 also have darker pixel values in the same regions. Due to less information in the middle as well as neighbouring frames, The models produced noisy texture in those region which is visualized in the zoomed sections in the bottom row. Therefore, our method similar to other approaches contains artifacts in this region. However, our multi-attention SKFHDRNet variants have less noisy estimated HDR scene than the other methods.*

6 Conclusion

We approach to the problem of HDR video reconstruction from alternating exposures by presenting a learning based technique. Using optical flow and multi-attentions along with PCD alignment module improved the model performance regarding image alignment and ghosting artifacts for HDR video reconstruction. We compared the performance of multi-attentions modules of our learning-based approach to the prior work both quantitatively and qualitatively. For recovering lost details in under and over-exposed regions, we merged the previously refined aligned features extracted by multi-attentions and PCD alignment module using a series of (DSKFRDBs) for estimating high quality final HDR scene with the addition of utilizing the global residual learning strategy. For optimizing the learning process of optimization algorithm we adopted L_1 and a combined L_1MS-SSIM loss function to minimize the error between the estimated and original HDR images. The proposed learning-based method is trained in an end-to-end manner. We trained our Multi-Attention SKFHDRNet on a synthetic training dataset composed of multiple publicly available HDR datasets with simulated limitations of conventional digital camera system. We demonstrate the performance of our method on a number of HDR test datasets containing challenging cases with over-exposed regions and large motions. The proposed method is evaluated on multiple image and video quality metrics which are specifically build for HDR content. Our learning-based method achieve better results in most cases from the recent state-of-the-art methods with model parameters half the size of the recent state of the art method.

6.1 Future work

Transformers as HDR video reconstruction:
As we already done an initial ablation study(see section 5.2). Where we used the work of Zamir et al. (2022) as our merge network with smaller no of Transformer blocks which showed promising performance in HDR scene reconstruction. In future we will be exploiting the capabilities of Transformer based architectures further.
Optimization for real-time performance:

Considering the real time scenarios, There is further research needed by making the model more interactive by minimizing the inference time of the model. As an example, Performing HDR video estimation with out optical flow network will further reduce the model inference time.

Improvement against challenging over-exposed cases:
Though our methodology showed improved performance regarding recovering details in over-exposed regions of LDR images. But further improvement is required as most of the prior work similar to our proposed method showed inferior performance in recovering missing details in challenging over-exposed examples (see Section 5.3).

Subjective evaluation:
Though we used HDR image and video quality metrics for evaluation which are specifically designed for finding differences between estimated and original HDR images in larger range of scene luminance. But each of the HDR quality metrics has its own limitations in relation to the HDR perceptual capabilities of HVS. It is important to get direct feedback or visual judgement from observers on the results of our proposed method against prior work for more better evaluation. In future, we will be extending the performance evaluation of our model against prior work through visual judgement by conducting psychophysical study and asking real observers to evaluate the model performance.

Additionally, it would be interesting to modify our system to work with different types of capturing setups, as an example, stereo cameras with various exposures.

A Appendix

Table A.1: *Represents major technical specifications of the scenes which was used from Froehlich et al. (2014) dataset for training are summarized in this table for further details about the dataset please refer to Froehlich et al. (2014)*

Data set for training	Frames	fps
Bistro x3	969	24
Car Closeshot	414	25
Car Fullshot	442	25
Car Longshot	820	25
Fire Place x2	952	24
hdr testimage	481	25
Show girl 1	776	25
Show girl 2	341	25
Smith Welding	1102	25
Smith Hammering	467	25
Test data		
Poker Fullshot	600	24
Carousel Fireworks	2536	25

Table A.2: *Represents specifications of the real world scenes captured by Chen et al. (2021a) which was used for evaluating our two alternating exposure model which are summarized in this table. for further details about the dataset please refer to Chen et al. (2021a)*

	Static Scenes w/ GT	Dynamic Scenes w/ GT	Dynamic Scenes w/o GT
	6 - 9 frames	5 - 7 frames	50 - 200 frames
Data Size	2-Exp 3-Exp	2-Exp 3-Exp	2-Exp 3-Exp
4096 × 2168	49 48	76 108	50 50

Bibliography

Aydın, T. O., Mantiuk, R., and Seidel, H.-P. (2008). Extending quality metrics to full luminance range images. In *Human vision and electronic imaging xiii*, volume 6806, pages 109–118. SPIE. (cited on pages 40 and 42)

Banterle, F., Ledda, P., Debattista, K., and Chalmers, A. (2006). Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356. (cited on pages 19 and 20)

Banterle, F., Ledda, P., Debattista, K., and Chalmers, A. (2008). Expanding low dynamic range videos for high dynamic range applications. In *Proceedings of the 24th Spring Conference on Computer Graphics*, pages 33–41. (cited on page 20)

Banterle, F., Ledda, P., Debattista, K., Chalmers, A., and Bloj, M. (2007). A framework for inverse tone mapping. *The Visual Computer*, 23(7):467–478. (cited on page 19)

Boitard, R., Bouatouch, K., Cozot, R., Thoreau, D., and Gruson, A. (2012). Temporal coherency for video tone mapping. In *Applications of Digital Image Processing XXXV*, volume 8499, pages 113–122. SPIE. (cited on page 34)

Chen, G., Chen, C., Guo, S., Liang, Z., Wong, K.-Y. K., and Zhang, L. (2021a). HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2502–2511. (cited on pages 22, 24, 25, 46, 47, 48, 49, 50, 60, 61, 62, 65, 66, 69, 70, 72, 76, 83, 89, 90, 91, 92, 93, 94, 95, 96, 105, and 122)

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., and Gao, W. (2021b). Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310. (cited on page 98)

CIE, C. (2001). Technical report: Improvement to industrial colordifference evaluation. *CIE Publication*, 142. (cited on page 67)

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773. (cited on pages 25, 58, and 59)

Daly, S. J. (1992). Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15. SPIE. (cited on page 41)

Daly, S. J. and Feng, X. (2003). Bit-depth extension using spatiotemporal microdither based on models of the equivalent input noise of the visual system. In *Color Imaging VIII: Processing, Hardcopy, and Applications*, volume 5008, pages 455–466. SPIE. (cited on pages 18 and 19)

De Simone, F., Valenzise, G., Lauga, P., Dufaux, F., and Banterle, F. (2014). Dynamic range expansion of video sequences: A subjective quality assessment study. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1063–1067. IEEE. (cited on page 19)

Debevec, P. E. and Malik, J. (2008). Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. (cited on pages 49, 64, and 69)

Deter, C. and Biehlig, W. (2004). Scanning laser projection display and the possibilities of an extended color space. In *Conference on Colour in Graphics, Imaging, and Vision*, volume 2004, pages 531–535. Society for Imaging Science and Technology. (cited on page 37)

Didyk, P., Mantiuk, R., Hein, M., and Seidel, H.-P. (2008). Enhancement of bright video features for HDR displays. In *Computer Graphics Forum*, volume 27, pages 1265–1274. Wiley Online Library. (cited on page 20)

Drago, F., Myszkowski, K., Annen, T., and Chiba, N. (2003). Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer graphics forum*, volume 22, pages 419–426. Wiley Online Library. (cited on page 35)

Eilertsen, G. (2018). *The high dynamic range imaging pipeline*, volume 1939. Linköping University Electronic Press. (cited on pages 10, 19, and 34)

Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R. K., and Unger, J. (2017). HDR image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15. (cited on pages 20, 21, 47, and 65)

Endo, Y., Kanamori, Y., and Mitani, J. (2017). Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1. (cited on page 20)

Fattal, R., Lischinski, D., and Werman, M. (2002). Gradient domain high dynamic range compression. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 249–256. (cited on page 35)

Ferwerda, J. and Luka, S. (2009). A high resolution, high dynamic range display system for vision research. *J. Vis.*, 9(8):346. (cited on page 37)

Ferwerda, J. A., Pattanaik, S. N., Shirley, P., and Greenberg, D. P. (1996). A model of visual adaptation for realistic image synthesis. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 249–258. (cited on pages 10 and 34)

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394. (cited on page 42)

Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., and Brendel, H. (2014). Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Digital photography X*, volume 9023, pages 279–288. SPIE. (cited on pages 16, 17, 46, 47, 49, 70, 89, 105, and 122)

Gallo, O., Tico, M., Manduchi, R., Gelfand, N., and Pulli, K. (2012). Metering for exposure stacks. In *Computer Graphics Forum*, volume 31, pages 479–488. Wiley Online Library. (cited on page 21)

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings. (cited on page 66)

Granados, M., Kim, K. I., Tompkin, J., and Theobalt, C. (2013). Automatic noise modeling for ghost-free hdr reconstruction. *ACM Transactions on Graphics (TOG)*, 32(6):1–10. (cited on page 23)

Gryaditskaya, Y., Pouli, T., Reinhard, E., Myszkowski, K., and Seidel, H.-P. (2015). Motion aware exposure bracketing for hdr video. In *Computer Graphics Forum*, volume 34, pages 119–130. Wiley Online Library. (cited on page 24)

Guarnieri, G., Albani, L., and Ramponi, G. (2008). Image-splitting techniques for a dual-layer high dynamic range lcd display. *Journal of Electronic Imaging*, 17(4):043009. (cited on page 38)

Guthier, B., Kopf, S., and Effelsberg, W. (2012). A real-time system for capturing hdr videos. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1473–1476. (cited on page 22)

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470. (cited on page 61)

Jack, T. and Holly, R. (1993). Tone reproduction for realistic images. *IEEE Computer Graphics and Applications*, 13(6):42–48. (cited on page 35)

Janesick, J. R., Elliott, T., Collins, S., Blouke, M. M., and Freeman, J. (1987). Scientific charge-coupled devices. *Optical Engineering*, 26(8):692–714. (cited on page 17)

Kajiya, J. T. (1986). The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150. (cited on page 15)

Kalantari, N. K. and Ramamoorthi, R. (2019). Deep hdr video from sequences with alternating exposures. In *Computer graphics forum*, volume 38, pages 193–205. Wiley Online Library. (cited on pages 22, 24, 46, 48, 49, 50, 60, 61, 65, 69, 70, 76, 83, 89, 90, 91, 92, 93, and 96)

Kalantari, N. K., Ramamoorthi, R., et al. (2017). Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1. (cited on pages 24, 46, 48, 49, 50, 69, 70, 83, 89, 90, 91, 92, and 93)

Kalantari, N. K., Shechtman, E., Barnes, C., Darabi, S., Goldman, D. B., and Sen, P. (2013). Patch-based high dynamic range video. *ACM Trans. Graph.*, 32(6):202–1. (cited on pages 23, 24, 49, 64, and 69)

Kang, S. B., Uyttendaele, M., Winder, S., and Szeliski, R. (2003). High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325. (cited on pages 22, 23, and 49)

Kavadias, S., Dierickx, B., Scheffer, D., Alaerts, A., Uwaerts, D., and Bogaerts, J. (2000). A logarithmic response cmos image sensor with on-chip calibration. *IEEE Journal of Solid-state circuits*, 35(8):1146–1152. (cited on pages 16 and 18)

Khan, E. A., Akyuz, A. O., and Reinhard, E. (2006). Ghost removal in high dynamic range images. In *2006 International Conference on Image Processing*, pages 2005–2008. IEEE. (cited on page 23)

Kovaleski, R. P. and Oliveira, M. M. (2009). High-quality brightness enhancement functions for real-time reverse tone mapping. *The Visual Computer*, 25(5):539–547. (cited on page 20)

Kovaleski, R. P. and Oliveira, M. M. (2014). High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56. IEEE. (cited on page 20)

Kronander, J., Gustavson, S., Bonnet, G., and Unger, J. (2013). Unified HDR reconstruction from raw cfa data. In *IEEE international conference on computational photography (ICCP)*, pages 1–9. IEEE. (cited on page 17)

Kronander, J., Gustavson, S., Bonnet, G., Ynnerman, A., and Unger, J. (2014). A unified framework for multi-sensor HDR video reconstruction. *Signal Processing: Image Communication*, 29(2):203–215. (cited on page 47)

Larson, G. W. (1998). Logluv encoding for full-gamut, high-dynamic range images. *Journal of Graphics Tools*, 3(1):15–31. (cited on page 28)

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690. (cited on page 63)

Li, X., Wang, W., Hu, X., and Yang, J. (2019). Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519. (cited on page 55)

Li, Y., Lee, C., and Monga, V. (2016). A maximum a posteriori estimation framework for robust high dynamic range video synthesis. *IEEE Transactions on Image Processing*, 26(3):1143–1157. (cited on pages 24 and 49)

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844. (cited on page 98)

Liu, Y.-L., Lai, W.-S., Chen, Y.-S., Kao, Y.-L., Yang, M.-H., Chuang, Y.-Y., and Huang, J.-B. (2020). Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660. (cited on page 21)

Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., and Liu, S. (2021). Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 463–470. (cited on page 24)

Lulé, T., Keller, H., Wagner, M., and Böhm, M. (1999). Lars ii-a high dynamic range image sensor with a-si: H photo conversion layer. In *1999 IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors, Nagano, Japan*. Citeseer. (cited on page 17)

Luo, M. R., Cui, G., and Rigg, B. (2001). The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350. (cited on pages 67 and 70)

Luzardo, G., Aelterman, J., Luong, H., Philips, W., and Ochoa, D. (2017). Real-time false-contours removal for inverse tone mapped hdr content. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1472–1479. (cited on page 19)

Mangiat, S. and Gibson, J. (2010). High dynamic range video with ghost removal. In *Applications of Digital Image Processing XXXIII*, volume 7798, pages 307–314. SPIE. (cited on pages 23 and 49)

Mangiat, S. and Gibson, J. (2011). Spatially adaptive filtering for registration artifact removal in HDR video. In *2011 18th IEEE International Conference on Image Processing*, pages 1317–1320. IEEE. (cited on pages 22, 23, and 70)

Mantiuk, R. (2015). *High dynamic range imaging*. (cited on pages 2, 3, 10, 11, 15, 16, 18, 23, 25, 26, 27, 28, 31, 37, 38, and 40)

Mantiuk, R., Daly, S., and Kerofsky, L. (2008). Display adaptive tone mapping. In *ACM SIGGRAPH 2008 papers*, pages 1–10. (cited on pages 33, 35, 66, and 70)

Mantiuk, R., Daly, S. J., Myszkowski, K., and Seidel, H.-P. (2005). Predicting visible differences in high dynamic range images: model and its calibration. In *Human Vision and Electronic Imaging X*, volume 5666, pages 204–214. SPIE. (cited on pages 41 and 97)

Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14. (cited on pages 41 and 66)

Mantiuk, R., Krawczyk, G., Myszkowski, K., and Seidel, H.-P. (2004). Perception-motivated high dynamic range video encoding. *ACM Transactions on Graphics (TOG)*, 23(3):733–741. (cited on pages 29 and 32)

Masia, B., Serrano, A., and Gutierrez, D. (2017). Dynamic range expansion based on image statistics. *Multimedia Tools and Applications*, 76(1):631–648. (cited on page 19)

McCann, J. J. (2008). Peceptual rendering of hdr in painting and photography. In *Human Vision and Electronic Imaging XIII*, volume 6806, pages 323–337. SPIE. (cited on page 41)

Meylan, L., Daly, S., and Süsstrunk, S. (2006). The reproduction of specular highlights on high dynamic range displays. In *Color and Imaging Conference*, volume 2006, pages 333–338. Society for Imaging Science and Technology. (cited on page 20)

Miller, S., Nezamabadi, M., and Daly, S. (2013). Perceptual signal coding for more efficient usage of bit codes. *SMPTE Motion Imaging Journal*, 122(4):52–59. (cited on pages 29 and 32)

Mitsunaga, T. and Nayar, S. K. (1999). Radiometric self calibration. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, volume 1, pages 374–380. IEEE. (cited on page 21)

Mukherjee, S., Su, G.-M., and Cheng, I. (2018). Adaptive dithering using curved markov-gaussian noise in the quantized domain for mapping sdr to hdr image. In *International Conference on Smart Multimedia*, pages 193–203. Springer. (cited on page 18)

Narwaria, M., Da Silva, M. P., and Le Callet, P. (2015). Hdr-vqm: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60. (cited on pages 42, 43, 66, and 70)

Narwaria, M., Da Silva, M. P., Le Callet, P., and Pépion, R. (2014). On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment. In *Human Vision and Electronic Imaging XIX*, volume 9014, pages 143–151. SPIE. (cited on page 97)

Nayar, S. K., Branzoi, V., and Boult, T. E. (2004). Programmable imaging using a digital micromirror array. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE. (cited on page 17)

Nayar, S. K. and Mitsunaga, T. (2000). High dynamic range imaging: Spatially varying pixel exposures. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 472–479. IEEE. (cited on pages 16 and 17)

Pattanaik, S. N., Ferwerda, J. A., Fairchild, M. D., and Greenberg, D. P. (1998). A multiscale model of adaptation and spatial vision for realistic image display. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 287–298. (cited on page 34)

Pattanaik, S. N., Tumblin, J., Yee, H., and Greenberg, D. P. (2000). Time-dependent visual adaptation for fast realistic image display. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 47–54. (cited on page 35)

Pharr, M., Jakob, W., and Humphreys, G. (2016). *Physically based rendering: From theory to implementation*. Morgan Kaufmann. (cited on page 15)

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F. (2009). Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45. (cited on page 97)

Ranjan, A. and Black, M. J. (2017). Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170. (cited on page 61)

Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., and Myszkowski, K. (2010). *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann. (cited on page 21)

Reinhard, E., Pouli, T., Kunkel, T., Long, B., Ballestad, A., and Damberg, G. (2012). Calibrated image appearance reproduction. *ACM Transactions on Graphics (TOG)*, 31(6):1–11. (cited on page 34)

Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J. (2002). Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 267–276. (cited on pages 4, 13, and 35)

Rempel, A. G., Trentacoste, M., Seetzen, H., Young, H. D., Heidrich, W., Whitehead, L., and Ward, G. (2007). Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. *ACM transactions on graphics (TOG)*, 26(3):39–es. (cited on page 20)

Robertson, M. A., Borman, S., and Stevenson, R. L. (2003). Estimation-theoretic approach to dynamic range enhancement using multiple exposures. *Journal of electronic imaging*, 12(2):219–228. (cited on page 21)

Seetzen, H., Heidrich, W., Stuerzlinger, W., Ward, G., Whitehead, L., Trentacoste, M., Ghosh, A., and Vorozcovs, A. (2004). Acm siggraph 2004 papers. *New York: ACM*. (cited on pages 37 and 38)

Seetzen, H., Li, H., Ye, L., Heidrich, W., Whitehead, L., and Ward, G. (2006). 25.3: Observations of luminance, contrast and amplitude resolution of displays. In *SID Symposium Digest of Technical Papers*, volume 37, pages 1229–1233. Wiley Online Library. (cited on page 36)

Seetzen, H., Whitehead, L. A., and Ward, G. (2003). 54.2: A high dynamic range display using low and high resolution modulators. In *SID Symposium Digest of Technical Papers*, volume 34, pages 1450–1453. Wiley Online Library. (cited on page 37)

Seger, U., Apel, U., and Höfflinger, B. (1999). HDRC-imagers for natural visual perception. *Handbook of Computer Vision and Application*, 1(223-235):2. (cited on pages 16 and 18)

Själander, M., Jahre, M., Tufte, G., and Reissmann, N. (2019). EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure. (cited on page 66)

Song, Q., Su, G.-M., and Cosman, P. C. (2016). Hardware-efficient debanding and visual enhancement filter for inverse tone mapped high dynamic range images and videos. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3299–3303. IEEE. (cited on page 19)

Sullivan, G. J., Yu, H., Sekiguchi, S.-i., Sun, H., Wedi, T., Wittmann, S., Lee, Y.-L., Segall, A., and Suzuki, T. (2007). New standardized extensions of mpeg4-avc/h. 264 for professional-quality video applications. In *2007 IEEE International Conference on Image Processing*, volume 1, pages I–13. IEEE. (cited on page 32)

Teed, Z. and Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer. (cited on page 61)

Tian, Y., Zhang, Y., Fu, Y., and Xu, C. (2020). Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369. (cited on page 58)

Tocci, M. D., Kiser, C., Tocci, N., and Sen, P. (2011). A versatile HDR video production system. *ACM Transactions on Graphics (TOG)*, 30(4):1–10. (cited on page 17)

Unger, J. and Gustavson, S. (2007). High-dynamic-range video for photometric measurement of illumination. In *Sensors, Cameras, and Systems for Scientific/Industrial Applications VIII*, volume 6501, pages 106–115. SPIE. (cited on page 22)

Van Hateren, J. H. (2006). Encoding of high dynamic range video with a model of human cones. *ACM Transactions on Graphics (TOG)*, 25(4):1380–1399. (cited on pages 33 and 35)

Wanat, R., Petit, J., and Mantiuk, R. (2012). Physical and perceptual limitations of a projector-based high dynamic range display. In *TPCG*, pages 9–16. (cited on page 37)

Wandell, B. A. (1995). *Foundations of vision.* Sinauer Associates. (cited on page 36)

Wang, X., Chan, K. C., Yu, K., Dong, C., and Change Loy, C. (2019). Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0. (cited on pages 58 and 59)

Wang, Z. and Bovik, A. C. (2006). Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–156. (cited on page 39)

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612. (cited on page 40)

Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., and Li, H. (2022). Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693. (cited on page 98)

Ward, G. (1994). A contrast-based scalefactor for luminance display. *Graphics Gems*, 4:415–21. (cited on page 35)

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19. (cited on page 54)

WU HR, R. (2005). Digital video image quality and perceptual coding (signal processing and communications). (cited on page 39)

Xu, D., Doutre, C., and Nasiopoulos, P. (2011). Correction of clipped pixels in color images. *IEEE Transactions on Visualization and Computer Graphics*, 17(3):333–344. (cited on page 20)

Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. T. (2019). Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125. (cited on page 47)

Yan, Q., Gong, D., Shi, Q., Hengel, A. v. d., Shen, C., Reid, I., and Zhang, Y. (2019). Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760. (cited on pages 24, 45, 50, 63, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 96, 98, 99, 120, and 121)

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739. (cited on pages 98 and 103)

Zhang, X. and Brainard, D. H. (2004). Estimation of saturated pixel values in digital color imaging. *JOSA A*, 21(12):2301–2310. (cited on page 20)

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018). Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481. (cited on pages 62, 64, and 120)

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57. (cited on pages 65, 66, 69, and 89)

BIBLIOGRAPHY

List of Figures

List of Tables