



Senior Project

**ON THE INCLUSION OF
VIDEO ATTENTION FOR THE QUALITY OF EXPERIENCE
EVALUATION IN VIRTUAL REALITY ENVIRONMENTS**

Israel de Araujo Nascimento

Brasilia, October 2022

UNIVERSITY OF BRASILIA
Technology Faculty
Undergraduation in Electrical Engineering

Senior Project

**ON THE INCLUSION OF
VIDEO ATTENTION FOR THE QUALITY OF EXPERIENCE
EVALUATION IN VIRTUAL REALITY ENVIRONMENTS**

Israel de Araujo Nascimento

*Report submitted as a partial requirement for obtaining
the degree of Electrical Engineer*

Examining Committee

Prof. Mylène C. Q. Farias, ENE/UnB
Advisor

Prof. Marcelo M. Carvalho, ENE/UnB
Internal Examiner

Prof. Henrique D. Garcia, ENE/UnB
External Examiner

Brasilia, October 2022

CATALOG SHEET

NASCIMENTO, Israel

On the inclusion of video attention for quality of experience evaluation in virtual reality environments

[Brasilia] 2022.

x, 101p., 297 mm (FT/UnB, Engineering, Electrical, 2022). Undergraduation Thesis – University of Brasilia.

1. Computer Vision

2.260 Videos

3. Visual Attention

I. Electrical/FT/UnB

II. Title (Series)

References

NASCIMENTO, I.A (2022). On the inclusion of video attention for quality of experience evaluation in virtual reality environments. Senior Project in Electrical Engineering, Technology Faculty, University of Brasilia, Brasilia, DF, 101p.

ASSIGNMENT OF RIGHTS

AUTHOR: Israel de Araujo Nascimento

TITLE: ON THE INCLUSION OF VIDEO ATTENTION FOR THE QUALITY OF EXPERIENCE EVALUATION IN VIRTUAL REALITY ENVIRONMENTS

DEGREE: Engineer

YEAR: 2022

Permission is granted to the University of Brasilia to reproduce copies of this Undergraduate Thesis and to loan or sell such copies for academic and scientific purposes only. The author reserves other publication rights and no part of this Undergraduate Thesis may be reproduced without the written permission of the author.

Israel de Araujo Nascimento

Darcy Ribeiro.

Brasilia, 70910-900 – DF – Brazil.

Dedication

To my mother, who has always encouraged me to study. To my high school math teacher Valdemir, for inspiring me to pursue the exact sciences, may he rest in peace. And to all those who helped me, I'm the outcome of a collective effort

Israel de Araujo Nascimento

Acknowledgements

First of all, I would like to thank my family, who always encouraged me towards my studies as means to a dignified future. I would like to thank Professor Mylène Farias for introducing me to Image Processing halfway through my undergraduate and again for receiving me for my senior project in the area. I would like to thank the professors at the Institut Fresnel, who provided me with experiences in the area and cemented my desire to research in Computer Vision. I thank the Digital Signal Processing Group (GPDS), in particular Henrique Garcia and Sana Alamgeer, for the time devoted to building my knowledge base in the area of salience detection.

I would also like to thank the many people who come into my life and helped shape the person I am. My mathematics teachers and the OBMEP committees, who made possible that my abilities to the exact sciences be shown from an early age. My foreign language teachers, who inspired me to learn and perfect foreign languages, which ultimately allowed me to have part of my education abroad and write the current text in English. To the people who made my under-graduation years remarkable, in particular my girlfriend. The University gifted me many things, she was the most special. To my friends from University of Brasilia, Ecole Centrale Marseille, Siemens Energy and what else the future reserves, I thank you all.

Finally, to everyone who passed through my life. You were all amazing.

Israel de Araujo Nascimento

RESUMO

O presente trabalho apresenta uma discussão sobre a aplicabilidade da inclusão de atenção visual para o cálculo de valores de métricas de qualidade de experiência. Para isso, dois métodos de cálculo de mapas de saliência foram utilizados: BMS360 e Cubepadding, e suas saliências foram incluídas aos frames utilizando-se uma técnica inspirada em fórmulas da literatura. Com essa inclusão pode-se comparar a qualidade de experiência prevista com e sem a inclusão de atenção visual para ver qual apresenta a melhor performance. Para essa avaliação, utilizou-se um framework recente de cálculo de métricas e estatísticas que permitiu um processamento facilitado e rápida comparação estatística. Por fim, os resultados obtidos mostram que há uma ligeira melhoria nessa inclusão, com grande potencial a uma melhoria ainda maior.

Palavras Chave: saliência, vídeos 360, VQA, atenção visual, viewport, HMD, métricas de qualidade, métricas de desempenho, distorção, VQA-ODV, framework.

ABSTRACT

This work presents a discussion about the applicability of the inclusion of visual attention for the calculation of quality metrics values. To do so, two methods for calculating saliency maps were used: BMS360 and Cubepadding, and their saliencies were included onto the frames using techniques inspired by formulas in the available literature. With this inclusion we were able to compare the predicted quality of experience with and without the inclusion of visual attention to assess which one presents a better performance. For this evaluation, a recent framework was used to calculate the metrics and respective statistics allowing easier processing and fast statistical comparison. At last, the obtained results show a slight improvement with this inclusion, with great potential for an even bigger improvement.

Keywords: saliency, 360 videos, VQA, visual attention, viewport, HMD, quality metrics, performance metrics, distortion, VQA-ODV, framework

SUMMARY

1	Introduction	1
1.1	CONTEXT	1
1.2	EXISTING VQA METHODS	4
1.3	PREDICTING WHERE PEOPLE WILL LOOK	5
1.4	GOALS OF THIS PROJECT	5
2	Visual Attention	7
2.1	VISUAL ATTENTION	7
2.1.1	TEMPORAL FEATURES	9
2.2	BOTTOM-UP AND TOP-DOWN MODELING	10
2.3	SALIENCY PREDICTION APPROACHES	10
2.3.1	HEURISTIC APPROACHES	10
2.3.2	DATA-DRIVEN APPROACHES	12
2.3.3	CUBE PADDING	14
2.4	SALIENCY PRECISION	16
3	Quality of Experience	17
3.1	SUBJECTIVE VQA METHODS	19
3.2	OBJECTIVE VQA METHODS	21
3.2.1	FULL-REFERENCE QUALITY ASSESSMENT METRICS	22
3.3	PERFORMANCE EVALUATION	26
4	Methods and Results	28
4.1	THE SOURCE MATERIAL: THE VQA-ODV DATASET	28
4.2	CHOSEN SALIENCIES	29
4.3	INCORPORATING SALIENCY AND THE VIDEO SIZE PROBLEM	30
4.4	QUALITY METRICS AND EVALUATION	32
4.5	SETUP AND FINE-TUNING	33
4.6	EXPERIMENTAL RESULTS	36
4.6.1	BMS360: CHOSEN VIDEOS AND SALIENCY INCORPORATION RESULTS	37
4.6.2	CUBE PADDING: CHOSEN VIDEOS AND SALIENCY INCORPORATION RESULTS	39
5	Conclusion	43

REFERENCES	46
6 Appendix	51

LIST OF FIGURES

1.1	(a)Model for 360° vision in an HMD. The sphere surrounding the apparatus denotes its omnidirectional nature. Note the typical aviation nomenclature. (b) Typical user wearing a state-of-the-art HMD device.....	2
1.2	Example of distortion in the sphere projection onto the plain: Mercator projection. Observe how the polar regions end up warped.	2
1.3	Examples of projections: Equirectangular (ERP) and Cubemap Projection (CMP) for the same frame. Figure taken from [46].	3
2.1	Example of the heatmap representation of saliency maps. The original image here is directly followed by three examples of maps from different algorithms.	8
2.2	Simple example of how movement is manifested from frame to frame.	9
2.3	Example image and its salience as generated by GBVS. The heatmap representation is sometimes used for visual purposes. Here, hotter spots represent where most people will tend to focus their attention when looking at the image.	11
2.4	Pipeline of BMS from [53]. The boolean maps are computed with the theory presented in [16] and have the attention maps as subproducts. By then summing the maps we get the mean attention map which gives enough information for a saliency map.	12
2.5	Example frame and its saliency as generated by BMS360 on a desktop computer. Notice that this saliency representation is shown in grayscale	13
2.6	Logic behind the idea of Cube Padding. By extending a face to include some of the contents from an adjacent face we can mitigate the discontinuity problem of having six different images per frame to process.	13
2.7	Comparison between the usage of zero-padding (ZP) and cube-padding (CP). Note how the strategy Cube Padding allows for a continuous response across the faces, whereas zero padding does not.	14
2.8	Example of a saliency map generated using Cube Padding [6].....	15
2.9	Example of how the elaboration of a saliency map is done in experiments with humans.	15
3.1	End-to-end 360-degree video processing pipeline	18
3.2	Examples of different popular HMD models.....	20
3.3	Subject participating in an experiment wearing an Oculus device during the pandemic. Notice the lack of physical obstacles near the chair.....	20

3.4	Comparison of images with same MSE (a) Original Image; (b) More contrast, MSSIM = 0.9168; (c) Displaced average MSSIM = 0.9900; (d) JPEG Compressed MSSIM = 0.6949; (e) Blurred Image MSSIM = 0.7052; (f) Salt and Pepper noise MSSIM = 0.7748.....	24
4.1	Saigg and Scholles’s framework ans shown in [36].....	33
4.2	The full flowchart representing the methods and subproducts of our analysis. Here, the rhombus represents inputs/outputs and the rectangles represent the operations.	34
4.3	Examples of saliency maps generated by the two algorithms analysed, Cube Padding and BMS360. The image on the left shows the BMS360 saliency and the image on the right shows the Cube Padding saliency map. We can easily see the borders of the cube in this projection.	35
4.4	Performance distribution for the BMS360 videos. We can see an upward trend, but with a lot of variance for PSNR and VMAF. Points for MS-SSIM tend to be accumulated towards the higher extremes.....	36
4.5	Performance distribution for BMS360 videos with saliency incorporation. The overall distribution is relatively similar. That means that visually we cannot see a clear improvement thanks to this incorporation.	37
4.6	Performance distribution for the Cube Padding videos.....	38
4.7	Performance Distribution for the Cubepadding Videos with Saliency Incorporation. The distribution is relatively similar to the one shown in figure 4.6.1, but in MS-SSIM and VMAF we see some improvement in terms of variance.	39
4.8	Variation for PSNR with different saliency inclusions.....	40
4.9	Variation for MS-SSIM with different saliency inclusions	41
4.10	Variation for VMAF with different saliency inclusions	42

LIST OF TABLES

4.1	List of all videos used from the VQA-ODV in this research, their respective dimensions and MOS values.	35
4.2	Table of VQA Metrics for the BMS360 videos with and without saliency incorporation.	39
4.3	Table of VQA Metrics for the Cubepadding videos with and without saliency incorporation.....	40
6.1	Full information table for BMS360 without saliency inclusion.	51
6.2	Full information table for Cubepadding without saliency inclusion.....	52
6.3	Full information table for BMS360 with saliency inclusion.	53
6.4	Full information table for Cubepadding with saliency inclusion.....	54

ABBREVIATIONS AND ACRONYMS

VR	<i>Virtual Reality</i>
HMD	<i>Head-Mounted Display</i>
QoE	<i>Quality of Experience</i>
VQA	<i>Visual Quality Assessment</i>
ERP	<i>Equirectangular Projection</i>
CMP	<i>CubeMapping Projection</i>
BMS	<i>Boolean Map Saliency</i>
CP	<i>Cube Padding</i>
PSNR	<i>Peak-Signal to Noise Ration</i>
SSIM	<i>Structural Similarity Index</i>
VMAF	<i>Video Multi-method Assesment Fusion</i>
HM	<i>Head Movement</i>
EM	<i>Eye Movement</i>
VQA-	<i>Visual Quality Assessment - OmniDirectional Videos dataset</i>
ODV	

Chapter 1

Introduction

Nowadays, 360° videos are becoming increasingly popular in communication. These videos aim to emulate human vision, allowing for a full 360° span and exploring the environment around the user, as shown in Figure 1.1. In this image we see that the aviation nomenclature of Pitch, Yaw and Roll is often used: pitch is related to the side-to-side head movement, yaw to the up-and-down movement and roll is the back and forth moment.

In Figure 1.1 it is shown the way the HMD device is worn. This kind of video is particularly useful in the context of Virtual Reality (VR) due to its immersive nature. In VR, users normally make use of a device called *Head Mounted Display (HMD)*, which mimics concrete 3D reality. This creates a few problems:

- The need to send high-fidelity information to the HMD requires high video resolution (usually above 4K) [18]. Added to that, a high *frame rate* makes transmission particularly challenging. To mitigate this problem, sending only information related to what the viewer can see at a given moment, i.e., to the *viewport* in high resolution, while sending other areas in low resolution is one of the main techniques proposed [46].
- Projecting 360° videos to a 2D surface in order to make use of well-established methods for 2D videos warps the signal due to oversampling in certain areas of the video frames. A classical example is the cartographic distortion as seen in Figure 1.2. This inspires the creation of a plethora of adaptations for the methods to work in the spherical conditions of 360° videos.

1.1 Context

Due to the massive amount of information in a 360° video, the most common approach today for processing it is the partitioning of the sphere into tiles [7] and the attempt to predict the users' fixation points through visual attention techniques called salience maps [2]. The latter and the discussion of the existing technique is the subject of Chapter 2. All these elements change

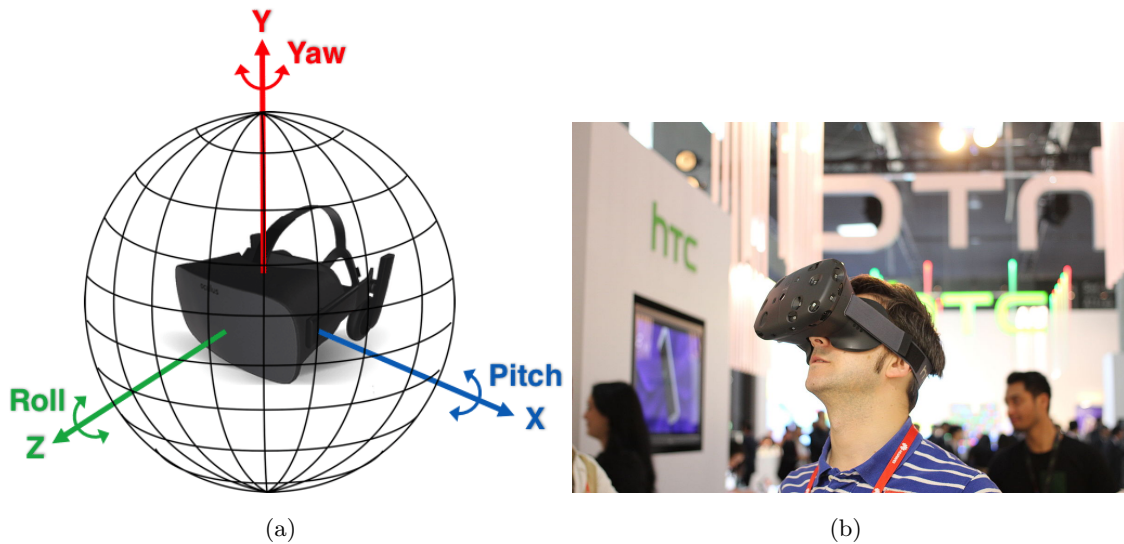


Figure 1.1: (a) Model for 360° vision in an HMD. The sphere surrounding the apparatus denotes its omnidirectional nature. Note the typical aviation nomenclature. (b) Typical user wearing a state-of-the-art HMD device.

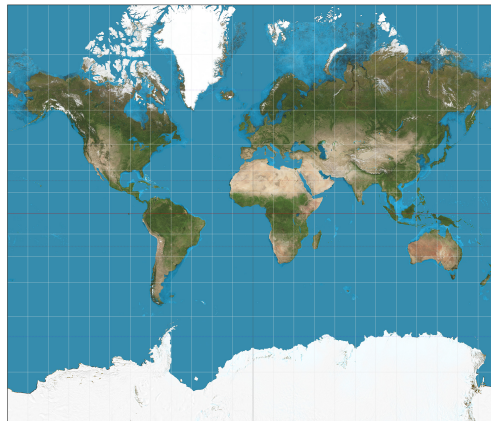


Figure 1.2: Example of distortion in the sphere projection onto the plain: Mercator projection. Observe how the polar regions end up warped.

how users perceive videos and their Quality of Experience (QoE). The video quality received after processing is subject to Visual Quality Assessment (VQA). The explanation of VQA concepts is explored in Chapter 3 and is the central theme of research in development.

The topic of distortion motivates the existence of various forms of projection. Projection is the term used to describe a broad set of transformations employed to represent the two-dimensional curved surface of a globe on a plane. In a map projection, coordinates, often expressed as latitude and longitude, of locations from the surface of the sphere are transformed to coordinates on a plane. Projection is a necessary step in creating a two-dimensional map and is one of the essential elements of cartography. Although the equirectangular projection (ERP)¹ is not ideal, it is usually

¹This is a type of projection for mapping a portion of the surface of a sphere to a flat image. It is also called the "non-projection", or plate carre, since the horizontal coordinate is simply longitude, and the vertical coordinate is

used for its simplicity and for historical reasons [7]. Its main problem is the dimensioning of poles with more pixels than the equator. However, this distortion may not be so detrimental to the final experience because of what is called equator bias, which asserts that humans tend to focus mainly on the equator of an image or video [7].



Projection	Illustration
Equirectangular Projection (ERP)	
Cubemap Projection (CMP)	

Figure 1.3: Examples of projections: Equirectangular (ERP) and Cubemap Projection (CMP) for the same frame. Figure taken from [46].

Another main projection method is the Cubemap Projection (CMP). It builds a cube around the spherical field of view and projects rays from the center of the sphere. Each ray crosses a single point on the surface of the sphere and the cube, resulting in the mapping. CMP is more efficient than ERP in terms of compression, and because of that, several new methods have been proposed based on this idea. Figure 1.1 compares both types of projection. In the first figure we can see how the sphere can be projected panoramically onto the plain, note that there is a continuity between the left and right borders. In the second figure the sphere is spread out onto the 6 faces of the cube. The result is a less distorted projections but with more discontinuities. The cubic projection is of particular interest in this work, as one of the methods described in Chapter 2 uses the cubic projection.

There are many types of projections, such as the conic and polyhedral projections, but because of their specialized nature they are beyond the scope of this work. These different projections tend to have a balance of the distortion-discontinuity trade-off tipped to one of the sides, but they are quite computationally complex. For a more detailed description, please refer to [7].

simply latitude, with no transformation or scaling applied.

1.2 Existing VQA methods

Visual Quality Assessment, that is, the evaluation of how good the content someone is seeing. In our case, we are mostly concerned with 360° videos, but the idea can be expanded to any form of visual content.

Distortion is a fundamental problem in image compression and transmission. In each stage of the 360-degree video communication system, distortions may be introduced. For example, distortions can be introduced in the acquisition, compression, and transmission of the content, leading to a degradation of the user’s VQA. In experiments, distortions are usually introduced at various strengths by varying the parameters of the system setup, obtaining effects similar to those in the real situations. All distortions are, in essence, a deviation from a reference signal that may or may not impair the user’s experience.

This distortion can impact how good is the content in the user’s perspective. A simple way to evaluate how good is the content people are seeing is by simply asking them. That is, we make them watch the content we want to evaluate and directly ask them to rate the quality of what they see on a scale of 1 to 5, or 1 to 100 for example, which is the idea behind the **subjective methods** for quality evaluation. Here the methods vary mostly in how viewers will watch the videos. If they are watching the video with some deterioration with nothing to compare, then we call it a Single Stimulus Experiment. Now if they have the video with no deteriorations (called reference video) to compare, then we have double stimulus.

Now we can very quickly see how this method of asking people to watch the videos can become cumbersome. We would have to invite a conglomerate of subjects to watch many video sequences and then ask them their opinion. Imagine we do this for every single existent 3D video? This difficulty motivates the approximation of the actual users’ VQA by computing metrics of signal fidelity.

This is an area quite well developed for 2D videos, and this development inspires metrics for 3D videos. Methods like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [15], multi-scale SSIM (MS-SSIM) [43], Visual Quality Model (VQM) [32]. It is important to note that none of these methods, although quite efficient in 2D, are not perfectly suitable for the particularities of 3D.

Some adaptations of 2D objective quality assessment methods have been proposed for 360° images and videos. For example, Spherical PSNR (S-PSNR) [51] tries to account for traditional spherical characteristics of omnidirectional videos. Its weighted version called WS-PSNR [54] tries to account for the existing distortion of the projection when assigning different weights to the different pixels according to their place in the frame. There are still more complex methods, like Perceptual Video Quality (PVQ) [55] and recent strategies using the machine learning method of Convolution Neural Networks (CNN), for example Viewport-based CNN (V-CNN), as proposed in [25].

Regardless of the chosen method, the idea is to make use of the subjective experiment as the ground truth and, therefore, to be able to analyze how the different methods compare in predicting

the VQA. This comparison is done using statistical methods for performance indication, such as the Pearson Correlation Coefficient(PCC) which is a measure of linear correlation between two sets of data.

The importance of analysing VQA is that the VQA metrics impact in the overall engagement of viewers and therefore determines the success of a particular video. The better the actual quality, the longer and more often users tend to tune in. It is worth pointing out that throughout this work, we will be analysing only VQA, that is, visual quality assessment. Quality of Experience (QoE) encompasses VQA, and it is important to note that in addition to the visual quality evaluated by VQA, it also takes into account other elements such as Quality of Service (QoS) [40] and other problems such as the sensation of presence and cybersickness [18]. These other concepts, although relevant for quality analysis, are not the focus of this text.

1.3 Predicting where people will look

Due to the quality in which videos have to be transmitted, being able to roughly predict where people will look in order to potentially transmit the most attractive regions with more quality is a good idea. This is the main motivation behind visual attention and saliency maps, which is the task of predicting where people will look. As we can imagine, this prediction is quite a complicated task, as human attention depends on several aspects, like personal interests, storytelling, movement, etc. It is a topic of debate with new models emerging all the time and will be the main topic of Chapter 2. Although it is a hard and computationally expensive endeavor, once we have an accurate model, it can be standardized in the producers' ends so that streaming services can know beforehand where best quality has to be directed.

Besides this goal, saliency prediction is a relevant topic in storytelling, since a director has in mind a story he is trying to transmit, and if people are not following that story in the 3D space, i.e., the main line does not capture their attention, the whole storytelling experience is ruined. Aligning the story with attention-catching mechanisms is therefore vital for a correct experience, particularly in virtual reality environments.

1.4 Goals of this project

The goal of this research is to integrate visual attention into methods for evaluating visual quality and compare them with the mean opinion scores (MOS) of volunteers who watched the same videos, asserting how visual attention impacts the prediction of the users' VQA. In order to do that, two saliency models are studied and incorporated onto the studied videos and the difference in the final quality metrics results is analysed. Due to limitations imposed by the sanitary conditions of the Covid-19 pandemic, the development of a data set of University of Brasilia authoring is not feasible. Therefore, it was decided to use Mai Xu's dataset [46], which is a very robust and complete dataset that includes 60 reference videos and viewers opinions as our source of videos and users' opinions.

The content is organized as follows:

- Chapter 2 is a study of the literature where we explore the main concepts behind visual attention, i.e. where people look at.
- Chapter 3 is a study of the literature in the concepts behind visual quality assessment, both experimental and computational.
- In Chapter 4 we discuss the concepts behind our proposal and the ensued results.
- Chapter 5 closes our text with the major conclusions and future ideas.

Chapter 2

Visual Attention

In this chapter we present an overview of visual attention, that is, where people tend to look, how we can predict their attention using saliency maps and how we can compare prediction with the actual places people look.

2.1 Visual Attention

Visual attention is the name of a wide area within psychology that tries to explain and model elements of human visual attraction. The human eye and brain do not form an indiscriminate machine for scene processing, so there are regions of greater and lesser attraction. Experiments attempting to empirically understand what attracts human attention have been conducted since the beginning of the 20th century.

This analysis is facilitated with the use of eye-trackers. These devices contain head support and cameras pointed at the user's eyes in order to register gaze fixations. These gaze fixations tend to be considered meaningful after the threshold of 200 milliseconds in a single spot [27]. The eye movement that generates these areas is commonly abridged as EM (Eye Movement).

Concerning computational modeling, the theory developed by Treisman and Gelade in [39] represents its most fundamental text. In their work, it is explained that visual information is processed in the human brain by combining different features to identify salient regions. Salient regions are those capable of attracting more of the human attention than others. From this arises the concept of saliency map, which attempts to convey the most relevant parts of the scene. This idea was initially formulated by Koch and Ullman in [21] in their model combining visual characteristics (color, intensity, orientation) to create maps. This model was subsequently implemented by Itti *et al.* [19], from where a whole new field of research originated.

Saliency map is an image in which the brightness of a pixel represents how attractive the pixel is relatively to the other pixels i.e., brightness of a pixel is directly proportional to its saliency. It is generally a grayscale image or heatmap image. But any sort of 2D array representation works.

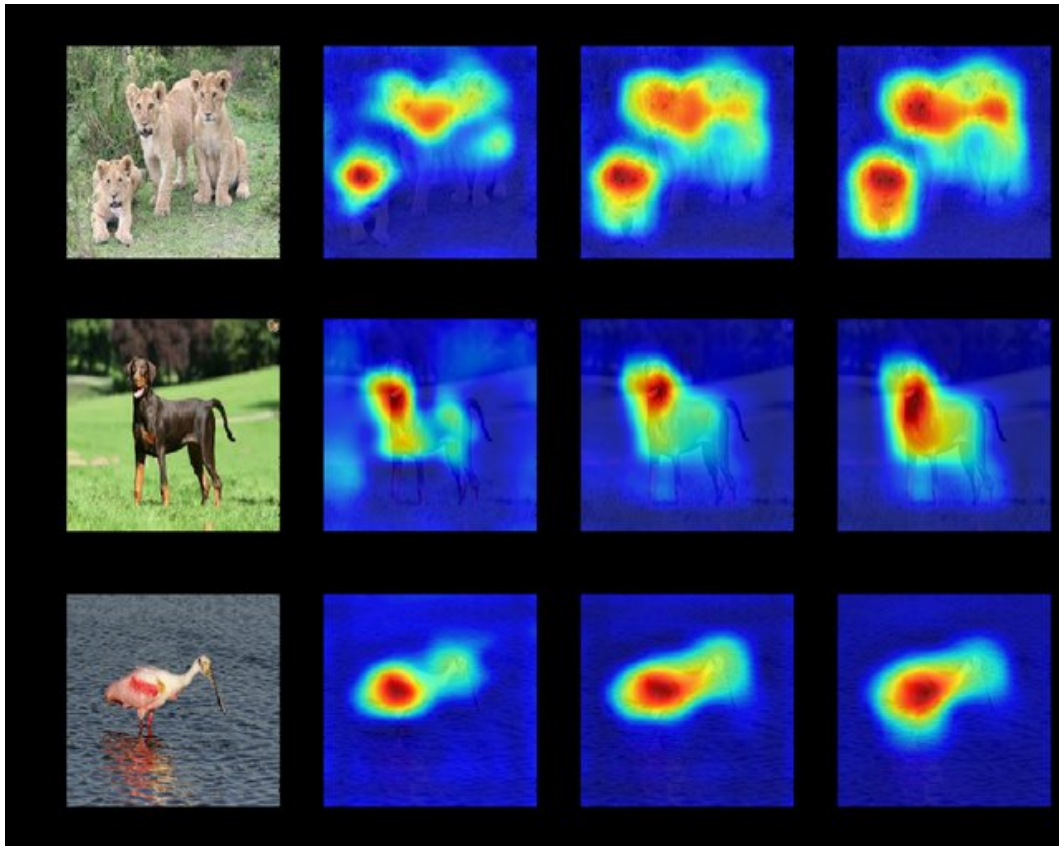


Figure 2.1: Example of the heatmap representation of saliency maps. The original image here is directly followed by three examples of maps from different algorithms.

The purpose of the saliency map is to find the regions which are prominent or noticeable at every location in the field of view and to guide the selection of important locations, based on the spatial distribution of saliency. Figure 2.1 shows an example of original images and their respective saliency maps. Notice that although different, all algorithms were able to tell that the animal in the picture is what draws people’s attention.

In experiments with 2D videos, the fixation points are taken by pointing a camera directly towards users’ eyes and projecting where they are looking. Since in the case of 360° videos users are visually surrounded by what is projected in their viewport (the framed area on a display screen for viewing information), unless we install cameras inside the HMDs we cannot proceed with the same type of analysis. So we can see how taking these measurements can be difficult.

What we could more easily do is measure the exact Head Movements from the users. If we could somehow correlate the Head Movements to the Eye Movements, then we would not have to install cameras inside HMDs. Rai *et al.* [34] found in their research that there is in fact a certain correlation. It was found they are different, but exploiting this correlation, some experiments have come about estimating saliency maps with only head movements, whereas more robust experiments take both head and eye movements into account for their computations.

With all that, many video databases and their respective Head and Eye movements have

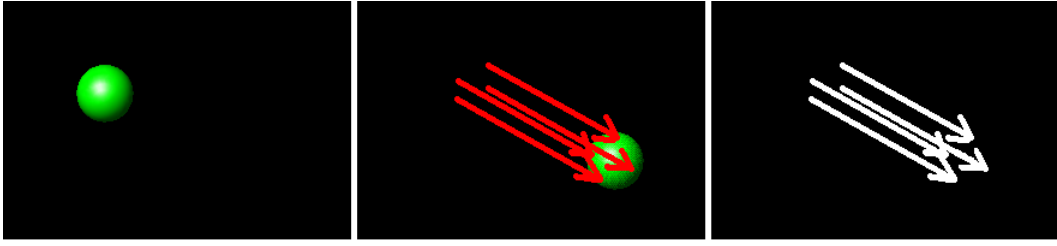


Figure 2.2: Simple example of how movement is manifested from frame to frame.

come into the scene. In addition to the answer regarding the correlation between the two, these databases showed that the saliency maps of different users were consistent among each other [37], which means that a single map was able to reasonably model the average user. It was also possible to notice the tendency of users to gaze at the center and equator of the image to the detriment of other areas, motivating the inclusion of biases to consider these characteristics in their analyses [37] [48].

2.1.1 Temporal features

When it comes to videos, the individual saliency maps very often cannot transmit the full attention users actually have. Movement is a very clear case of an attention attracting mechanism which is not captured in a frame by frame basis. This is the idea behind the computation of temporal features. This gives rise to dynamic models (which include temporal features) as opposed to static ones. Here, we briefly present the idea behind optical flow, which tackles this problem.

The implementation of movement in the prediction of saliency has many challenges compared to the same task without this inclusion. Whereas static models use features like color, intensity, and orientation for their predictions, the dynamic predictions must focus on moving things, as human beings tend to have their attention gravitate towards moving objects [3].

Usually, movement is included via the addition of an optical flow channel. Optical flow describes the displacements that occurred between two consecutive frames in a video sequence. This algorithm is normally implemented in the discrete domain through a vector map. Figure 2.2 shows a simple example of how these vectors are drawn. The pixels of the moving sphere have an associated movement vector that represents the direction and intensity of the movement between two frames. The vector map is the optical flow for the first frame.

The use of models based on DNNs (Deep Neural Networks) is at the state of the art both for static and dynamic models. An example of the usage of this idea can be seen in [31], where the temporal information is extracted through the optical flow between consecutive frames and different ways of using the additional information in the saliency prediction with a DNN are explored, with two output streams, a spatial output and a temporal output, which are later combined. The idea of these two streams comes from the hypothesis of the two types of human perception, which states that the processing of appearance and movement is done separately in the human brain [12]. There are many other ways to include movement, many of which do not

use deep learning..

2.2 Bottom-Up and Top-Down Modeling

Our surroundings give rise to a vast amount of sensory information that is more than our brain can process simultaneously. Selecting the most relevant stimuli in the physical world for processing while filtering out less relevant information allows us to respond quickly to critical environmental changes and achieve behavioral goals more efficiently.

As shown by Borji and Itti [20], attention is commonly categorized into two distinct functions. The *bottom-up* attention is based on the visual characteristics of the scene (stimulus-based), whereas the *top-down* attention is based on a certain objective, with a task in mind, such as scene recognition, expectations, rewards, and current goals.

The regions of interest in a bottom-up approach must be sufficiently different from the surrounding characteristics, for example, a highlighted color. Bottom-up attention is quick, involuntary, and very likely an open-loop system (that is, the output is not processed as feedback to the input) [2]. A simpler example of bottom-up attention is looking at a scene with a single horizontal bar and several vertical bars. Here, the horizontal bar will be more prominent, as its orientation is different from the others. Most models are in this category, but bottom-up features cannot explain well the human fixations, which are mostly based on tasks.

The top-down attention is slow, voluntary, and of closed loop. The key element being the task orientation. Yarbus in 1967 [50] conducted the experiment of tracing the observers' eye movements looking at a scene after asking them to keep a question in mind. The kind of eye movements observed when the question was "How old are the people in the scene?" was distinct from those observed through the question "Estimate the material means of the people in the scene". It is expected that in the first case the observers look more towards people's faces in the scene whereas in the second one they look towards the objects.

In spite of being the most influential mechanism of human attention, top-down approaches are inherently more complex than the bottom-up ones because they depend on a task. This creates a context for the rise of target-driven attention guidance [30], where an object becomes the target and lures the spectator's attention. Furthermore, the combination of top-down and bottom-up approaches in a single algorithm is considered the most realistic way to model human attention, as shown by the example in [5].

2.3 Saliency Prediction Approaches

2.3.1 Heuristic Approaches

There are two basic approaches in the context of generating saliency maps: heuristic, which searches for the specific characteristics in the image to generate the maps according to the human

visual system; and data-driven, which uses real attention data to perform the learning part of the algorithms. The use of these two methods has an interesting historical separation, since with the advent of machine learning, more and more data-driven methods have emerged.

The model of heuristic methods, for example, elements like contrast or surprise are considered [47]. Another element considered heuristic is movement, as seen before. This is because it is well known that movement, in particular, attracts human attention. The heuristic methods are, however, very dependent on understanding how human perception works, an area of psychology still in its infancy, and therefore there are not many methods in use.

Because most of the approaches for salience prediction until 2008 were for 2D methods, many authors tried to adapt already existing methods to the 360° environment. Abreu *et al.*'s method [8] directly applies SALICON [17] to different projected rotations of the images and merges the obtained maps. GBVS360, which is an adaptation of GBVS [13] used for 2D images, applies the GBVS method to various possible viewport images. GBVS was the first saliency method explored in the learning process to this work and an example of it can be shown in Figure 2.3.



Figure 2.3: Example image and its salience as generated by GBVS. The heatmap representation is sometimes used for visual purposes. Here, hotter spots represent where most people will tend to focus their attention when looking at the image.

Below, we present the main heuristic approach considered in this work, which is the BMS360 model described in Lebreton *et al.*'s work in [22].

2.3.1.1 BMS360

In their work, Lebreton and Raake [22] have mainly focused on the inclusion of the equatorial prior (which is a weight matrix that prioritizes the equatorial region) in order to improve the prediction capabilities of the 2D versions of BMS, GBVS and ProSal. In their analysis they concluded that BMS360 is the best performing of these methods, and therefore was the chosen one for our analysis.

BMS (Boolean Map Saliency) [53] was chosen to be adapted to 3D as it is one of the top-performing models in the MIT saliency benchmark database. It is based on the notion of surroundedness, and creates boolean maps ¹ by thresholding feature maps. They assume that

¹Which is a spatial representation that partitions a visual scene into two distinct and complementary regions: the region that is selected and the region that is not selected [16]

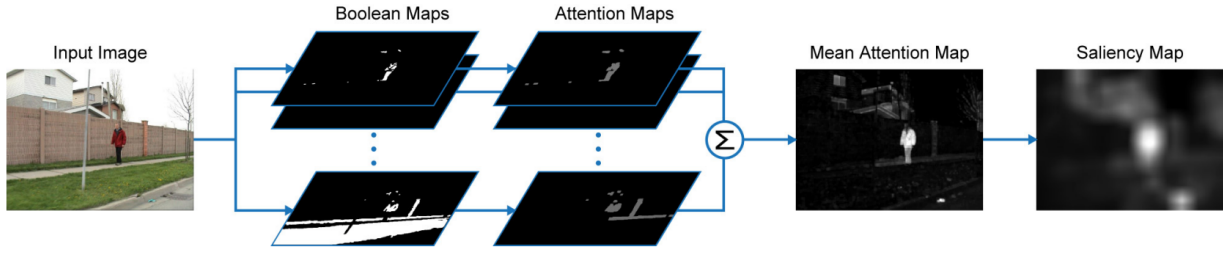


Figure 2.4: Pipeline of BMS from [53]. The boolean maps are computed with the theory presented in [16] and have the attention maps as subproducts. By then summing the maps we get the mean attention map which gives enough information for a saliency map.

Boolean maps in BMS are generated from randomly selected feature channels, and the influence of a Boolean map B on visual attention can be represented by an Attention Map $A(B)$, which highlights regions on B that attract visual attention. Then the saliency is modeled by the mean attention map A_{mean} over randomly generated Boolean map. A_{mean} can be further post-processed to form a final saliency map S for some specific task. The full pipeline is illustrated in Figure 2.4 as taken from Zhang et al.’s paper in [53].

Here are the required adaptations for 360°

- Handling of borders in the equirectangular domain: the original model does not account for spots in contact with the border, which is a problem in 360° as the borders are technically connected. In this case, multifusion saliency (average of saliencies of a succession of horizontal shifts) was applied.
- The problem of oversampling of the polar regions: this is corrected by modeling this distortion as the inverse of the pitch’s cosine and then applying the L2 norm as the regular BMS does.

This model is originally conceived for images, but since static saliency models consider frames as single images we can apply the BMS360 algorithm to every single frame normally. So, the only tweak necessary was using the software Ffmpeg to separate the reference videos into frames using a command like the one below and then performing the saliency evaluation to every frame. We create the frames in the bitmap format as to minimize compression deterioration and have a more accurate saliency map.

BMS360² can currently only run in windows, and as such was the only algorithm run in a desktop. All the other algorithms below were run at GPDS’s server.

2.3.2 Data-driven approaches

Saliency models can also be inferred by training with existing data. With the era of deep learning, many new methods based on neural networks came into place. These novelties originally

²<https://github.com/Telecommunication-Telemedia-Assessment/GBVS360-BMS360-ProSal>

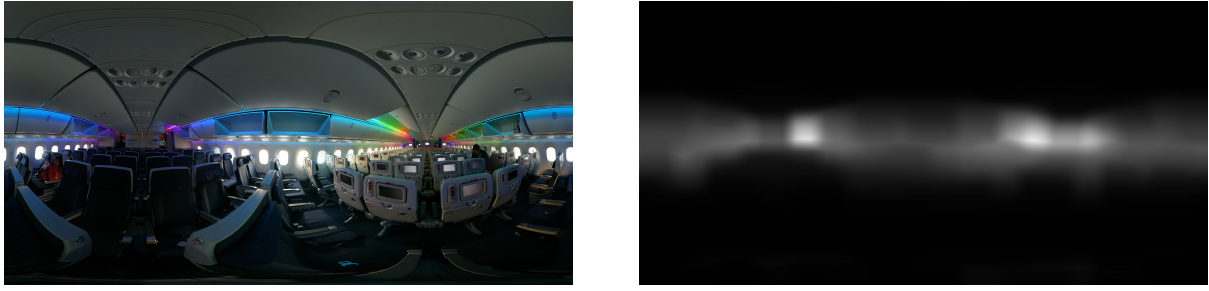


Figure 2.5: Example frame and its saliency as generated by BMS360 on a desktop computer. Notice that this saliency representation is shown in grayscale

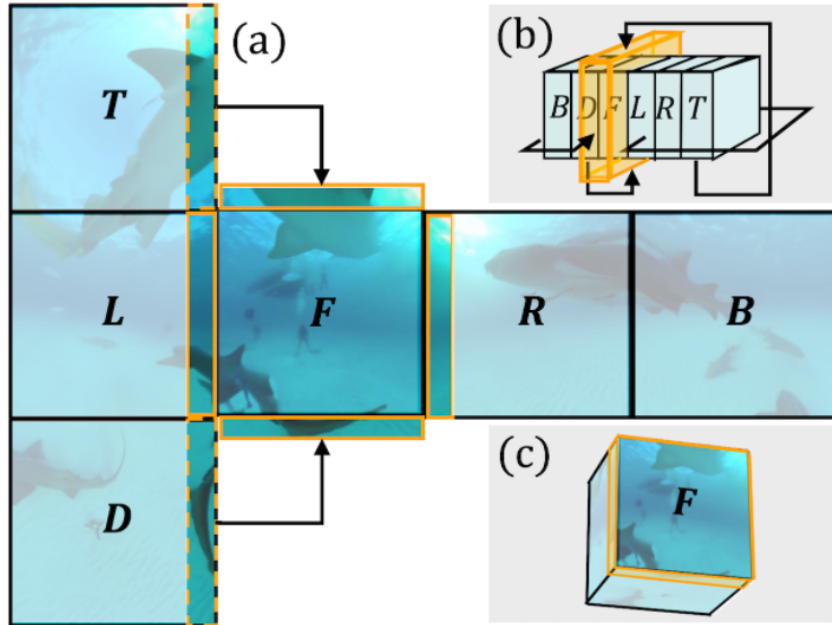


Figure 2.6: Logic behind the idea of Cube Padding. By extending a face to include some of the contents from an adjacent face we can mitigate the discontinuity problem of having six different images per frame to process.

came for 2D analysis, but they were quickly adapted to 3D. Among the various DNNs, we have convolution neural networks (CNNs), generative adversarial networks (GANs), Long Short-Term Memory (LSTM) for videos, and others. Many of these initial approaches were adaptations of the SalNet method for 2D [29]. Data-driven approaches tend to perform better in generalist videos because of their rather uncontrollable nature and are therefore mostly bottom-up [47]. For example, facial recognition is a complicated task in this area, and most CNNs performing this task tend to capture top-down features. However, a lot of recent work has been performed in top-down models, with many innovations coming from this area.

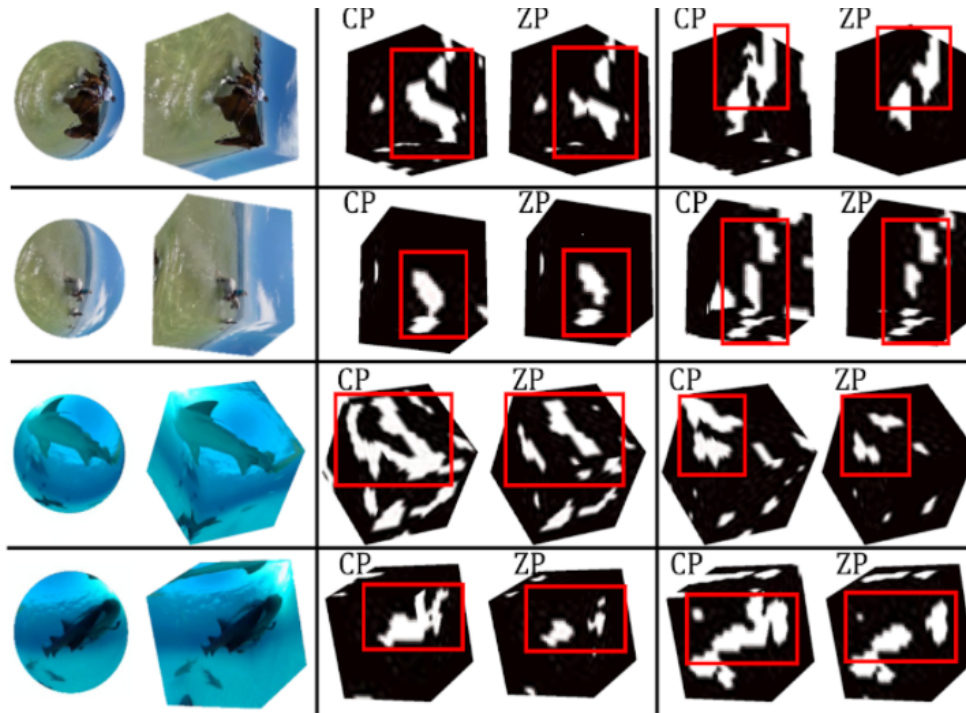


Figure 2.7: Comparison between the usage of zero-padding (ZP) and cube-padding (CP). Note how the strategy Cube Padding allows for a continuous response across the faces, whereas zero padding does not.

2.3.3 Cube Padding

The generation of saliency maps occurs mostly using the equirectangular projection for image processing. But, in Cube Padding [6] the authors claim that this projection generates distortion in the image borders, which makes the saliency extraction difficult. With that in mind, they propose an approach that divides the image into smaller parts. Although it avoids distortion, it will introduce more image boundaries. By dividing the 360° sphere into multiple “overlapping” perspective images they only take the saliency prediction in a center sub-region in order to combine all predictions onto the whole sphere. This process is described in Figure 2.6. However, this requires many more perspective images and significantly slows down the prediction process.

In the Cube Padding algorithm, prediction is based on a DNN ³ and has both a static and a temporal model. The static model predicts the saliency map for each frame in the video, and the temporal model adjusts the output of the static model based on temporal features. The static model obtains the saliency maps in the face of a cube using Convolution Neural Networks (CNNs) ResNet-50 and VGG-16 ⁴, which are well-known pre-trained models. The authors point out that using these CNNs without the aforementioned enlargement of the faces results in an inefficient processing of the borders, and the continuity between the borders would be lost. That is why the

³Deep Neural Network: An Artificial Neural Network is a sequence of neural nodes, usually perceptrons, displayed in layers. A deep network is a network with many layers

⁴ResNet-50 [14] was the winner of the 2015 ImageNet challenge [10], challenge that also saw the emergence of VGG-16. ResNet-50 is the most cited NN in the 21st century

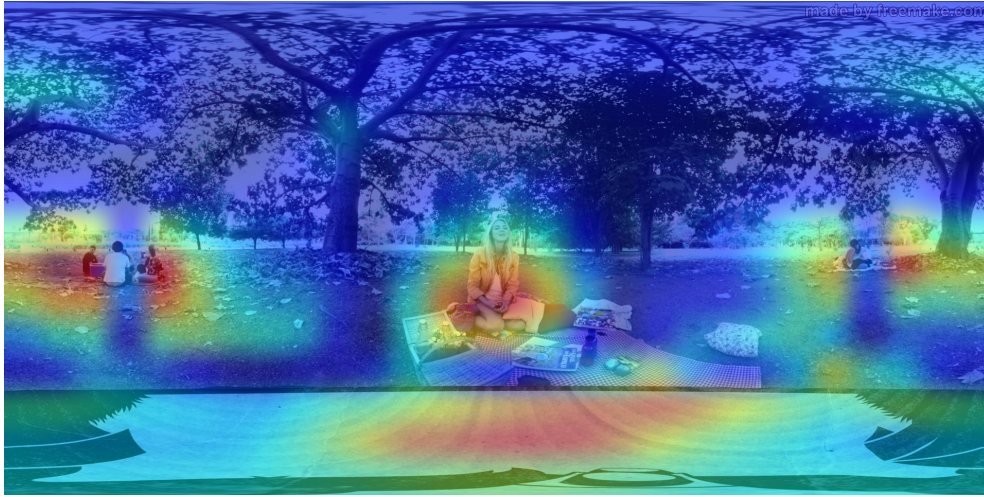


Figure 2.8: Example of a saliency map generated using Cube Padding [6].

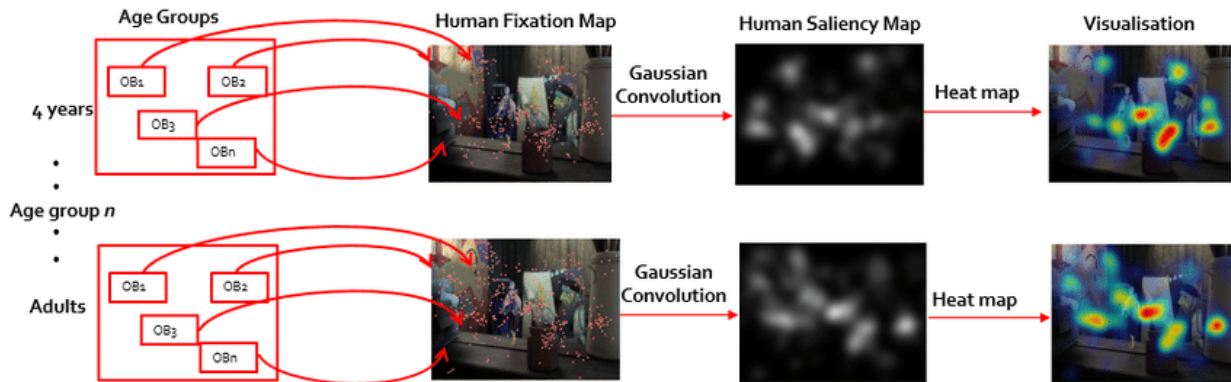


Figure 2.9: Example of how the elaboration of a saliency map is done in experiments with humans.

cubic projection with the chosen padding works, as shown in Figure 2.7, which shows how the cube padding is superior to the zero-padding (the absence of incorporation of the content of adjacent faces). In the black and white hexagons we see the frames centered around a vertex of the cube, where in all of the cases shown an object happened to be there. In the highlighted boxes we can see how Cube Padding manages to keep the object mostly together whereas zero-padding separates them. This is especially noticeable in the case of the sharks. Figure 2.8 shows an example output frame. The hotter regions are the most visually attractive and they highlight the people in the scene as expected.

For the temporal model, LSTMs are used, as 360° videos can be considered a sequence of 360° frames. Training was weakly supervised as the method does not use pixel annotation or bounding boxes for labeling. The reasoning behind this method is to allow the method to be scalable. In their article, they also propose the Wild-360 dataset, which contains challenging videos with their respective saliency map annotations. The method showed good performance compared to similar methods. Performance was evaluated using the AUC-Judd, AUC-Borji, and Pearson correlation coefficient.

2.4 Saliency Precision

To have a base for performance comparison for a certain saliency model, subjective experiments must be conducted on the image or video in question. In these experiments, the eye-tracker annotates fixation sequences of many individuals and amalgamates them in a ground-truth map. With that, many datasets could emerge with coordinates of the fixations of these experiments [11]. It is also common to perform an operation called foveation, where a circular function (e.g. a Gaussian function) is convoluted with each fixation point. This is usually done because the human visual system does not focus on only one pixel at a time, but rather on a small region centered around this pixel. The entire process of elaborating a subjective saliency map is shown in Figure 2.9. Normally, various age groups are chosen, their fixations annotated, and after an aggregation process the saliency map is created. Here, the map in gray scale is the saliency map, also represented as a heatmap in a more intuitive way of seeing. This way, the task of precision evaluation of any new saliency model is reduced to comparing the saliency map of an experiment with the one generated by the method using classical comparison methods. Below we present an introduction to the most common metrics in the literature.

Chapter 3

Quality of Experience

In this chapter we go through relevant topics to understand Visual Quality Assessment (VQA), starting with what deteriorates videos, then going through methods for assessing VQA and finally talking about how to compute the correlation between methods.

Visual Quality Assessment (VQA) is the most relevant measurement in the scope of transmission performance for both 2D and 3D videos. VQA is an inherently subjective concept as it depends on the perception of those who watch the video. There is a lot of research for the area of traditional videos and there are unique challenges for 360° videos.

In Figure 3.1 we see the overall pipeline for 360-degree video processing [9]. We see certain elements already seen before in Chapter 2, such as the projection in encoding and decoding, but also some new elements of the pipeline we have not seen before. We can basically describe the whole process (with their respective possible distortion problems) in 4 steps:

1. **Acquisition:** Here, we deal with how the video is produced. Production of such videos normally uses 360-degree cameras with multi-sensor systems. These systems can be modeled as central cameras that project in the 3D space to a point on a spherical imaging surface. In practice, the omnidirectional output signal is the result of a “stitching” algorithm, which merges the overlapping field-of-view signals acquired by all sensors to produce a panorama media. In the case of video, additional video synchronization and stabilization may be needed. After acquisition, the signal is usually stored in Equirectangular format (as seen in section 1.1). Normal distortions caused at this level come from limitations of the acquisition device and also faulty stitching.
2. **Encoding:** The goal of the encoding step is to reduce the redundancy in the signal so that it can be better stored and transmitted. Most of the current 360-degree video systems reuse the same encoding tools as classical video solutions, such as H.264, H.265, VP9 or AV1. One of the main challenges then lies in mapping the content into rectangular frames that are the

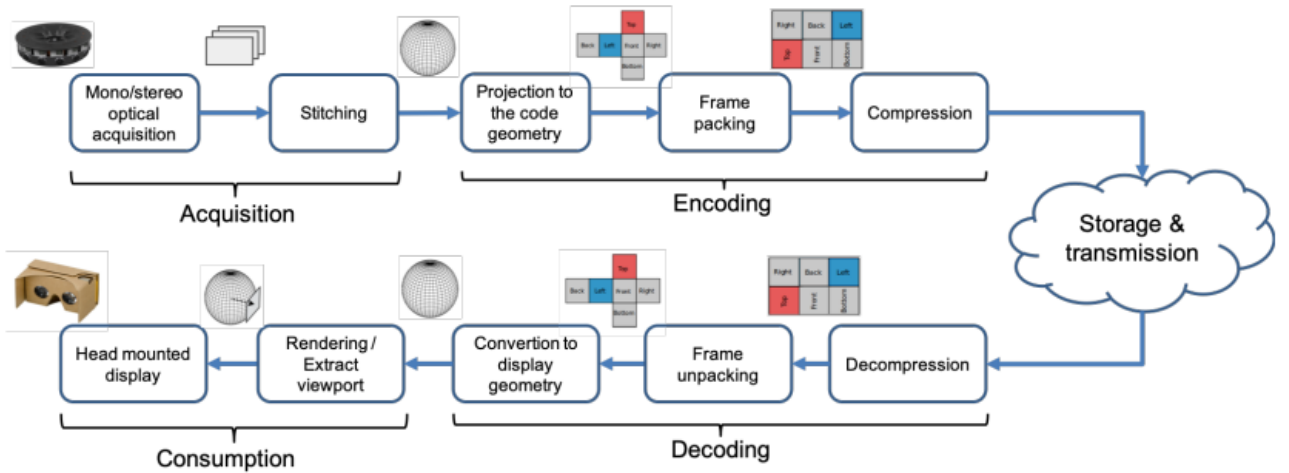


Figure 3.1: End-to-end 360-degree video processing pipeline

input of these video encoders. Frame Packing refers to the combination of two frames, one for the left eye and the other for the right eye, into a single “packed” frame that consists of these two individual sub-frames. In 360° videos, warping due to the projection type also exist, in addition to conventional artifacts, such as blurring, color bleeding, blocking, and ringing; .

3. **Transmission:** The biggest problem here is the high fidelity necessary for 360 degree videos (usually at least 4K) and the ensuing large bandwidth necessary. To reduce quality demands, viewport-dependent projection and tile-based approaches have come into play. Transmission delays and communication losses affect the streaming of omnidirectional video sequences, similar to how they affect traditional videos. When not considering the viewport, the user may perceive typical streaming distortions such as delay, rebuffering, events, and quality fluctuation.
4. **Consumption:** At the client end, all the processes mentioned have to be inverted. Here, an inverse mapping from the plane to the sphere has to be performed. Users usually watch on head-mounted displays or more recently on a computer or smartphone. The main distortions we face in this case are related to the capabilities of the display, such as aliasing, blurring, motion blur, etc.

Visual distortions that occur in images and videos captured by perspective cameras and undergoing compression and transmission have been largely characterized and analyzed in the literature, both for standard 2D and stereoscopic 3D signals [9]. However, new types of distortion can occur in 360-degree visual signal dataflows.

It is worth pointing out that, although for 2D videos a certain degree of distortion can be accepted, in 3D environments the same level of distortion may not be acceptable. The exact impact of distortions on perceived quality and immersive experience is still unknown and is therefore an open research topic.

In this work, we mainly focus on the effects present in encoding compression, in particular, the quantization parameter (QP) and its influence on the quality of the video.

Encoding Compression: Quantization Parameter (QP) and BitRate Trade-off

The Quantization Parameter controls the amount of compression for every macroblock in a frame. Large values mean that there will be higher quantization, more compression, and lower quality. Lower values mean the opposite. QP ranges from 0 to 51 in H.264 encoding. BitRate refers to the bits per second consumed by a sequence of pictures, that is, $\text{bitrate} = (\text{average bits per frame}) \times (\text{frames per second})$. In practice, it is equated to the reliable network bandwidth that is provisioned or available for the stream.

Block-based hybrid video encoding schemes, such as MPEG and H.264/H.265 are inherently lossy processes. They achieve compression not only by removing truly redundant information from the bitstream but also by making small quality compromises in ways that are intended to be minimally perceptible. In particular, the quantization parameter QP regulates how much data can be saved. When QP is very small, almost all the details are kept. As QP increases, some details are aggregated and the bit rate drops, but at the price of a distortion increase and some loss of quality. Therefore, we arrive at a trade-off where if you want to lower the bit rate, you can do so by increasing the QP at the cost of increased distortion. The increase of the spatial activity¹ can allow for a better trade-off, where you have to sacrifice less of the bitrate to get a better QP, and vice-versa [38].

3.1 Subjective VQA methods

In order to test any VQA prediction algorithm, extensive research has to be done with subjects in order to assess their actual opinions. This research is done by inviting large numbers of people, usually from the university where the research is conducted, but also community members, mostly those associated with the researchers. The research also has a set of requirements that it has to follow, for example those from ITU-T and MPEG. Experiments which aim to evaluate only the visual quality elements of the video will usually involve a swivel chair with enough room for 360° rotation. In cases where immersiveness and interaction are also evaluated, enough room for physical exploration has to be allowed. There are certain situations where 360° treadmills can be incorporated as well. Participants wear HMDs like Oculus Quest or some other similar apparatus. Figure 3.2 shows some of the other popular HMDs used in experiments.

In these experiments, individuals have a scheduled time to watch the sequence of videos. Figure 3.3 shows one such volunteer in another experiment performed in University of Brasilia. If each video is shown without another video to compare to, the methodology is called Single Stimulus

¹This term is used to quantify the amount of spatial variation within a part of the picture, normally a block of N pixels. In other words the spatial activity is the sample variance of a block's values. It is the measure for local complexity used in MPEG-2.



Figure 3.2: Examples of different popular HMD models.



Figure 3.3: Subject participating in an experiment wearing an Oculus device during the pandemic. Notice the lack of physical obstacles near the chair.

Impairment Scale (SSIS), whereas if we have a video for comparison, we talk rather of a Double Stimulus Impairment Scale (DSIS).

Normally each video lasts just a few seconds and is followed by an evaluation moment where subjects are asked their opinion on a predefined numerical scale. This is the raw score associated to each person. By averaging the raw scores of all the participants we arrive at the Mean Opinion Score, or MOS for short. The MOS score is calculated using the formula in Equation 3.1:

$$MOS_j = \frac{1}{I_j} \sum_{i=1}^{I_j} S_{ij} \quad (3.1)$$

where S_{ij} is the raw score that subject i assigns to sequence j and I_j is the number of valid subjects viewing sequence j .

Depending on the number of video sequences, experiments can last hours for each participant, with pauses every couple of minutes in order to avoid eye soreness and confusion and to have the cleanest results possible.

As we can see, the research on the topic is relatively difficult as they depend on the correct equipment and a large body of individuals willing to participate. Recently, due to the Covid-19 pandemic, conditions were further exacerbated and new experiments also depend on robust sanitary protocols. Because of that, having a dataset with the ground truths for this research of University of Brasilia’s authoring was particularly challenging, and Li et al.’s [24] dataset was chosen instead.

3.2 Objective VQA Methods

Objective VQA methods attempt to mathematically estimate the score a user would give after watching a video. Extensive research has been done in images and 2D videos has been done, as shown in [46]. The methods have evolved from simple computations of errors, like the Mean Squared Error, to very robust methods based on deep learning. These metrics can be calculated in a multitude of configurations, and below we present some of the most influential and relevant to our work.

In our work, we explore only Full-Reference metrics, that is, metrics which take into account the original signal, but we also have other types of metrics. The No-Reference Quality Assessment Metrics are an ensemble of metrics that do not need the original signal to estimate the VQA. They compute the quality score based on statistics of the expected image. In fact, they possess an inferior evaluation capability compared to Full-Reference metrics and constitute a minority in the field of VQA. They have the advantage of needing less bandwidth to perform their measurements and do not need to be temporally aligned with the original signal. As examples of No-Reference metrics, we can cite BRISQUE (Blind/Referenceless Image Quality Evaluator) and NIQE (Naturalness Image Quality Evaluator). Both algorithms train a model using identical and predictable statistical features called Natural Scene Statistics (NSS, not to be confused with Natural Scanpath Saliency). NSS are based on the luminance coefficients normalized in the spatial domain and are modeled as

a multidimensional Gaussian distribution. The BRISQUE model is trained in an image database with known distortions, and the method evaluates the quality of the images with the same kind of distortion. BRISQUE accounts for the users' opinions and is therefore opinion-aware, which means that the subjective quality scores have to come along with the training images. The NIQE model is trained with non-distorted images, but it is capable of evaluating the image quality with any distortion. NIQE is opinion-unaware and therefore does not correlate as well as BRISQUE with the human perception.

Reduced Reference Metrics, on the other hand, strive for a middle ground between the restrictions applied to Full Reference and the difficult modeling of No Reference metrics. In it, only some characteristics of the original video are transmitted for evaluation. These metrics present some advantages and disadvantages in each extreme and are performed according to each application case.

3.2.1 Full-Reference Quality Assessment Metrics

The Full-Reference metrics are the ones which need a reference signal for comparison. The availability of the original signal is not always a given, which implies that the application of these metrics is restricted. Because of that, these kinds of metrics work better in an offline quality context. In this work, we explore only full-reference metrics. In a general sense, 360 quality metrics are adapted from their traditional 2D counterparts, in the same way that video metrics are adaptations of the same metrics for images. Some of them are based purely on the data, while others take into account properties of the human visual system.

In order to account for the non-uniform sampling density from the sphere to the plane, a problem which does not occur in 2D, spherical quality metrics were created [52]. This motivates the usage of pixel weights according to their position. Furthermore, in 360°, only the FoV and its periphery are seen at any given moment [24] and the distortion of the salient regions starts to hamper the visual quality. Therefore, there is a division of 360° metrics by type, one attempting to solve the sampling problem and the other to incorporate human perception.

The two most well-known ways of evaluating objective VQA for images are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [15]. Both manipulate the difference between received and reference frame pixels, and then do the average between the two. For 360° videos, a second average between all frames can also be implemented. Let the Mean Squared Error (MSE) be defined as:

$$MSE = \frac{\sum_{M,N}[x(m,n) - y(m,n)]^2}{MN}, \quad (3.2)$$

where x represents the reference image and y the distorted image, m and n are the spatial coordinates.

With this error at hand, we can calculate PSNR:

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right), \quad (3.3)$$

where R is the most intense pixel value in the reference image (in 8-bit images, $R = 255$). PSNR is extremely popular in the signal processing area, but because there are too many parameters that highly influence the PSNR value and barely affect the visual quality its correlation with actual data tends to be worse as there is no particular pixel manipulation (like bias or weight attribution). Therefore, it usually does not reflect well the real VQA value.

In 2004, Wang *et al.* [43] proposed the Structured Similarity Index (SSIM), which uses the fact that human perception is highly adapted to extract structural information from the scene. This means that the pixels have a high dependence, especially when they are close in space, and this dependence has information about the structure of objects in the scene. The idea of evaluating the structural information change and that it can be a good approximation of the perceived distortion in the image.

SSIM evaluates the visual impact of three features in the image: luminance, contrast, and structure and is computed as follows:

$$SSIM(x, y) = [l(x, y)]^\alpha [c(x, y)]^\beta [s(x, y)]^\gamma, \quad (3.4)$$

where x and y are the reference and distorted images, respectively, and $l(x, y)$, $c(x, y)$ and $s(x, y)$ are the luminance, contrast, and structure features, respectively. We can also write expressions for $l(x, y)$, $c(x, y)$, and $s(x, y)$ as:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad (3.5)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad (3.6)$$

and

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \quad (3.7)$$

Here, μ_x , μ_y , σ_x , σ_y and σ_{xy} are the local means, standard variation, and cross covariance, respectively, for the images x and y . As a standard, $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$. With these constants, the expression in equation 3.4 is simplified to:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (3.8)$$

For VQA analysis, it is more interesting to apply the SSIM index locally than globally. In [43], an 11x11 Gaussian kernel window is used, but frameworks such as ffmpeg use 8x8 windows for SSIM moving pixel by pixel in the image. This approach can be problematic, as it can create block artifacts. There are other ways to define the window, but the final idea is the same: averaging the SSIM indexes for each of the windows to obtain a final index called Mean-SSIM (MSSIM), as shown in the following equation:

$$MSSIM = \frac{1}{M} \sum_{j=1}^M SSIM(x_j, y_j). \quad (3.9)$$

Figure 3.4 from Wang's work in [43] shows how SSIM can capture differences that PSNR cannot. In all of them, $MSE = 210$, but, as we can see, the images in the second row have a lower SSIM index, indicating less structural similarity. This is much closer to what we can actually see.

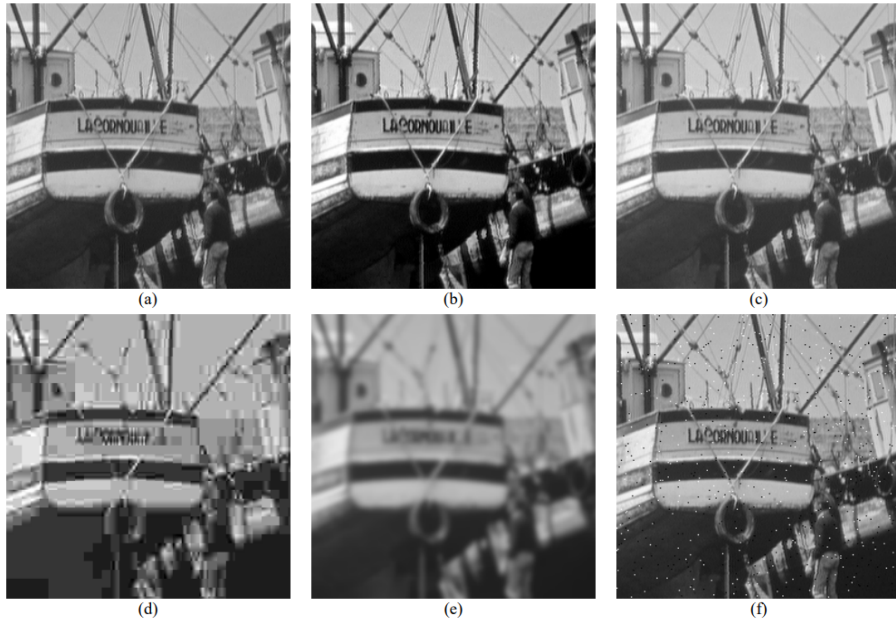


Figure 3.4: Comparison of images with same MSE (a) Original Image; (b) More contrast, MSSIM = 0.9168; (c) Displaced average MSSIM = 0.9900; (d) JPEG Compressed MSSIM = 0.6949; (e) Blurred Image MSSIM = 0.7052; (f) Salt and Pepper noise MSSIM = 0.7748

Wang *et al.* also proposed multiscale SSIM (MS-SSIM), which combined the SSIM of various versions of the image, on different scales and in a multistep subsampling process. MS-SSIM can be more robust compared to SSIM when it comes to variation in visual conditions. There are other variants based on SSIM, such as multicomponent SSIM [23] and complex wavelet SSIM [42].

The adaptation of 2D metrics for 3D is natural here in the same manner as was for saliency prediction methods. To solve the sampling density problem, one solution would be to project the 360° image/video to another domain with uniform sampling. Yu *et al.* [51] have proposed the S-PSNR, where the PSNR is calculated in the spherical domain. Specifically, the expression for PSNR in Equation 3.3 has values of m and n replaced by a set of points uniformly distributed in the spherical domain. $M \times N$, the number of original pixels, is replaced by the number of these new points N_{S-PSNR} . For each point location p in the sphere, its value is obtained in the corresponding location of the pixel i of the original projection through nearest-neighbor search (so-called S-PSNR-NN) or bilinear interpolation methods (S-PSNR-I). The main problem with this adaptation and others that have arisen is the high complexity of these approaches. Taking that into account, many approaches consider weights (therefore, with initial W in their abbreviations) according to the pixel locations. From this we have WS-PSNR and AW-PSNR (area-weighted PSNR) [54]. If we consider the same parameters as in equations 3.2 and 3.3, with the addition of a matrix weight $W(m, n)$, equation 3.10 shows the equation for computing WS-PSNR.

$$WSPSNR = 10 \log \left(\frac{R^2}{WMSE} \right) \quad (3.10)$$

where

$$WMSE = \frac{1}{\sum_{M,N} W(m,n)} \sum_{M,N} (x(m,n) - y(m,n))^2 W(m,n) \quad (3.11)$$

Just like PSNR adaptations, SSIM was adapted to take into account the peculiarities of the 360° images/videos. Methods with weight assignment (W-SSIM and WMS-SSIM [33]) and methods that calculate the similarity between windows in each pixel (Spherical SSIM [4]) were proposed. In the case of the last one, in order to calculate similarity between windows in each pixel location, small images centered around the viewport in the spherical location corresponding to the pixel are extracted both for the reference and distorted videos. The similarity is then calculated between the two extracted viewport images. S-SSIM also has weight allocation. For videos, it is possible to use a single similarity metric that joins all the indexes of all the frames. It is called Video SSIM or VSSIM [44]:

$$Q_i = \frac{\sum_{j=1}^{R_S} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_S} w_{ij}}, \quad (3.12)$$

where Q_i is the quality index for the i -th frame, w_{ij} is the weight attributed (in the case of W-SSIM) and R_S is the number of sampling frames. We then calculate $VSSIM$ for a video with N frames as:

$$VSSIM = \frac{\sum_{i=1}^N W_i Q_i}{\sum_{i=1}^N W_i}, \quad (3.13)$$

where W_i the weight for the i -th frame based on the global movement and in w_{ij} .

To incorporate perception in a 2D video or image, the simplest method is using the saliency map as a set of weights attributed to the aforementioned VQA metrics. This idea was adapted to 360° in the works of Xu *et al.* [45]. In their work, they have also explored a convolution by the viewport region to generate a non-content-based weight map. The idea of incorporating saliency maps into the VQA metrics is the main topic of discussion in Chapter 4. There are also other models attempting to extract the perception characteristics directly, such as pixel-level characteristics, superpixel, and semantic segmentation, for example as in Yang *et al.*'s work [49]. Currently, many models based on machine learning have been proposed, as in [46].

3.2.1.1 VMAF

VMAF is an Emmy-winning perceptual video quality assessment algorithm developed by Netflix. This software package includes a stand-alone C library `libvmaf` and its wrapping Python library. The Python library also provides a set of tools that allows a user to train and test a custom VMAF model [26]. It is a full reference perceptual video quality assessment model that combines quality-aware features to predict perceptual quality. VMAF combines human vision modeling with machine learning, offering a good prediction of video QoE. The development of VMAF started between Netflix and Professor C.C. Jay Kuo from the University of Southern California. In June 2016, VMAF was first open sourced on GitHub (<https://github.com/Netflix/vmaf>). VMAF uses existing image quality metrics to predict video quality, such as Visual Information Fidelity (VIF) or Detail Loss Metric (DLM). These features are combined using a supervised learning regression

model to provide a single output result, called the VMAF score. This score ranges from 0 to 100 per video frame, with 100 being the quality of a video identical to the reference.

An early version of VMAF has been shown to outperform other image and video quality metrics such as SSIM, PSNR-HVS and VQM-VFD on three of four datasets in terms of prediction accuracy, when compared to subjective ratings. Its performance has also been analyzed in another paper, which found that VMAF did not perform better than SSIM and MS-SSIM on a video dataset [1]. In 2017, engineers from RealNetworks reported good reproducibility of Netflix’ performance findings [35]. In MSU video quality metrics benchmark, where its various versions (including VMAF NEG) were tested, VMAF outperformed all other metrics on all compression standards.

It can be shown that VMAF is also a reasonable metric when working in 3D environments, as shown in [28]. As such, it presents a very robust evaluation for our research and is in line with the state-of-the-art in the field.

3.3 Performance Evaluation

The same way as for 2D, performance evaluation in 360° videos is done calculating the correlation and error between the subjective values and the objective prediction. There are then many metrics capable of assessing the result: PCC (Pearson Linear Correlation Coefficient), Spearman Correlation Coefficient, and Kendall Correlation Coefficient (Kendall’s τ), the root-mean squared error (RMSE) and the Mean Absolute Error (MAE). Unlike the saliency case, where there are many databases with available human eye fixation positions, for VQA there are not many open databases. This means that comparing solutions is difficult since they are not evaluated on the same data.

For the examples of evaluation metrics below, it is possible to consider that one of the vectors is MOS and the other vector is the objective VQA according to the chosen methods.

- **Pearson Correlation Coefficient (PCC)** - Most common correlation metric. For a column X_a in matrix $[X]$ and a column Y_b in matrix $[Y]$, with means $\bar{X}_a = \sum_{i=1}^n (X_{a,i})/n$ and $\bar{Y}_b = \sum_{j=1}^n (Y_{b,j})/n$, we calculate the PCC as follows:

$$\rho(a, b) = \frac{\sum_{i=1}^n (X_{a,i} - \bar{X}_a)(Y_{b,i} - \bar{Y}_b)}{\sqrt{\sum_{i=1}^n (X_{a,i} - \bar{X}_a)^2 \sum_{j=1}^n (Y_{b,j} - \bar{Y}_b)^2}}, \quad (3.14)$$

where n is the size of each column. Correlation values are contained in the interval $[-1, 1]$. -1 indicates a perfect negative correlation, while $+1$ indicates a perfect positive correlation. 0 indicates the absence of correlation. PCC, like many other conventionally used coefficients, is not robust to outliers. It is also not capable of asserting any nonlinear relationship between variables.

- **Spearman’s Rank Correlation Coefficient (SCC)** - This coefficient is equal to PCC applied to the matrix rank of columns X_a and Y_b . If all ranks in each column are distinct,

the equation that defines SCC is the following:

$$\rho(a, b) = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}, \quad (3.15)$$

where d is the difference between the ranks of the two columns and n is the length of the column. The Spearman coefficient also varies within the range $[-1, 1]$, with its sign indicating the direction of a growth trend of a variable in relation to the other. Therefore, the two perfect correlations (-1 e 1) mean that Y is a monotonous function of X . That is, in practice, it brings more information than PCC.

While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

- **Kendall's Rank Correlation Coefficient** - It is based on counting the number of pairs (i, j) , where $i < j$ are concordant. That is, for which (i, j) we have $X_{a,i} - X_{a,j}$ and $Y_{b,i} - Y_{b,j}$ of the same sign. For a column X_a in matrix $[X]$ and column Y_b in matrix $[Y]$, the Kendall coefficient τ is defined as:

$$\tau = \frac{2K}{n(n-1)}, \quad (3.16)$$

where $K = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \xi^*(X_{a,i}, X_{a,j}, Y_{b,i}, Y_{b,j})$, and

$$\xi^*(X_{a,i}, X_{a,j}, Y_{b,i}, Y_{b,j}) = \begin{cases} 1, & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) > 0 \\ 0, & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) = 0 \\ -1, & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) < 0. \end{cases} \quad (3.17)$$

The correlation values can vary between -1 and 1 . The value -1 indicates that the rank of a column is the inverse of the other, whereas the value $+1$ indicates that both ranks are the same. The value 0 indicates the absence of relation between the columns.

Chapter 4

Methods and Results

In this chapter we present the main ideas of the research as well as the means to achieve them. We join the theoretical basis shown in the previous two chapters in order to propose a way to answer the question: Does knowing facts about spectators' attentions make our prediction of their experience more accurate? After that, we present the relevant results from our evaluation.

4.1 The Source Material: The VQA-ODV dataset

Called Visual Quality Assessment for OmniDirectional Videos (VQA-ODV), Li and his team [24] collected data for 60 reference videos and 540 impaired sequences, providing not only the MOS scores but also HM and EM data. Furthermore, they developed a deep learning model embedding HM and EM for objective VQA (as discussed in the previous chapter, this is one of the new models on an objective metrics based on deep-learning). With all that, this research provides a goldmine for experimentation on saliency (head and eye movement based) and quality so their videos are the ones chosen for our methodology. The only problem we have for this dataset is the sheer size of the videos, the biggest of which surpassing 12GB (therefore, any processing on a personal computer is ill-advised).

Returning to the dataset, the 60 reference videos are divided into 10 groups of 6 videos each and range in resolution from 4K (3840x1920 pixels) to 8K (7680x3840 pixels). The impaired sequences vary in compression with different Quantization Parameters (QP = 27,37 and 42) and different map projections: Equirectangular (ERP), Reshaped Cubemap Projection (RCMP) and Truncated Square Pyramid Projection (TSP).

In the experiment, HTC Vive was used as an HMD and the eye-tracking module aGlasse DKI was embedded to the HMD. A GUI for control and programs for HM and EM capturing were developed. The sessions were divided into training and then testing sessions, where at first subjects

were shown with 1 video and 9 impaired sequences how they should approach the experiment and after a break they were shown 6 reference videos and their corresponding 54 impaired sequences. Videos were displayed at random and therefore are characterized as single stimulus.

There were 221 people participating in this experiment, 143 males and 78 females, with ages ranging from 19 to 35. the subjects were divided in 10 groups as to match the grouping of videos and so that each subject would watch just one group of videos. The dataset provides both individual raw scores and DMOS measurements for the experiment. MOS is then calculated as in Equation 3.1.

To simplify our analysis, we tried to choose two videos from each one of the groups so to contemplate the largest number of people and resolutions possible while still being able to perform analysis in a reasonable amount of time. We also only consider the influence of QP as impairment, opting therefore to discard any non-ERP video projection. With that sampling performed, we perform the analysis below on 20 reference videos and 60 distorted sequences.

4.2 Chosen Saliencies

Our goal with the saliency maps is using incorporating them in our video frames as to bias our quality of experience metrics towards considering more of what the user is actually paying attention.

In order to assert exactly which kind of saliency was the best performer in our analysis, we went on to choose the most different ones possible, with the restriction that they be specific to 360° videos. We originally wanted to analyse a heuristic, a static data-driven and temporal data-driven model. After considering the processing time for the temporal model, we have ultimately decided to discard it.

Since we can consider that the streamer has access to the reference video and can evaluate saliency beforehand, the most logical approach is extracting the saliency maps from the original video and applying them to both original and distorted videos. Of course, the mathematical applicability of this saliency map incorporation has to be evaluated and it will be seen in a case-by-case situation in Section 4.3. But, for now, we can keep in mind this idea.

Along with BMS360, Cube Padding is one of the chosen models for evaluation in this work. The original idea was to use it as a temporal model to contrast with the other two which are static. But, this model is particularly slow. The processing of a 3 minute video with Cube Padding takes 4 hours in the static model, so the temporal one takes even longer and was therefore not considered an object of this work.

Cube Padding is a video algorithm and its inputs are mp4 videos. So, just by providing the video we are able to extract the saliency already. Its outputs are the individual saliency frames. Due to its complexity and because it can be run on Linux, this algorithm was run remotely in our server. Originally its outputs were heatmaps which were overlaid to the video frames, but with minor tweaks we were able to get grayscale saliency maps, which are the saliency format for our

analysis as explained in Section 4.3. BMS360 is originally a 360-degree image saliency metric, so in order to run it on a video, we first have to extract the individual frames from the video, then we perform the frame-wise saliency evaluation, and finally the output is a similar frame-wise saliency map.

It is important to note that Cube Padding has specific dependencies that makes it hard to run for the unaware user. Because of that there has been an effort of our team in expanding the instructions present in their GitHub page¹, in particular for our modifications.

For this work, many parallel resources were used:

- As processing units a personal computer with a Windows 10, Intel(R) Core(TM) i5-5200U CPU at 2.20GHz processor with 4,00 GB installed RAM was used as well as GPDS server with an AMD Ryzen threadripper 3990X, 64 cores, 3Ghz, 128GB ram ddr4. Processing in the personal computer was much slower than in the server.
- The algorithms used are described in their respective sections. For the saliency map models, the original cubepadding without the overlay function was used, and for the BMS360 model we used the aforementioned video adaptation. This adaptation was run with the help of the Ffmpeg application through the PC terminal with a Python3 code by the author
- The saliency incorporation algorithms were originally run and tested with Google Colab using the Python3 language and later implemented on the PC and server. This is also an algorithm written by the author
- The BMS360 adaptation for videos was run on windows on the personal computer whereas all other algorithms were tested on the personal computer and run remotely on the server. The task of transferring files between PC and server was facilitated with the user of the FileZilla application (which allows for the transferring of files using FTP and encrypted FTP such as FTPS (server and client) and SFTP). The server containers were managed through Portainer, which is a powerful, GUI-based Container-as-a-Service solution that helps organizations manage and deploy cloud-native applications easily and securely.

4.3 Incorporating saliency and the video size problem

After we run the saliency models we have to find a way to incorporate the saliency maps onto the videos without incurring in a mathematical incoherence. The simplest incoherence we can think of is one where the sum of all weights is not normalized and so it invalidates the comparison scale. In order to proceed with this incorporation, we look back at the literature to see how weights and biases are usually included in the models. Looking back at the models WS-PSNR and WS-SSIM, which manage to incorporate weights into the video in a way that does not invalidate their mathematical logic, we find inspiration for this task.

¹<http://aliensunmin.github.io/project/360saliency/>

For example, let us consider the formulation below for WS-PSNR using the conventions from Equations 3.10 and 3.11. The corresponding saliency-weighted metrics are:

$$\text{WS-PSNR}(x, y) = 10 \log_{10} \left(\frac{R^2}{\text{WS-MSE}} \right) \quad (4.1)$$

where

$$\text{WS-MSE} = \sum_{M,N} \frac{W \cdot |\text{Error}(m, n)|^2}{MN} \quad (4.2)$$

and

$$\text{Error}(m, n) = x(m, n) - y(m, n) \quad (4.3)$$

Here, x and y are the original and distorted images, and M and N are the width and height in pixels of the frame, and x and y the reference and deteriorated images respectively. By viewing our saliencies as a normalized weight matrix S (i.e., by dividing the complete saliency map by 255 so that the pixels are confined between 0 and 1), we can make use of them directly as weights, that is, $W = S$. We can incorporate the weight matrix S into the error by taking its square root and distributing it on the reference and impaired images. This procedure is seen below, which takes the expression inside the sum in Equation 4.2 and distributes the weight matrix.

$$\text{WS-MSE} = \sum_{M,N} \frac{W \cdot |\text{Error}(m, n)|^2}{MN}, W = S \quad (4.4)$$

$$\text{WS-MSE} = \sum_{M,N} \frac{S \cdot |\text{Error}(m, n)|^2}{MN}, S \geq 0$$

$$\text{WS-MSE} = \sum_{M,N} \frac{|\sqrt{S} \cdot \text{Error}(m, n)|^2}{MN}$$

As the saliency pixel values are not negative, $|S| = S$. This new error is given by the following equation:

$$\text{Error}_{sal}(m, n) = \sqrt{S}x(m, n) - \sqrt{S}y(m, n). \quad (4.5)$$

Therefore, we show that a saliency map inclusion similar to WS-PSNR works. In practice, for every video, we first divide it into its individual frames. Then we run the corresponding saliency method and, when incorporating the saliency, we proceed with the pseudo-code in Algorithm 1 using an Ffmpeg writer object.

A brief note about the `ffmpeg` writer object. It is accessed through the `skvideo` library and it is a very powerful tool to avoid memory overload due to the processing of several frames at a time. By using the attribute `writeFrame`, it sends `ndarray` frames to Ffmpeg software, which is run in the command line. By successively using this method we can create a video from the individual frames, and by using the attribute `close` we can save the video.

In the sequence of this work, we will work with a fixed framework and, therefore, consider the inputs $\sqrt{S}x(m, n)$ and $\sqrt{S}y(m, n)$ to calculate three main metrics: PSNR, MS-SSIM, and VMAF. Although it can be argued that this type of inclusion may not be valid for the latter two metrics,

Algorithm 1 Saliency Computation and Inclusion Steps

```
1: for Video in Videos List do
2:   if Method = BMS360 then
3:     Perform frame separation
4:     Evaluate frame-wise saliency evaluation using BMS360 code
5:     frames  $\leftarrow$  BMS360 frame-wise saliency maps
6:   end if
7:   if Method = Cubepadding then
8:     Perform a lossless MP4 video conversion (to avoid further deterioration)
9:     Run Cube Padding algorithm
10:    frames  $\leftarrow$  Cube Padding frame-wise saliency maps
11:  end if
12:  for Frame in frames do
13:    Extract corresponding frame from Video
14:    Incorporate Saliency according to Equation 4.1
15:    Write modified frame at the end of a ffmpeg writer object
16:  end for
17:  Save writer object with video name
18: end for
```

we shall consider the question: Does including information about the user’s attention improve the accuracy of our estimations? If that is the case, then we start with other incorporation methods.

Similarly to the BMS360 algorithm, Cube Padding outputs the estimated saliency maps in image frames. However, its input its input is an mp4 video, which is different from a video in YUV format. The MP4 file keeps the data and can potentially compress the data in a lossy way unless we specify it not to do so. Therefore, a lossless conversion has to be performed to evaluate the saliency with Cube Padding. To incorporate the saliency maps using both these methods, we face the problem of the large size of the video. Therefore, in order to manipulate a video frame-by-frame, we make use of the Python3 library `skvideo`.

4.4 Quality metrics and evaluation

After we have the videos with the saliency maps included, we can calculate the three chosen metrics (PSNR, MS-SSIM, and VMAF) to compare whether we can improve the accuracy of our predictions compared to the absence of such incorporation. To do this evaluation, we use Saigg and Scholles’ visual quality framework [36], which standardizes the calculation of metrics and its correlation coefficients, presenting the user a framework for these metrics. The framework allows for the calculation of a total of 11 visual quality metrics. From these metrics, PSNR, MS-SSIM and VMAF were chosen because of their ubiquity in the literature and because they consist of inherently different methods. The flow chart for this framework is shown in Figure 4.1. In step 1, they prepare the video data into pandas dataframes, then they use the Ffmpeg and OpenCV

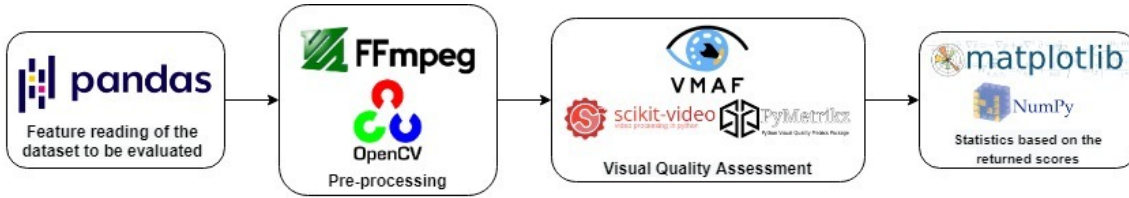


Figure 4.1: Saigg and Scholles’s framework as shown in [36].

libraries to preprocess the data so that the quality metrics can be calculated in step 3. Step 4 calculates the correlations between the MOS values and the quality metrics from step 3 and also is the plotting mechanism which makes use of `matplotlib`. The general flow chart used for our analysis is shown in Figure 4.2.

After we have the quality metrics computed, we compute the performance metrics: Pearson, Spearman, and Kendall correlation coefficients, as well as the RMSE. After that, we generate a graph with the predicted scores versus the MOS scores (computed from the raw scores of the videos from the data-set) is generated for a point-wise visualization. It is important to note that the proposed framework was originally made for 2D videos. As of the time of the writing of this text, no metric specific to 360° environment had been implemented to the framework. As some of the 2D metrics can also be used for 3D videos, granted they are not the most reliable, we deemed reasonable to use the current metrics presented in the framework.

In order to process the videos in parallel with the framework, we have divided all of them into four batches, one for the videos used in the BMS360 processing, one for its saliency inclusion, one for the videos used in the Cube Padding processing and one for its saliency inclusion. Each batch needed a respective `.csv` file to run the set of specified reference and distorted videos, with information about the spatial and temporal resolution of these videos. To compute the correlation coefficients, we used the individual MOS values for each test content provided by the data set. So, after the quality framework finishes running the metrics, we join the `.csv` of objective and subjective quality scores, and the statistics are computed. The framework performs a statistical analysis outputting correlation coefficients in order to compare performance.

4.5 Setup and Fine-tuning

For this part, 20 videos of the dataset were selected, which was mainly a choice of time limitation. For the respective distorted videos, we chose the ERP projection and the evaluation was based upon varying the QP. The QPs in MaiXu’s dataset were set to 27, 37 and 42, therefore spanning across a wide range. The videos chosen were those in the Table 4.1.

A different set of videos was chosen for the evaluation of BMS360 and Cube Padding performance. This is due to a limitation hardcoded in the cube-padding algorithm where the width of the frame must necessarily be twice the height of the frame. This could be solved by stretching one of the dimensions so that we get that proportion, but this would introduce another level of distortion,

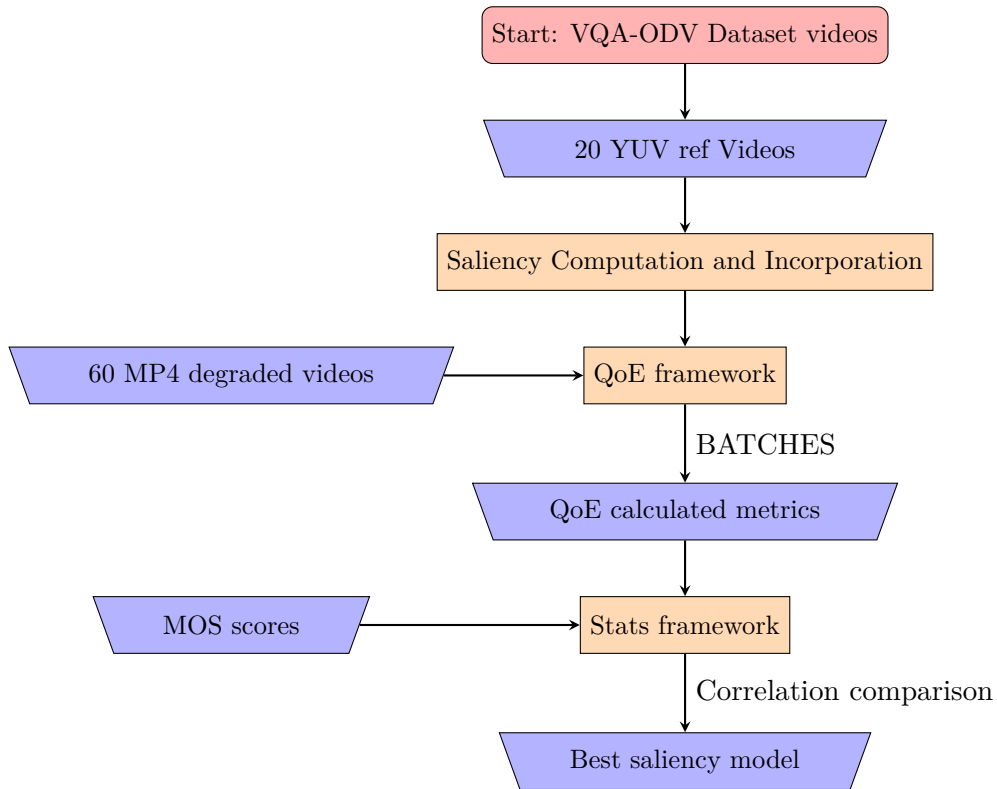


Figure 4.2: The full flowchart representing the methods and subproducts of our analysis. Here, the rhombus represents inputs/outputs and the rectangles represent the operations.

which could further bias our results. Since videos are well spread out among classes, by choosing videos roughly spread out, we aimed to mitigate any advantage given to any of the methods.

The author deemed such a solution the best option due to the time limitations for the publication of this work. The main idea is that, since we are comparing the difference the inclusion of saliency makes, and not the difference from a saliency method to the other, there should not be any problems. As a perspective work, comparing more videos and the same videos would be the best approach.

Table 4.1 shows the chosen videos. We attempted to choose videos well spread-out among classes, with different environments and contexts. Due to a runtime error, there were no 8K videos chosen for the BMS360 processing. The videos name scheme is: Group (G1 to G10) + Name of video _ width x height _ frames per second. Column MOS shows the computation from the data set provides spreadsheets with raw score values from each viewer using Equation 3.1.

For each one of these reference videos, we calculate the corresponding saliency maps, and this saliency is incorporated in a frame-by-frame basis to both the reference and the distorted video. Recall that we have three distorted videos for every reference video. One of the constraints of the framework is a limited option of resolution. For example, we decided against the analysis of the video G4BikingInSaalbach_5600x2800 due to these constraints.

In terms of runtime problems, in addition to the time taken to run all videos (a little over a

Table 4.1: List of all videos used from the VQA-ODV in this research, their respective dimensions and MOS values.

BMS360				Cubepadding			
Video File	height	width	MOS	Video File	height	width	MOS
G1BajaCalifornia_3840x2160_fps23.976	2160	3840	64.91	G10BoatInPark_4096x2048_fps30	2048	4096	67.93
G1BikingToWork_3840x2160_fps23.976	2160	3840	67.15	G10BuddhaCave_4096x2048_fps30	2048	4096	69.92
G2AstonVillaGoal_3840x2048_fps24	2048	3840	74.56	G10XiaoGuang_4096x2048_fps30	2048	4096	71.59
G3BackcountrySkiing_3840x1920_fps25	1920	3840	71.33	G1AbandonedKingdom_7680x3840_fps30	3840	7680	74.13
G3GetYoGurl_3840x1920_fps29.97	1920	3840	58.55	G1Aerial_7680x3840_fps25	3840	7680	77.67
G2ForgottenBook_7680x3840_fps30	3840	7680	64.00	G2ForgottenBook_7680x3840_fps30	3840	7680	64.00
G4WingsuitFlight_3840x2048_fps29.97	2048	3840	53.55	G2FormationPace_7680x3840_fps29.97	3840	7680	80.80
G5EbinShader_7168x3584_fps30	3584	7168	70.69	G3BackcountrySkiing_3840x1920_fps25	1920	3840	71.33
G5Neighborhood_3840x1920_fps23.976	1920	3840	66.35	G3GetYoGurl_3840x1920_fps29.97	1920	3840	58.55
G6DragonTale_3840x2160_fps30	2160	3840	74.39	G4CliffsideMansion_7680x3840_fps30	3840	7680	66.02
G6GTRDriving_3840x2160_fps30	2160	3840	66.68	G5Neighborhood_3840x1920_fps23.976	1920	3840	66.35
G7DragonCastleAttatck_3840x2048_fps24	2048	3840	78.94	G5ResistMarch_3840x1920_fps29.97	1920	3840	70.17
G7PressConference_4096x2048_fps30	2048	4096	77.31	G6AngelFallsClimbing_7680x3840_fps29.97	3840	7680	84.42
G8AlpsParagliding_3840x1920_fps25	1920	3840	70.35	G7OrchestraOfSpheres_7680x3840_fps24	3840	7680	83.10
G8ANewEmpire_3840x2048_fps29.97	2048	3840	74.05	G8DivingWithSharks_7680x3840_fps29.97	3840	7680	74.86
G9ConcertLive_4096x2048_fps30	2048	4096	78.21	G8YourMan_7680x3840_fps29.97	3840	7680	73.60
G10BoatInPark_4096x2048_fps30	2048	4096	67.93	G9ConcertLive_4096x2048_fps30	2048	4096	78.21
G10BuddhaCave_4096x2048_fps30	2048	4096	69.92	G9DrivingInCity_3840x1920_fps30	1920	3840	71.63
G10XiaoGuang_4096x2048_fps30	2048	4096	71.59	G4HachaWaterfall_3840x1920_fps29.97	1920	3840	68.31
				G7UcaimaWaterfall_3840x1920_fps29.97	1920	3840	72.52

month), some of the saliency incorporation did not run for all frames. After confirming which videos had these problems we replaced them with videos of similar dimension and returned to regular processing.

Figure 4.3 contains two example frames, one with the incorporation of Cube Padding saliency and the other with the incorporation of the BMS360 saliency. Notice that the video frame is in grayscale. This was decided as a computation strategy so that we would have to process just one channel as opposed to three. This process is accomplished as follows: the video is first converted to grayscale using the pillow `Image.convert('L')` function and then multiplied frame by frame by the square-root of the normalized saliency map. Then the incorporated frames are joined together.



Figure 4.3: Examples of saliency maps generated by the two algorithms analysed, Cube Padding and BMS360. The image on the left shows the BMS360 saliency and the image on the right shows the Cube Padding saliency map. We can easily see the borders of the cube in this projection.

4.6 Experimental Results

All of the analysis done here is performed with the help of the full information tables presented in Chapter 6. In those tables, we have the reference video for every sequence under the label `refFile`, the distorted video sequence under `testFile`, as well as the MOS of the distorted videos, their dimensions and the calculated PSNR, MS-SSIM and VMAF values. The first thing we notice from these values is the drop in MOS: the higher the QP is the lower the MOS is. The QP value is explicit in the full name of the video. For QP = 27, the average MOS is 68.52, for QP = 37, average MOS is 56.57 and finally for QP = 42, average MOS is 42.04, indicating a rapid degradation in perceived quality. Table 4.1 shows the MOS values for all the reference files used.

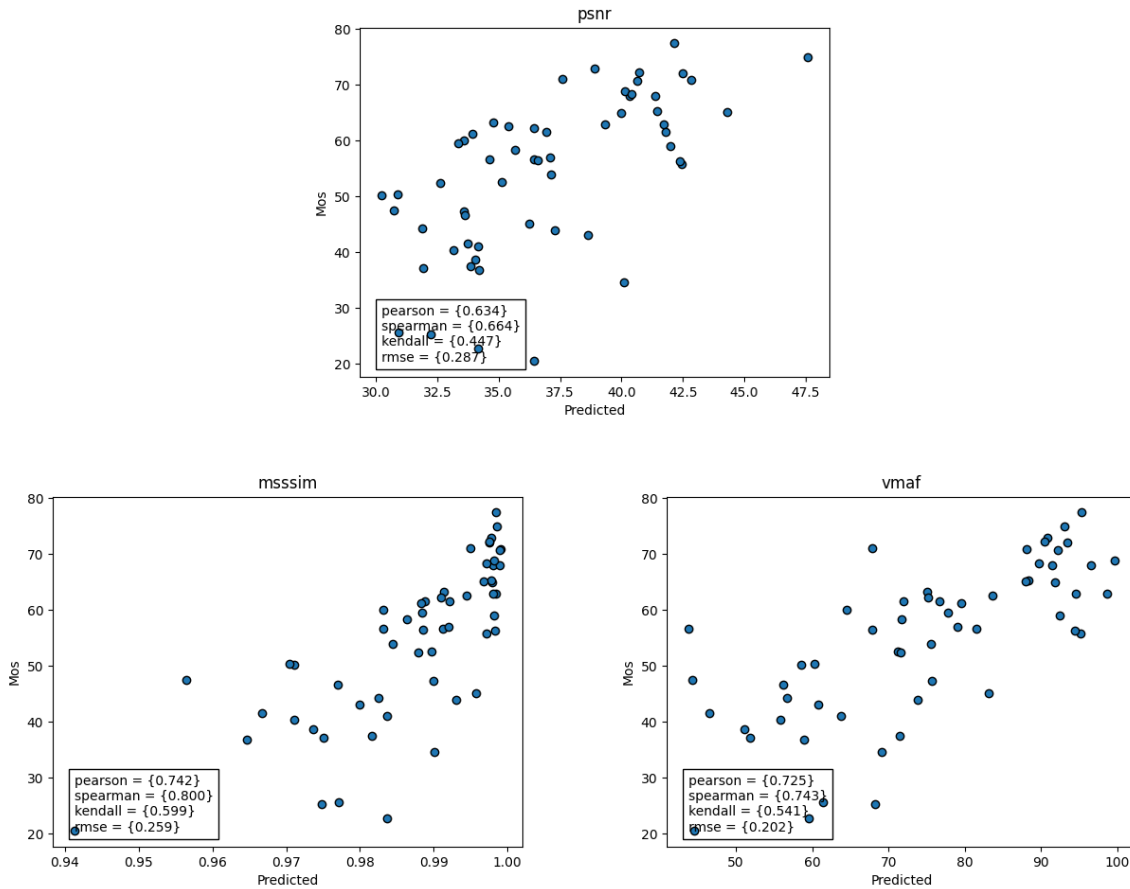


Figure 4.4: Performance distribution for the BMS360 videos. We can see an upward trend, but with a lot of variance for PSNR and VMAF. Points for MS-SSIM tend to be accumulated towards the higher extremes.

An interesting question to ask is if there is any correlation between the resolution of the video (8K versus 4K, for example, and the perceived visual quality) and the perceived MOS. By taking average MOS values in Table 4.1 for 8K and 4K videos, we see that 4K videos have an average MOS of 69.92 whereas 8K videos have an average MOS of 74.92, which means a 5-point difference. These results show that, at least for this batch of videos, 4K videos tend to present just a slightly lower quality level than 8K videos, corroborating the fact that 4K is generally an acceptable

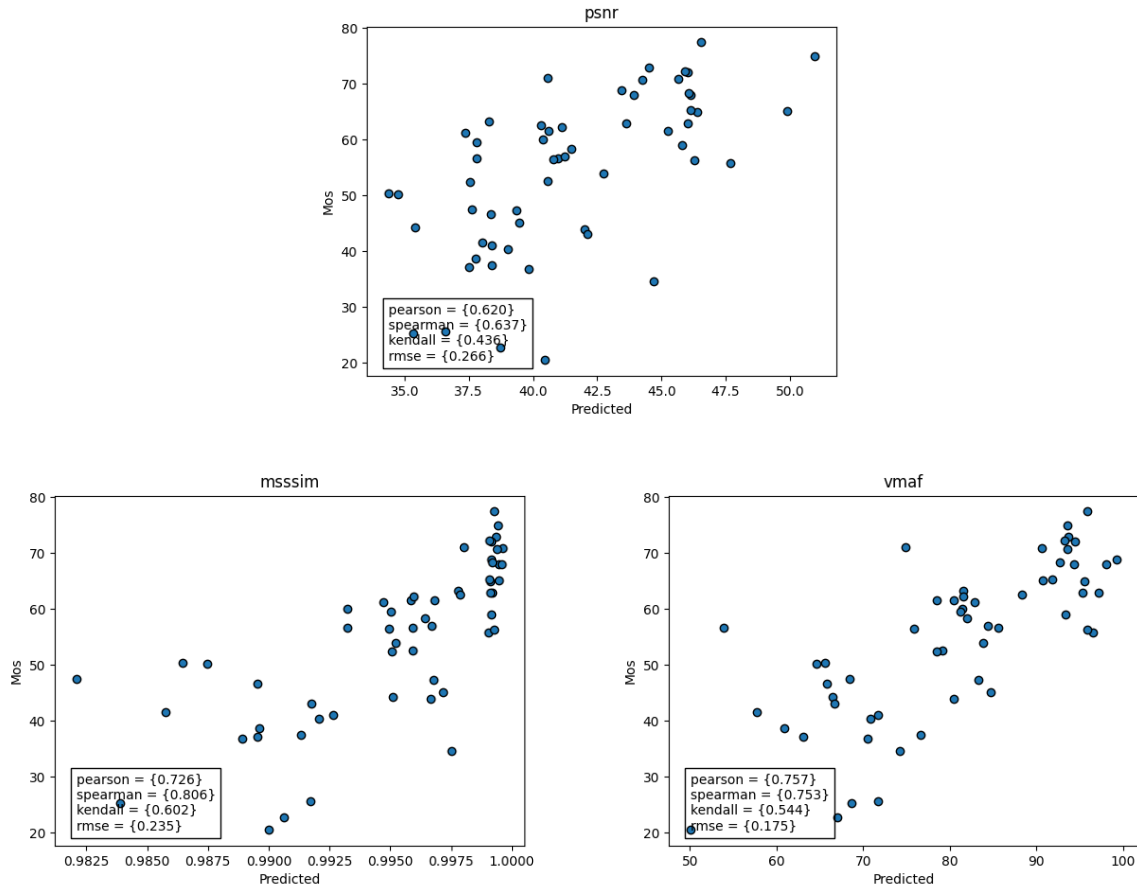


Figure 4.5: Performance distribution for BMS360 videos with saliency incorporation. The overall distribution is relatively similar. That means that visually we cannot see a clear improvement thanks to this incorporation.

resolution for virtual reality videos.

Now we analyse whether the incorporation of saliency in the frame allows better results in our quality prediction.

4.6.1 BMS360: Chosen videos and saliency incorporation results

Table 4.2 shows the prediction results for the videos used in the BMS360 incorporation. Overall, we see a better accuracy for the MS-SSIM metric across the board. As discussed before, this is due to the fact that the SSIM based metrics take into account many human inspired perceptions. VMAF improves with the inclusion of saliency as it is a metric that accounts for information, and the inclusion of saliency apportis more information. This table also shows the results of the correlation coefficients obtained for the incorporation of the saliency into the three metrics. Results remain roughly the same, with a minor advantage for the usage of VMAF in our estimations.

The relationships shown in Figure 4.4 show less variance in VMAF and MS-SSIM, with a

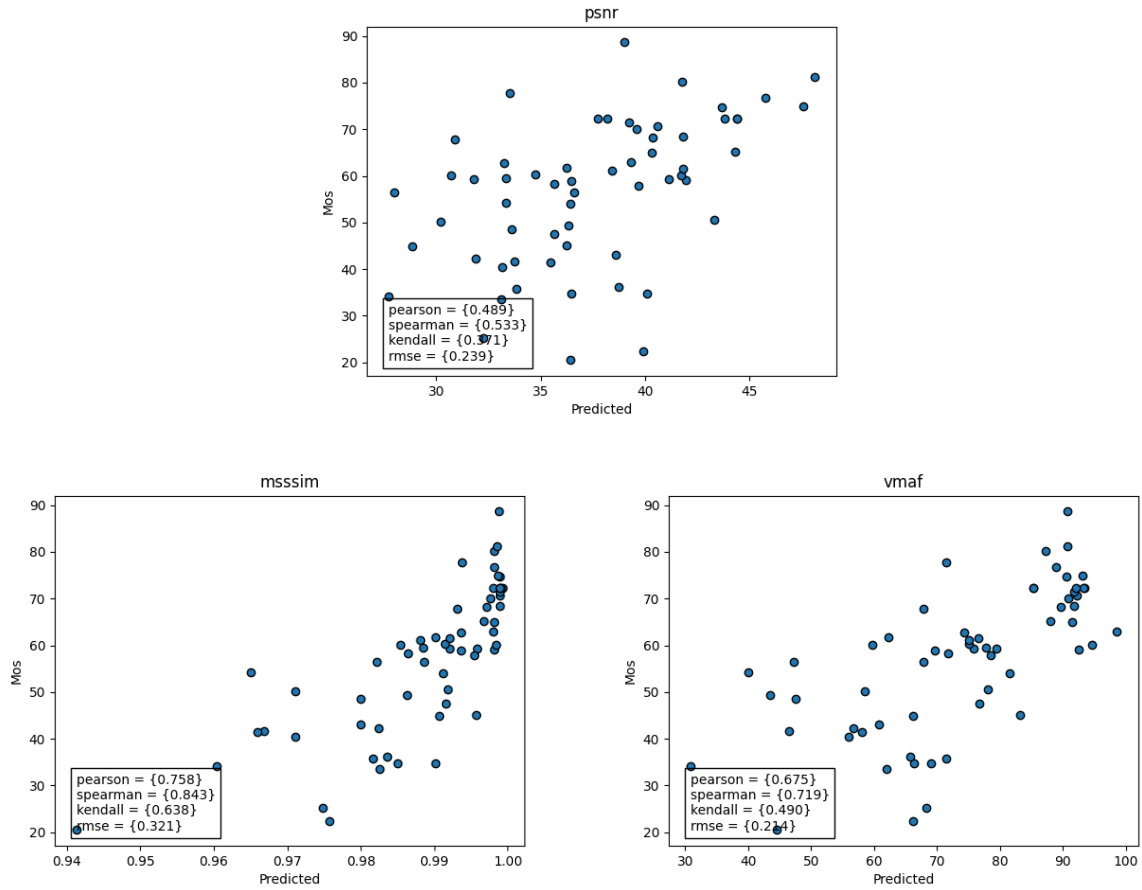


Figure 4.6: Performance distribution for the Cube Padding videos.

general linear relationship between values. PSNR presents the worst performance, as it has the smallest complexity. This graph also tells us about 4K videos, as they constitute almost the totality of analysed videos. In the graphs of Figure 4.6 it is easy to see that, for VMAF, there is an overall smaller variance in the instances.

MS-SSIM, on the other hand, presents an accumulation of high values, where the actual MOS does not show that. In a deeper analysis of all graphs and also the values obtained in the tables presented in our appendix, we can see how the MS-SSIM values are actually not evenly scattered, being concentrated in the higher ranges of the scale. An interesting fact is that although concentrated, the computed MS-SSIM values are the ones with the best difference between Spearman and Linear coefficients of all the metrics, pointing to a more pronounced non-linear relationship with the MOS values.

Comparing both tables, we see a very similar distribution, boxplot shape and quality metrics results. Therefore, overall, the addition of visual attention information in the case of BMS360 is considered roughly irrelevant, so the extra computation effort should be discarded.

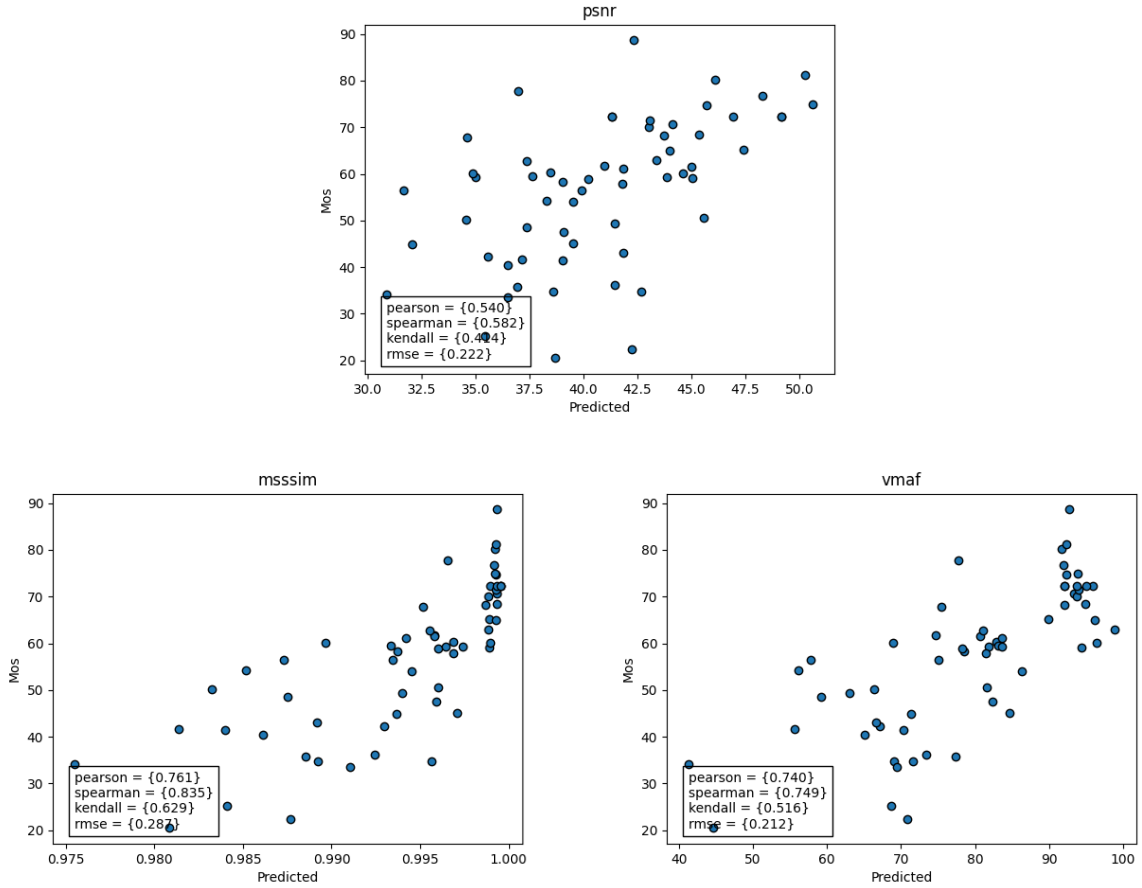


Figure 4.7: Performance Distribution for the Cubepadding Videos with Saliency Incorporation. The distribution is relatively similar to the one shown in figure 4.6.1, but in MS-SSIM and VMAF we see some improvement in terms of variance.

Table 4.2: Table of VQA Metrics for the BMS360 videos with and without saliency incorporation.

BMS360 No Saliency				BMS360 With Saliency			
Correlation	psnr	msssim	vmaf	Correlation	psnr	msssim	vmaf
Pearson	0.634	0.742	0.725	Pearson	0.625	0.726	0.757
Spearman	0.664	0.8	0.743	Spearman	0.637	0.806	0.753
Kendall	0.447	0.599	0.541	Kendall	0.436	0.602	0.544
RMSE	0.287	0.259	0.202	RMSE	0.266	0.235	0.175

4.6.2 Cube Padding: Chosen videos and saliency incorporation results

The Cube Padding videos differ from the previous set of videos as they consider a more even distribution of 4K and 8K videos. Table 4.3 shows MS-SSIM and VMAF maintain a good correlation with the actual MOS values, while PSNR has greatly diminished accuracy considering Pearson. Note how in the graphs of Figure 4.6.1 we see a more random distribution of points for PSNR, whereas in VMAF and particularly MS-SSIM the data distribution has a less pronounced variance. In this case, we have a very similar distribution to BMS360 for PSNR and VMAF,

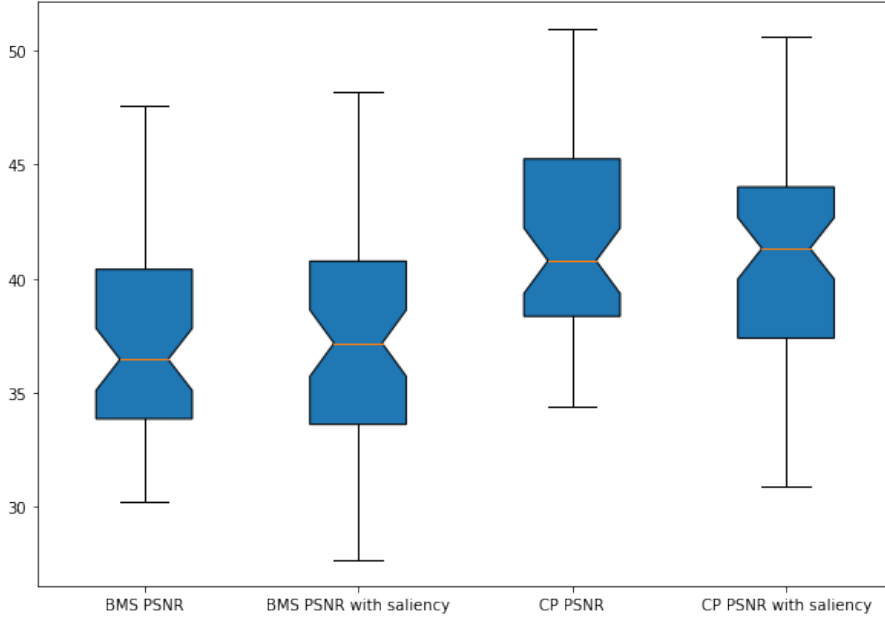


Figure 4.8: Variation for PSNR with different saliency inclusions

and an even greater concentration of points towards the higher extremes in MS-SSIM. This is expected as cube-padding chosen video set has more 8K videos (and therefore, more videos of better quality) than the BMS360 chosen video set.

Differently from the BMS360 situation, Cube Padding is more effective in its application as it allows for a better correlation across the board, with noticeable results in PSNR and VMAF correlation values (7.3 and 6% better correlation on average, respectively) and a roughly level situation for MS-SSIM. This is considering we are talking about a simple saliency incorporation onto the video. If we were to consider a an incorporation of saliency more adapted for each quality metric surely the results would be better.

Table 4.3: Table of VQA Metrics for the Cubepadding videos with and without saliency incorporation.

CP No Saliency				CP With Saliency			
Correlation	psnr	msssim	vmaf	Correlation	psnr	msssim	vmaf
Pearson	0.489	0.758	0.675	Pearson	0.54	0.761	0.74
Spearman	0.533	0.843	0.719	Spearman	0.582	0.835	0.749
Kendall	0.371	0.638	0.49	Kendall	0.414	0.629	0.516
RMSE	0.239	0.321	0.214	RMSE	0.222	0.287	0.212

Figures 4.6.2, 4.6.2 and 4.6.2 compare the boxplots of the different metrics with ad without saliency incorporation. We can see, as said before, a higher predicted value for the Cube Padding videos, as this set of videos is, on average, of higher resolution. Another thing we notice is that, visually, the saliency does not seem to interfere in the overall distribution shape, which means the results do not get biased by the saliency incorporation process. Another important aspect is

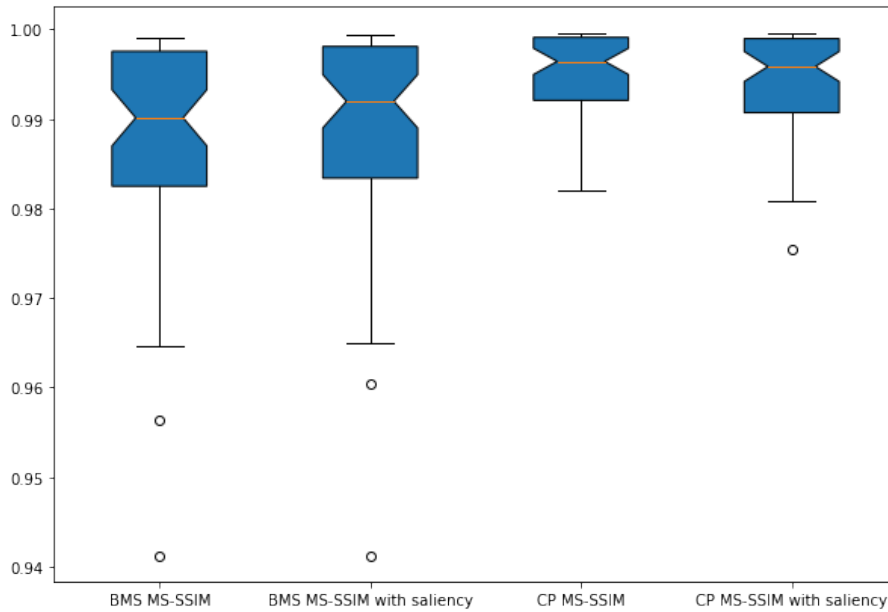


Figure 4.9: Variation for MS-SSIM with different saliency inclusions

the fact that the results do not change that much with the saliency inclusion, showing that the improvements it attains are marginal.

Overall we see that the saliency incorporation as weights is a viable option to improve the accuracy of our quality estimation if we have the computational power in the streamer side to perform this evaluation. Nonetheless, the improvement is minor so the computational time lost has to be taken into account when deciding for this strategy.

As exposed before, an interesting option to improve the result of saliency inclusion now that we know that it can improve our quality prediction is to mathematically analyse how we can incorporate the saliency maps onto the videos in a more adequate manner for each metric, perhaps even as a pre-processing step particular to each metric. This could be included inside the quality metric framework. In our case, since we used the predefined metrics from Saigg and Scholles, the preprocess step had to be done beforehand, and therefore was not perfectly adapted to all metrics.

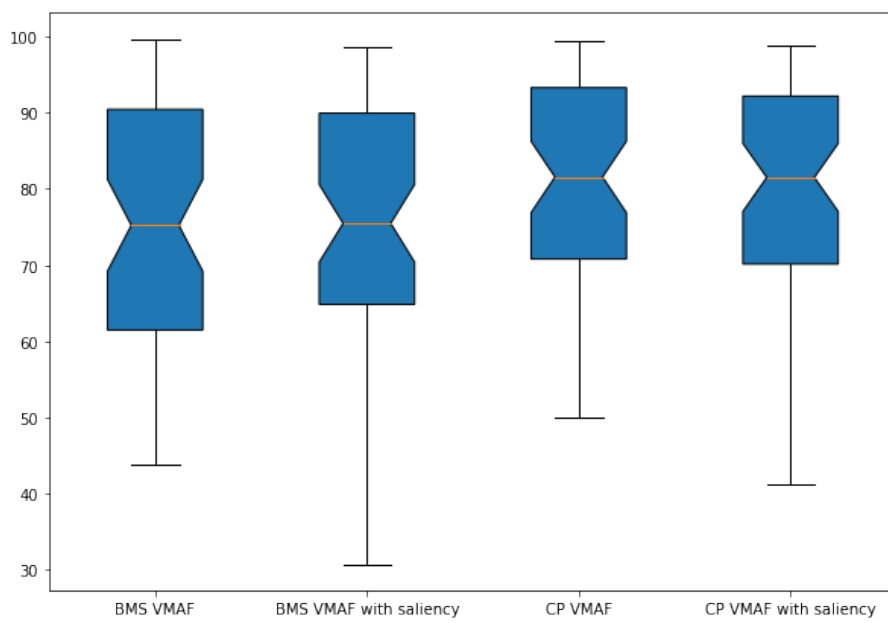


Figure 4.10: Variation for VMAF with different saliency inclusions

Chapter 5

Conclusion

In this work, the ensemble of concepts necessary for the development of techniques which use visual attention combined with quality of experience metrics was explored. Initially, the idea of visual attention was studied and two main saliency map estimation techniques were examined (BMS360 and Cube Padding) and the necessary elements for their understanding.

In a second moment, the concept of Visual Quality Assessment was explored in many of its facets, stating from what causes video degradation, how experiments to assess quality of experience are performed and how we can attempt to get rough estimation with computational methods such as PSNR, MS-SSIM and VMAF. Ways to show how the computations compare to the actual experiments were also explored at this moment.

In light of the theory researched, this work showed the materials with which it intended to answer its main question. Starting with a quick exploration of the VQA-ODV dataset and the BMS360 and Cube Padding adaptations for our case. This work then proposes a way to incorporate the saliency maps into the metrics inspired by the WS-PSNR metric, which is a robust attempt of this incorporation and allows for an already tested weights multiplication. Finally, a framework for the calculation of the VQA metrics and of the statistical correlations are shown.

As subproducts of our research, we saw the difference QP and resolution make in the MOS values. The lower the QP and the higher the resolution, the better the video perception is. By analysing the images from the different saliency methods, we also saw how they differ and how visually the saliency is manifested. We also were able to understand the different processing times, the importance of the standardization for the sake of the research.

Finally, by calculating the MOS values, the PSNR, MS-SSIM and the VMAF values for both saliency incorporated and regular videos this work managed to show that there is a slight advantage of using the Cube Padding saliency no advantage in using BMS360. the advantage was of as much as 7%. This is most likely due to the data-driven nature of Cube Padding.

Because Cube Padding is also relatively fast to run, and we can suppose that is the case for other data-driven approaches running on a powerful computer. Therefore, a streamer can be interested in implementing such an algorithm to predict the expected quality of experience of the

user. Since the VQA metrics directly affect both sessions lengths and viewer engagement and can determine the success or failure of your streaming video venture. Viewers experiencing poor quality are more likely to tune out while those with a high quality of experience become repeat viewers.

Proposed Ideas

There are several ideas we could explore in a future context:

- In the area of Visual Attention, we could explore several other saliency models, in particular temporal models which take into account the optical flow in the video. Visual attention models which explore more Top-down ideas can be an interesting addition as it brings more information, in particular regarding the movement in the scenes. More accurate saliency maps mean better information addition and therefore more accuracy in the final result.
- The creation of an original saliency model incorporating new ideas appearing in the visual attention modelling and machine learning domains, such as transformers and alternatives to LSTMs. In order to create this saliency model, several videos with their respective fixation annotations would be necessary, as well as a thorough study of the current state-of-the-art in this and other data-driven domains. This idea could be explored in partnership with Saigg and Scholles, expanding their framework to contemplate 3D specific metrics and other types too.
- In the area of Quality Metrics, the exploration of other objective metrics is an evident addition; in particular the spherical metrics and the newer machine learning metrics, the latter being the state-of-the-art in the area such as the one proposed in [41]. Methods such as S-PSNR, S-SSIM and other similar methods which attribute weights to the quality metrics provide a more accurate representation of how quality is perceived in the 360° environment.
- In the specific case of the exploration of the VQA-ODV dataset, we can expand the analysis to the full set of 60 reference videos as well as the exploration of the distortions caused by the Cubic and TSP projections. Still on the exploration of this dataset, an analysis of the DMOS values and of the HM and EM measurements can be interesting to get more accurate results for our metrics and ground truth saliency maps to compare with the saliency estimations of all the other methods.
- A more robust way of incorporating the saliency maps can also be explored, taking into account more of the mathematical peculiarities of each one of the Quality of Experience metrics at play. A good idea would be to make this incorporation inside the framework for calculating metrics. With that, the framework would just receive the reference, impaired videos and saliency maps and for each quality metric the framework would include the saliency maps in an adapted way for that metric.

- In our discussion, a deeper dive into the runtime errors, into the saliency maps themselves and also into the graphic plots in order to more closely inspect all of the facets of the results, how to solve any problems that arise and how each aspect influences the final result. By doing this, we can get a more complete understanding of all aspects playing in this research.
- Finally, as the Covid-19 pandemic safety measures get relaxed, a dataset of University of Brasilia's authoring could motivate a whole new batch of research in the area.

In summary, this work has just scratched the surface of the potential surrounding the area of visual attention and quality of experience, an area so new and which motivates so many ideas towards a future where fully immersive media will be widespread. We can expect more and more research in the field as the new area becomes increasingly important and the user experience and retention more relevant as the amount of information being transmitted through immersive media increases.

REFERENCES

- [1] C. BAMPIS AND A. BOVIK, *Learning to predict streaming video qoe: Distortions, rebuffering and memory*, (2017).
- [2] A. BORJI AND L. ITTI, *State-of-the-art in visual attention modeling*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 185–207.
- [3] K. CHEN, H. J. CHOI, AND D. BREN, *Visual attention and eye movements*, 2008.
- [4] S. CHEN, Y. ZHANG, Y. LI, Z. CHEN, AND Z. WANG, *Spherical structural similarity index for objective omnidirectional video quality assessment*, 07 2018, pp. 1–6.
- [5] Z. CHEN, J. YUAN, AND Y.-P. TAN, *Hybrid saliency detection for images*, IEEE Signal Processing Letters, 20 (2013), pp. 95–98.
- [6] H.-T. CHENG, C.-H. CHAO, J.-D. DONG, H.-K. WEN, T.-L. LIU, AND M. SUN, *Cube padding for weakly-supervised saliency prediction in 360° videos*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2018).
- [7] F. CHIARIOTTI, *A survey on 360-degree video: Coding, quality of experience and streaming*, ArXiv, abs/2102.08192 (2021).
- [8] A. DE ABREU, C. OZCINAR, AND A. SMOLIC, *Look around you: Saliency maps for omnidirectional images in vr applications*, 05 2017.
- [9] R. G. DE ALBUQUERQUE AZEVEDO; NEIL BIRKBECK; FRANCESCA DE SIMONE; IVAN JANATRA, B. ADSUMILLI, AND P. FROSSARD, *Visual distortions in 360° videos*, IEEE Transactions on Circuits and Systems for Video Technology, 30 (2020), pp. 2524–2537.
- [10] J. DENG, W. DONG, R. SOCHER, L.-J. LI, K. LI, AND L. FEI-FEI, *Imagenet: A large-scale hierarchical image database*, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
- [11] U. ENGELKE, M. BARKOWSKY, P. LE CALLET, AND H.-J. ZEPERNICK, *Modelling saliency awareness for objective video quality assessment*, 07 2010, pp. 212 – 217.
- [12] M. GOODALE AND A. MILNER, *Separate visual pathways for perception and action*, Trends in Neurosciences, 15 (1992), pp. 20–25.

- [13] J. HAREL, C. KOCH, AND P. PERONA, *Graph-based visual saliency*, in Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, Cambridge, MA, USA, 2006, MIT Press, p. 545–552.
- [14] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
- [15] A. HORÉ AND D. ZIOU, *Image quality metrics: Psnr vs. ssim*, in 2010 20th International Conference on Pattern Recognition, 2010, pp. 2366–2369.
- [16] L. HUANG AND H. PASHLER, *A boolean map theory of visual attention.*, Psychological review, 114 3 (2007), pp. 599–631.
- [17] X. HUANG, C. SHEN, X. BOIX, AND Q. ZHAO, *Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks*, in 2015 International Conference on Computer Vision, ICCV 2015, Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., Feb. 2015, pp. 262–270. 15th IEEE International Conference on Computer Vision, ICCV 2015 ; Conference date: 11-12-2015 Through 18-12-2015.
- [18] T. HUYEN, N. PHAM NGOC, C. PHAM, Y. JUNG, AND T. CONG THANG, *A subjective study on user perception aspects in virtual reality*, Applied Sciences, 9 (2019).
- [19] L. ITTI, C. KOCH, AND E. NIEBUR, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 20 (1998), p. 1254–1259.
- [20] L. ITTI, C. KOCH, AND E. NIEBUR, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (1998), pp. 1254–1259.
- [21] C. KOCH AND S. ULLMAN, *Shifts in selective visual attention: towards the underlying neural circuitry.*, Human neurobiology, 4 4 (1985), pp. 219–27.
- [22] P. LEBRETON AND A. RAAKE, *Gbus360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images*, Signal Processing: Image Communication, 69 (2018), pp. 69–78.
- [23] C. LI AND A. BOVIK, *Content-weighted video quality assessment using a three-component image model*, Journal of Electronic Imaging, 29 (2010).
- [24] C. LI, M. XU, X. DU, AND Z. WANG, *Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model*, 07 2018.
- [25] C. LI, M. XU, L. JIANG, S. ZHANG, AND X. TAO, *Viewport proposal cnn for 360deg video quality assessment*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

- [26] T.-J. LIU, Y.-C. LIN, W. LIN, AND C.-C. J. KUO, *Visual quality assessment: recent developments, coding applications and future trends*, APSIPA Transactions on Signal and Information Processing, 2 (2013), p. e4.
- [27] B. MANOR AND E. GORDON, *Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks*, Journal of neuroscience methods, 128 (2003), pp. 85–93.
- [28] M. ORDUNA, C. DÍAZ, L. MUÑOZ, P. PÉREZ, I. BENITO, AND N. GARCÍA, *Video multimethod assessment fusion (vmaf) on 360vr contents*, IEEE Transactions on Consumer Electronics, 66 (2020), pp. 22–31.
- [29] J. PAN, E. SAYROL, X. GIRO-I NIETO, K. MCGUINNESS, AND N. E. O’CONNOR, *Shallow and deep convolutional networks for saliency prediction*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
- [30] N. PAQUETTE AND M. NEIDER, *Attentional guidance in visual search: Examining the interaction between goal driven and stimulus driven information in natural images*, Journal of Vision, 13 (2013), pp. 162–162.
- [31] A. PATRONE, C. VALUCH, U. ANSORGE, AND O. SCHERZER, *Dynamical optical flow of saliency maps for predicting visual attention*, (2016).
- [32] M. PINSON AND S. WOLF, *A new standardized method for objectively measuring video quality*, Broadcasting, IEEE Transactions on, 50 (2004), pp. 312 – 322.
- [33] M. QUELUZ, F. LOPES, J. ASCENSO, AND A. RODRIGUES, *Subjective and objective quality assessment of omnidirectional video*, 09 2018, p. 25.
- [34] Y. RAI, J. GUTIÉRREZ, AND P. LE CALLET, *A dataset of head and eye movements for 360 degree images*, 06 2017, pp. 205–210.
- [35] R. RASSOOL, *Vmaf reproducibility: Validating a perceptual practical video quality metric*, in 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2017, pp. 1–2.
- [36] C. SAIGG AND B. SCHOLLES, *Framework-for-objective-visual-quality-assessment-fovqa*, available at <https://github.com/scholles007/framework-for-objective-visual-quality-assessment-fovqa>, 2022.
- [37] V. SITZMANN, A. SERRANO, A. PAVEL, M. AGRAWALA, D. GUTIERREZ, B. MASIA, AND G. WETZSTEIN, *How do people explore virtual environments?*, IEEE Transactions on Visualization and Computer Graphics, (2017).
- [38] P. TOOLS, *Rate control and h.264*, available at https://www.pixeltools.com/rate_control_paper.html.
- [39] A. TREISMAN AND G. GELADE, *A feature-integration theory of attention*, Cognitive Psychology, 12 (1980), pp. 97–136.

- [40] M. VARELA, L. SKORIN-KAPOV, AND T. EBRAHIMI, *Quality of Service vs. Quality of Experience*, 03 2014, pp. 85–96.
- [41] V. VASILEV, J. LEGUAY, S. PARIS, L. MAGGI, AND M. DEBBAH, *Predicting qoe factors with machine learning*, in 2018 IEEE International Conference on Communications (ICC), 2018, pp. 1–6.
- [42] Z. WANG AND A. BOVIK, *Bovik, a.c.: Mean squared error: love it or leave it? - a new look at signal fidelity measures. iee sig. process. mag.* 26, 98-117, Signal Processing Magazine, IEEE, 26 (2009), pp. 98 – 117.
- [43] Z. WANG, A. BOVIK, H. SHEIKH, AND E. SIMONCELLI, *Image quality assessment: from error visibility to structural similarity*, IEEE Transactions on Image Processing, 13 (2004), pp. 600–612.
- [44] Z. WANG, L. LU, AND A. C. BOVIK, *Video quality assessment based on structural distortion measurement*, Signal Processing: Image Communication, 19 (2004), pp. 121–132.
- [45] M. XU, C. LI, Z. CHEN, Z. WANG, AND Z. GUAN, *Assessing visual quality of omnidirectional videos*, IEEE Transactions on Circuits and Systems for Video Technology, PP (2018), pp. 1–1.
- [46] M. XU, C. LI, S. ZHANG, AND P. L. CALLET, *State-of-the-art in 360° video/image processing: Perception, assessment and compression*, IEEE Journal of Selected Topics in Signal Processing, 14 (2020), p. 5–26.
- [47] M. XU, Y. REN, Z. WANG, J. LIU, AND X. TAO, *Saliency detection in face videos: A data-driven approach*, IEEE Transactions on Multimedia, 20 (2018), pp. 1335–1349.
- [48] M. XU, Y. SONG, J. WANG, M. QIAO, L. HUO, AND Z. WANG, *Predicting head movement in panoramic video: A deep reinforcement learning approach*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PP (2018), pp. 1–1.
- [49] S. YANG, J. ZHAO, T. JIANG, J. WANG, T. RAHIM, B. ZHANG, Z. XU, AND Z. FEI, *An objective assessment method based on multi-level factors for panoramic videos*, in 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1–4.
- [50] A. YARBUS, *Eye Movement and Vision*, 1967.
- [51] M. YU, H. LAKSHMAN, AND B. GIROD, *A framework to evaluate omnidirectional video coding schemes*, 2015 IEEE International Symposium on Mixed and Augmented Reality, (2015), pp. 31–36.
- [52] M. YU, H. LAKSHMAN, AND B. GIROD, *A framework to evaluate omnidirectional video coding schemes*, 2015 IEEE International Symposium on Mixed and Augmented Reality, (2015), pp. 31–36.
- [53] J. ZHANG AND S. SCLAROFF, *Saliency detection : A boolean map approach supplementary materials*, 2013.

- [54] Y. ZHOU, M. YU, H. MA, H. SHAO, AND G. JIANG, *Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video*, 08 2018, pp. 54–57.
- [55] W. ZOU, F. YANG, AND S. WAN, *Perceptual video quality metric for compression artefacts: from two-dimensional to omnidirectional*, IET Image Process., 12 (2018), pp. 374–381.

Chapter 6

Appendix

Table 6.1: Full information table for BMS360 without saliency inclusion.

Reference Video	Distorted Video	Mos	height	width	psnr	msssim	vmaf
G1BajaCalifornia_3840x2160_fps23.976	G1BajaCalifornia_ERP_3840x2160_fps23.976_qp27_41186k	64.942	2160	3840	39.988	0.998	91.863
G1BajaCalifornia_3840x2160_fps23.976	G1BajaCalifornia_ERP_3840x2160_fps23.976_qp37_12633k	60.041	2160	3840	33.600	0.983	64.470
G1BajaCalifornia_3840x2160_fps23.976	G1BajaCalifornia_ERP_3840x2160_fps23.976_qp42_5570k	47.554	2160	3840	30.745	0.956	44.335
G1BikingToWork_3840x2160_fps23.976	G1BikingToWork_ERP_3840x2160_fps23.976_qp27_12306k	55.828	2160	3840	42.450	0.997	95.194
G1BikingToWork_3840x2160_fps23.976	G1BikingToWork_ERP_3840x2160_fps23.976_qp37_3596k	54.019	2160	3840	37.138	0.984	75.539
G1BikingToWork_3840x2160_fps23.976	G1BikingToWork_ERP_3840x2160_fps23.976_qp42_1855k	36.841	2160	3840	34.200	0.965	58.895
G2AstonVillaGoal_3840x2048_fps24	G2AstonVillaGoal_ERP_3840x2048_fps24_qp27_11596k	72.924	2048	3840	38.892	0.998	90.817
G2AstonVillaGoal_3840x2048_fps24	G2AstonVillaGoal_ERP_3840x2048_fps24_qp37_2141k	47.285	2048	3840	33.601	0.990	75.628
G2AstonVillaGoal_3840x2048_fps24	G2AstonVillaGoal_ERP_3840x2048_fps24_qp42_872k	25.691	2048	3840	30.922	0.977	61.476
G3BackcountrySkiing_3840x1920_fps25	G3BackcountrySkiing_ERP_3840x1920_fps25_qp27_19131k	68.007	1920	3840	40.336	0.998	91.464
G3BackcountrySkiing_3840x1920_fps25	G3BackcountrySkiing_ERP_3840x1920_fps25_qp37_4153k	63.254	1920	3840	34.765	0.991	75.065
G3BackcountrySkiing_3840x1920_fps25	G3BackcountrySkiing_ERP_3840x1920_fps25_qp42_1608k	44.197	1920	3840	31.881	0.982	56.697
G3GetYoGurl_3840x1920_fps29.97	G3GetYoGurl_ERP_3840x1920_fps29.97_qp27_14177k	62.896	1920	3840	41.726	0.998	94.595
G3GetYoGurl_3840x1920_fps29.97	G3GetYoGurl_ERP_3840x1920_fps29.97_qp37_3395k	56.638	1920	3840	36.445	0.991	81.489
G3GetYoGurl_3840x1920_fps29.97	G3GetYoGurl_ERP_3840x1920_fps29.97_qp42_1688k	37.557	1920	3840	33.847	0.982	71.522
G2ForgottenBook_7680x3840_fps30	G2ForgottenBook_ERP_7680x3840_fps30_qp27_41804k	59.033	3840	7680	41.972	0.998	92.490
G4WingsuitFlight_3840x2048_fps29.97	G4WingsuitFlight_ERP_3840x2048_fps29.97_qp27_12696k	56.289	2048	3840	42.370	0.998	94.392
G5EbinShader_7168x3584_fps30	G5EbinShader_ERP_7168x3584_fps30_qp27_31040k	70.924	3584	7168	42.855	0.999	88.110
G5Neighborhood_3840x1920_fps23.976	G5Neighborhood_ERP_3840x1920_fps23.976_qp27_24718k	62.925	1920	3840	39.328	0.998	98.583
G6DragonTale_3840x2160_fps30	G6DragonTale_ERP_3840x2160_fps30_qp27_13868k	72.102	2160	3840	42.475	0.998	93.437
G6GTRDriving_3840x2160_fps30	G6GTRDriving_ERP_3840x2160_fps30_qp27_28748k	68.049	2160	3840	41.353	0.999	96.504
G7DragonCastleAttack_3840x2048_fps24	G7DragonCastleAttack_ERP_3840x2048_fps24_qp27_15133k	77.435	2048	3840	42.155	0.998	95.265
G7PressConference_4096x2048_fps30	G7PressConference_ERP_4096x2048_fps30_qp27_4909k	65.319	2048	4096	41.454	0.998	88.344
G8AlpsParagliding_3840x1920_fps25	G8AlpsParagliding_ERP_3840x1920_fps25_qp27_8437k	72.284	1920	3840	40.714	0.997	90.428
G8ANewEmpire_3840x2048_fps29.97	G8ANewEmpire_ERP_3840x2048_fps29.97_qp27_33327k	68.867	2048	3840	40.121	0.998	99.690
G9ConcertLive_4096x2048_fps30	G9ConcertLive_ERP_4096x2048_fps30_qp27_7102k	74.963	2048	4096	47.596	0.999	93.107
G10BoatInPark_4096x2048_fps30	G10BoatInPark_ERP_4096x2048_fps30_qp27_14547k	68.308	2048	4096	40.402	0.997	89.711
G10BuddhaCave_4096x2048_fps30	G10BuddhaCave_ERP_4096x2048_fps30_qp27_1289k	65.164	2048	4096	44.319	0.997	88.013
G10XiaoGuang_4096x2048_fps30	G10XiaoGuang_ERP_4096x2048_fps30_qp27_2502k	70.666	2048	4096	40.622	0.999	92.191
G2ForgottenBook_7680x3840_fps30	G2ForgottenBook_ERP_7680x3840_fps30_qp37_9445k	56.442	3840	7680	36.594	0.989	67.885
G4WingsuitFlight_3840x2048_fps29.97	G4WingsuitFlight_ERP_3840x2048_fps29.97_qp37_3545k	57.062	2048	3840	37.082	0.992	79.073
G5EbinShader_7168x3584_fps30	G5EbinShader_ERP_7168x3584_fps30_qp37_4094k	71.059	3584	7168	37.584	0.995	67.915
G5Neighborhood_3840x1920_fps23.976	G5Neighborhood_ERP_3840x1920_fps23.976_qp37_7721k	59.554	1920	3840	33.347	0.988	77.766
G6DragonTale_3840x2160_fps30	G6DragonTale_ERP_3840x2160_fps30_qp37_3984k	61.631	2160	3840	36.954	0.989	71.995
G6GTRDriving_3840x2160_fps30	G6GTRDriving_ERP_3840x2160_fps30_qp37_8912k	62.608	2160	3840	35.409	0.994	83.600
G7DragonCastleAttack_3840x2048_fps24	G7DragonCastleAttack_ERP_3840x2048_fps24_qp37_4272k	62.215	2048	3840	36.454	0.991	75.216
G7PressConference_4096x2048_fps30	G7PressConference_ERP_4096x2048_fps30_qp37_1038k	43.919	2048	4096	37.300	0.993	73.822
G8AlpsParagliding_3840x1920_fps25	G8AlpsParagliding_ERP_3840x1920_fps25_qp37_1995k	52.499	1920	3840	35.139	0.990	71.184
G8ANewEmpire_3840x2048_fps29.97	G8ANewEmpire_ERP_3840x2048_fps29.97_qp37_10885k	61.291	2048	3840	33.914	0.988	79.572
G9ConcertLive_4096x2048_fps30	G9ConcertLive_ERP_4096x2048_fps30_qp37_2256k	61.501	2048	4096	41.814	0.992	76.633
G10BoatInPark_4096x2048_fps30	G10BoatInPark_ERP_4096x2048_fps30_qp37_3270k	58.305	2048	4096	35.661	0.986	71.708
G10BuddhaCave_4096x2048_fps30	G10BuddhaCave_ERP_4096x2048_fps30_qp37_236k	34.698	2048	4096	40.105	0.990	69.059
G10XiaoGuang_4096x2048_fps30	G10XiaoGuang_ERP_4096x2048_fps30_qp37_644k	45.178	2048	4096	36.256	0.996	83.191
G2ForgottenBook_7680x3840_fps30	G2ForgottenBook_ERP_7680x3840_fps30_qp42_4374k	41.642	3840	7680	33.739	0.967	46.469
G4WingsuitFlight_3840x2048_fps29.97	G4WingsuitFlight_ERP_3840x2048_fps29.97_qp42_1877k	41.085	2048	3840	34.179	0.984	63.753
G5EbinShader_7168x3584_fps30	G5EbinShader_ERP_7168x3584_fps30_qp42_1383k	56.605	3584	7168	34.643	0.983	43.739
G5Neighborhood_3840x1920_fps23.976	G5Neighborhood_ERP_3840x1920_fps23.976_qp42_3894k	50.164	1920	3840	30.215	0.971	58.551
G6DragonTale_3840x2160_fps30	G6DragonTale_ERP_3840x2160_fps30_qp42_2057k	38.656	2160	3840	34.046	0.974	51.099
G6GTRDriving_3840x2160_fps30	G6GTRDriving_ERP_3840x2160_fps30_qp42_4245k	52.463	2160	3840	32.633	0.988	71.596
G7DragonCastleAttack_3840x2048_fps24	G7DragonCastleAttack_ERP_3840x2048_fps24_qp42_2084k	46.729	2048	3840	33.613	0.977	56.169
G7PressConference_4096x2048_fps30	G7PressConference_ERP_4096x2048_fps30_qp42_509k	22.707	2048	4096	34.170	0.984	59.569
G8AlpsParagliding_3840x1920_fps25	G8AlpsParagliding_ERP_3840x1920_fps25_qp42_964k	37.148	1920	3840	31.929	0.975	51.891
G8ANewEmpire_3840x2048_fps29.97	G8ANewEmpire_ERP_3840x2048_fps29.97_qp42_5571k	50.376	2048	3840	30.904	0.970	60.319
G9ConcertLive_4096x2048_fps30	G9ConcertLive_ERP_4096x2048_fps30_qp42_1246k	43.130	2048	4096	38.627	0.980	60.741
G10BoatInPark_4096x2048_fps30	G10BoatInPark_ERP_4096x2048_fps30_qp42_1507k	40.372	2048	4096	33.153	0.971	55.883
G10BuddhaCave_4096x2048_fps30	G10BuddhaCave_ERP_4096x2048_fps30_qp42_170k	20.475	2048	4096	36.452	0.941	44.491
G10XiaoGuang_4096x2048_fps30	G10XiaoGuang_ERP_4096x2048_fps30_qp42_393k	25.278	2048	4096	32.241	0.975	68.229

Table 6.2: Full information table for Cubepadding without saliency inclusion.

Reference Video	Distorted Video	Mos	height	width	psnr	msssim	vmaf
G10BoatInPark_4096x2048_fps30	G10BoatInPark_ERP_4096x2048_fps30_qp27_14547k	68.308	2048	4096	40.402	0.997	89.711
G10BuddhaCave_4096x2048_fps30	G10BuddhaCave_ERP_4096x2048_fps30_qp27_1289k	65.164	2048	4096	44.319	0.997	88.013
G10XiaoGuang_4096x2048_fps30	G10XiaoGuang_ERP_4096x2048_fps30_qp27_2502k	70.666	2048	4096	40.622	0.999	92.191
G1AbandonedKingdom_7680x3840_fps30	G1AbandonedKingdom_ERP_7680x3840_fps30_qp27_45406k	68.452	3840	7680	41.841	0.999	91.781
G1Aerial_7680x3840_fps25	G1Aerial_ERP_7680x3840_fps25_qp27_18646k	80.186	3840	7680	41.809	0.998	87.243
G2ForgottenBook_7680x3840_fps30	G2ForgottenBook_ERP_7680x3840_fps30_qp27_41804k	59.033	3840	7680	41.972	0.998	92.490
G2FormationPace_7680x3840_fps29.97	G2FormationPace_ERP_7680x3840_fps29.97_qp27_15300k	76.729	3840	7680	45.795	0.998	88.911
G3BackcountrySkiing_3840x1920_fps25	G3BackcountrySkiing_ERP_3840x1920_fps25_qp27_19131k	64.916	1920	3840	40.336	0.998	91.464
G3GetYoGurl_3840x1920_fps29.97	G3GetYoGurl_ERP_3840x1920_fps29.97_qp27_14177k	60.037	1920	3840	41.726	0.998	94.595
G4CliffsideMansion_7680x3840_fps30	G4CliffsideMansion_ERP_7680x3840_fps30_qp27_21044k	72.270	3840	7680	43.840	0.998	93.426
G5Neighborhood_3840x1920_fps23.976	G5Neighborhood_ERP_3840x1920_fps23.976_qp27_24718k	62.925	1920	3840	39.328	0.998	98.583
G5ResistMarch_3840x1920_fps29.97	G5ResistMarch_ERP_3840x1920_fps29.97_qp27_19898k	72.314	1920	3840	38.205	0.999	93.211
G6AngelFallsClimbing_7680x3840_fps29.97	G6AngelFallsClimbing_ERP_7680x3840_fps29.97_qp27_54581k	88.617	3840	7680	39.041	0.999	90.635
G7OrchestraOfSpheres_7680x3840_fps24	G7OrchestraOfSpheres_ERP_7680x3840_fps24_qp27_4824k	81.154	3840	7680	48.172	0.999	90.710
G8DivingWithSharks_7680x3840_fps29.97	G8DivingWithSharks_ERP_7680x3840_fps29.97_qp27_42160k	72.322	3840	7680	44.434	0.999	85.266
G8YourMan_7680x3840_fps29.97	G8YourMan_ERP_7680x3840_fps29.97_qp27_10412k	74.758	3840	7680	43.695	0.999	90.592
G9ConcertLive_4096x2048_fps30	G9ConcertLive_ERP_4096x2048_fps30_qp27_7102k	74.963	2048	4096	47.596	0.999	93.107
G9DrivingInCity_3840x1920_fps30	G9DrivingInCity_ERP_3840x1920_fps30_qp27_11315k	70.135	1920	3840	39.608	0.998	90.854
G4HachaWaterfall_3840x1920_fps29.97	G4HachaWaterfall_ERP_3840x1920_fps29.97_qp27_27203k	71.501	1920	3840	39.230	0.999	91.684
G7UcaimaWaterfall_3840x1920_fps29.97	G7UcaimaWaterfall_ERP_3840x1920_fps29.97_qp27_46307k	72.326	1920	3840	37.733	0.999	91.983
G10BoatInPark_4096x2048_fps30	G10BoatInPark_ERP_4096x2048_fps30_qp37_3270k	58.305	2048	4096	35.661	0.986	71.708
G10BuddhaCave_4096x2048_fps30	G10BuddhaCave_ERP_4096x2048_fps30_qp37_236k	34.698	2048	4096	40.105	0.990	69.059
G10XiaoGuang_4096x2048_fps30	G10XiaoGuang_ERP_4096x2048_fps30_qp37_644k	45.178	2048	4096	36.256	0.996	83.191
G1AbandonedKingdom_7680x3840_fps30	G1AbandonedKingdom_ERP_7680x3840_fps30_qp37_9283k	58.828	3840	7680	36.466	0.994	69.652
G1Aerial_7680x3840_fps25	G1Aerial_ERP_7680x3840_fps25_qp37_3307k	61.639	3840	7680	36.248	0.990	62.317
G2ForgottenBook_7680x3840_fps30	G2ForgottenBook_ERP_7680x3840_fps30_qp37_9445k	56.442	3840	7680	36.594	0.989	67.885
G2FormationPace_7680x3840_fps29.97	G2FormationPace_ERP_7680x3840_fps29.97_qp37_3823k	59.280	3840	7680	41.177	0.992	75.882
G3BackcountrySkiing_3840x1920_fps25	G3BackcountrySkiing_ERP_3840x1920_fps25_qp37_4153k	60.379	1920	3840	34.765	0.991	75.065
G3GetYoGurl_3840x1920_fps29.97	G3GetYoGurl_ERP_3840x1920_fps29.97_qp37_3395k	54.064	1920	3840	36.445	0.991	81.489
G4CliffsideMansion_7680x3840_fps30	G4CliffsideMansion_ERP_7680x3840_fps30_qp37_4698k	61.179	3840	7680	38.414	0.988	74.992
G5Neighborhood_3840x1920_fps23.976	G5Neighborhood_ERP_3840x1920_fps23.976_qp37_7721k	59.554	1920	3840	33.347	0.988	77.766
G5ResistMarch_3840x1920_fps29.97	G5ResistMarch_ERP_3840x1920_fps29.97_qp37_3644k	59.341	1920	3840	31.775	0.996	79.477
G6AngelFallsClimbing_7680x3840_fps29.97	G6AngelFallsClimbing_ERP_7680x3840_fps29.97_qp37_5981k	77.647	3840	7680	33.508	0.994	71.398
G7OrchestraOfSpheres_7680x3840_fps24	G7OrchestraOfSpheres_ERP_7680x3840_fps24_qp37_1330k	50.493	3840	7680	43.351	0.992	78.120
G8DivingWithSharks_7680x3840_fps29.97	G8DivingWithSharks_ERP_7680x3840_fps29.97_qp27_42160k	72.322	3840	7680	44.434	0.999	85.266
G8YourMan_7680x3840_fps29.97	G8YourMan_ERP_7680x3840_fps29.97_qp37_2519k	57.932	3840	7680	39.679	0.995	78.506
G9ConcertLive_4096x2048_fps30	G9ConcertLive_ERP_4096x2048_fps30_qp37_2256k	61.501	2048	4096	41.814	0.992	76.633
G9DrivingInCity_3840x1920_fps30	G9DrivingInCity_ERP_3840x1920_fps30_qp37_2350k	47.478	1920	3840	35.675	0.992	76.687
G4HachaWaterfall_3840x1920_fps29.97	G4HachaWaterfall_ERP_3840x1920_fps29.97_qp37_6897k	62.788	1920	3840	33.265	0.994	74.363
G7UcaimaWaterfall_3840x1920_fps29.97	G7UcaimaWaterfall_ERP_3840x1920_fps29.97_qp37_13551k	67.724	1920	3840	30.901	0.993	67.835
G10BoatInPark_4096x2048_fps30	G10BoatInPark_ERP_4096x2048_fps30_qp42_1507k	40.372	2048	4096	33.153	0.971	55.883
G10BuddhaCave_4096x2048_fps30	G10BuddhaCave_ERP_4096x2048_fps30_qp42_170k	20.475	2048	4096	36.452	0.941	44.491
G10XiaoGuang_4096x2048_fps30	G10XiaoGuang_ERP_4096x2048_fps30_qp42_393k	25.278	2048	4096	32.241	0.975	68.229
G1AbandonedKingdom_7680x3840_fps30	G1AbandonedKingdom_ERP_7680x3840_fps30_qp42_4140k	48.628	3840	7680	33.634	0.980	47.491
G1Aerial_7680x3840_fps25	G1Aerial_ERP_7680x3840_fps25_qp42_1216k	54.221	3840	7680	33.329	0.965	39.949
G2ForgottenBook_7680x3840_fps30	G2ForgottenBook_ERP_7680x3840_fps30_qp42_4374k	41.642	3840	7680	33.739	0.967	46.469
G2FormationPace_7680x3840_fps29.97	G2FormationPace_ERP_7680x3840_fps29.97_qp42_1749k	36.226	3840	7680	38.728	0.984	65.679
G3BackcountrySkiing_3840x1920_fps25	G3BackcountrySkiing_ERP_3840x1920_fps25_qp42_1608k	42.188	1920	3840	31.881	0.982	56.697
G3GetYoGurl_3840x1920_fps29.97	G3GetYoGurl_ERP_3840x1920_fps29.97_qp42_1688k	35.850	1920	3840	33.847	0.982	71.522
G4CliffsideMansion_7680x3840_fps30	G4CliffsideMansion_ERP_7680x3840_fps30_qp42_2239k	41.513	3840	7680	35.484	0.966	58.012
G5Neighborhood_3840x1920_fps23.976	G5Neighborhood_ERP_3840x1920_fps23.976_qp42_3894k	50.164	1920	3840	30.215	0.971	58.551
G5ResistMarch_3840x1920_fps29.97	G5ResistMarch_ERP_3840x1920_fps29.97_qp42_1484k	44.988	1920	3840	28.832	0.991	66.122
G6AngelFallsClimbing_7680x3840_fps29.97	G6AngelFallsClimbing_ERP_7680x3840_fps29.97_qp42_1663k	34.241	3840	7680	27.686	0.960	30.777
G7OrchestraOfSpheres_7680x3840_fps24	G7OrchestraOfSpheres_ERP_7680x3840_fps24_qp42_683k	22.367	3840	7680	39.933	0.976	66.186
G8DivingWithSharks_7680x3840_fps29.97	G8DivingWithSharks_ERP_7680x3840_fps29.97_qp42_5645k	49.405	3840	7680	36.352	0.986	43.541
G8YourMan_7680x3840_fps29.97	G8YourMan_ERP_7680x3840_fps29.97_qp42_1193k	34.826	3840	7680	36.452	0.985	66.267
G9ConcertLive_4096x2048_fps30	G9ConcertLive_ERP_4096x2048_fps30_qp42_1246k	43.130	2048	4096	38.627	0.980	60.741
G9DrivingInCity_3840x1920_fps30	G9DrivingInCity_ERP_3840x1920_fps30_qp42_1069k	33.589	1920	3840	33.132	0.983	61.944
G4HachaWaterfall_3840x1920_fps29.97	G4HachaWaterfall_ERP_3840x1920_fps29.97_qp42_2776k	60.077	1920	3840	30.687	0.985	59.734
G7UcaimaWaterfall_3840x1920_fps29.97	G7UcaimaWaterfall_ERP_3840x1920_fps29.97_qp42_5233k	56.358	1920	3840	27.997	0.982	47.292

Table 6.3: Full information table for BMS360 with saliency inclusion.

Reference Video	Distorted Video	Mos	height	width	psnr	mssim	vmaf
G1BajaCalifornia_3840x2160_fps23.976_preprocess	G1BajaCalifornia_ERP_3840x2160_fps23.976_qp27_41186k_preprocess	64.942	2160	3840	46.399	0.999	95.550
G1BajaCalifornia_3840x2160_fps23.976_preprocess	G1BajaCalifornia_ERP_3840x2160_fps23.976_qp37_12633k_preprocess	60.041	2160	3840	40.371	0.993	81.456
G1BajaCalifornia_3840x2160_fps23.976_preprocess	G1BajaCalifornia_ERP_3840x2160_fps23.976_qp42_5570k_preprocess	47.554	2160	3840	37.607	0.982	68.394
G1BikingToWork_3840x2160_fps23.976_preprocess	G1BikingToWork_ERP_3840x2160_fps23.976_qp27_12306k_preprocess	55.828	2160	3840	47.686	0.999	96.595
G1BikingToWork_3840x2160_fps23.976_preprocess	G1BikingToWork_ERP_3840x2160_fps23.976_qp37_3596k_preprocess	54.019	2160	3840	42.729	0.995	83.870
G1BikingToWork_3840x2160_fps23.976_preprocess	G1BikingToWork_ERP_3840x2160_fps23.976_qp42_1855k_preprocess	36.841	2160	3840	39.830	0.989	70.525
G2AstonVillaGoal_3840x2048_fps24_preprocess	G2AstonVillaGoal_ERP_3840x2048_fps24_qp27_11596k_preprocess	72.924	2048	3840	44.531	0.999	93.681
G2AstonVillaGoal_3840x2048_fps24_preprocess	G2AstonVillaGoal_ERP_3840x2048_fps24_qp37_2141k_preprocess	47.285	2048	3840	39.350	0.997	83.203
G2AstonVillaGoal_3840x2048_fps24_preprocess	G2AstonVillaGoal_ERP_3840x2048_fps24_qp42_872k_preprocess	25.691	2048	3840	36.603	0.992	71.703
G3BackcountrySkiing_3840x1920_fps25_preprocess	G3BackcountrySkiing_ERP_3840x1920_fps25_qp27_19131k_preprocess	68.007	1920	3840	43.918	0.999	94.390
G3BackcountrySkiing_3840x1920_fps25_preprocess	G3BackcountrySkiing_ERP_3840x1920_fps25_qp37_4153k_preprocess	63.254	1920	3840	38.283	0.998	81.577
G3BackcountrySkiing_3840x1920_fps25_preprocess	G3BackcountrySkiing_ERP_3840x1920_fps25_qp42_1608k_preprocess	44.197	1920	3840	35.397	0.995	66.433
G3GetYoGurl_3840x1920_fps29.97_preprocess	G3GetYoGurl_ERP_3840x1920_fps29.97_qp27_14177k_preprocess	62.896	1920	3840	46.040	0.999	95.331
G3GetYoGurl_3840x1920_fps29.97_preprocess	G3GetYoGurl_ERP_3840x1920_fps29.97_qp37_3395k_preprocess	56.638	1920	3840	40.992	0.996	85.604
G3GetYoGurl_3840x1920_fps29.97_preprocess	G3GetYoGurl_ERP_3840x1920_fps29.97_qp42_1688k_preprocess	37.557	1920	3840	38.391	0.991	76.690
G2ForgottenBook_7680x3840_fps30_preprocess	G2ForgottenBook_ERP_7680x3840_fps30_qp27_41804k_preprocess	59.033	3840	7680	45.793	0.999	93.406
G4WingsuitFlight_3840x2048_fps29.97_preprocess	G4WingsuitFlight_ERP_3840x2048_fps29.97_qp27_12696k_preprocess	56.289	2048	3840	46.293	0.999	95.926
G5EbinShader_7168x3584_fps30_preprocess	G5EbinShader_ERP_7168x3584_fps30_qp27_31040k_preprocess	70.924	3584	7168	45.641	1.000	90.652
G5Neighborhood_3840x1920_fps23.976_preprocess	G5Neighborhood_ERP_3840x1920_fps23.976_qp27_24718k_preprocess	62.925	1920	3840	43.634	0.999	97.222
G6DragonTale_3840x2160_fps30_preprocess	G6DragonTale_ERP_3840x2160_fps30_qp27_13868k_preprocess	72.102	2160	3840	46.007	0.999	94.477
G6GTRDriving_3840x2160_fps30_preprocess	G6GTRDriving_ERP_3840x2160_fps30_qp27_28748k_preprocess	68.049	2160	3840	46.119	1.000	98.063
G7DragonCastleAttack_3840x2048_fps24_preprocess	G7DragonCastleAttack_ERP_3840x2048_fps24_qp27_15133k_preprocess	77.435	2048	3840	46.525	0.999	95.918
G7PressConference_4096x2048_fps30_preprocess	G7PressConference_ERP_4096x2048_fps30_qp27_4909k_preprocess	65.319	2048	4096	46.124	0.999	91.850
G8AlpsParagliding_3840x1920_fps25_preprocess	G8AlpsParagliding_ERP_3840x1920_fps25_qp27_8437k_preprocess	72.284	1920	3840	45.921	0.999	93.279
G8ANewEmpire_3840x2048_fps29.97_preprocess	G8ANewEmpire_ERP_3840x2048_fps29.97_qp27_33327k_preprocess	68.867	2048	3840	43.429	0.999	99.327
G9ConcertLive_4096x2048_fps30_preprocess	G9ConcertLive_ERP_4096x2048_fps30_qp27_7102k_preprocess	74.963	2048	4096	50.979	0.999	93.614
G10BoatInPark_4096x2048_fps30_preprocess	G10BoatInPark_ERP_4096x2048_fps30_qp27_14547k_preprocess	68.308	2048	4096	46.068	0.999	92.672
G10BuddhaCave_4096x2048_fps30_preprocess	G10BuddhaCave_ERP_4096x2048_fps30_qp27_1289k_preprocess	65.164	2048	4096	49.897	0.999	90.696
G10XiaoGuang_4096x2048_fps30_preprocess	G10XiaoGuang_ERP_4096x2048_fps30_qp27_2502k_preprocess	70.666	2048	4096	44.266	0.999	93.564
G2ForgottenBook_7680x3840_fps30_preprocess	G2ForgottenBook_ERP_7680x3840_fps30_qp37_9445k_preprocess	56.442	3840	7680	40.779	0.995	75.925
G4WingsuitFlight_3840x2048_fps29.97_preprocess	G4WingsuitFlight_ERP_3840x2048_fps29.97_qp37_3545k_preprocess	57.062	2048	3840	41.244	0.997	84.392
G5EbinShader_7168x3584_fps30_preprocess	G5EbinShader_ERP_7168x3584_fps30_qp37_4094k_preprocess	71.059	3584	7168	40.563	0.998	74.898
G5Neighborhood_3840x1920_fps23.976_preprocess	G5Neighborhood_ERP_3840x1920_fps23.976_qp37_7721k_preprocess	59.554	1920	3840	37.815	0.995	81.223
G6DragonTale_3840x2160_fps30_preprocess	G6DragonTale_ERP_3840x2160_fps30_qp37_3984k_preprocess	61.631	2160	3840	40.612	0.996	78.450
G6GTRDriving_3840x2160_fps30_preprocess	G6GTRDriving_ERP_3840x2160_fps30_qp37_8912k_preprocess	62.608	2160	3840	40.303	0.998	88.395
G7DragonCastleAttack_3840x2048_fps24_preprocess	G7DragonCastleAttack_ERP_3840x2048_fps24_qp37_4272k_preprocess	62.215	2048	3840	41.117	0.996	81.557
G7PressConference_4096x2048_fps30_preprocess	G7PressConference_ERP_4096x2048_fps30_qp37_1038k_preprocess	43.919	2048	4096	42.001	0.997	80.419
G8AlpsParagliding_3840x1920_fps25_preprocess	G8AlpsParagliding_ERP_3840x1920_fps25_qp37_1995k_preprocess	52.499	1920	3840	40.580	0.996	79.194
G8ANewEmpire_3840x2048_fps29.97_preprocess	G8ANewEmpire_ERP_3840x2048_fps29.97_qp37_10885k_preprocess	61.291	2048	3840	37.356	0.995	82.878
G9ConcertLive_4096x2048_fps30_preprocess	G9ConcertLive_ERP_4096x2048_fps30_qp37_2256k_preprocess	61.501	2048	4096	45.245	0.997	80.484
G10BoatInPark_4096x2048_fps30_preprocess	G10BoatInPark_ERP_4096x2048_fps30_qp37_3270k_preprocess	58.305	2048	4096	41.495	0.996	82.017
G10BuddhaCave_4096x2048_fps30_preprocess	G10BuddhaCave_ERP_4096x2048_fps30_qp37_236k_preprocess	34.698	2048	4096	44.705	0.998	74.275
G10XiaoGuang_4096x2048_fps30_preprocess	G10XiaoGuang_ERP_4096x2048_fps30_qp37_644k_preprocess	45.178	2048	4096	39.467	0.997	84.684
G2ForgottenBook_7680x3840_fps30_preprocess	G2ForgottenBook_ERP_7680x3840_fps30_qp42_4374k_preprocess	41.642	3840	7680	38.044	0.986	57.704
G4WingsuitFlight_3840x2048_fps29.97_preprocess	G4WingsuitFlight_ERP_3840x2048_fps29.97_qp42_1877k_preprocess	41.085	2048	3840	38.398	0.993	71.732
G5EbinShader_7168x3584_fps30_preprocess	G5EbinShader_ERP_7168x3584_fps30_qp42_1383k_preprocess	56.605	3584	7168	37.822	0.993	53.884
G5Neighborhood_3840x1920_fps23.976_preprocess	G5Neighborhood_ERP_3840x1920_fps23.976_qp42_3894k_preprocess	50.164	1920	3840	34.752	0.987	64.662
G6DragonTale_3840x2160_fps30_preprocess	G6DragonTale_ERP_3840x2160_fps30_qp42_2057k_preprocess	38.656	2160	3840	37.772	0.990	60.930
G6GTRDriving_3840x2160_fps30_preprocess	G6GTRDriving_ERP_3840x2160_fps30_qp42_4245k_preprocess	52.463	2160	3840	37.538	0.995	78.522
G7DragonCastleAttack_3840x2048_fps24_preprocess	G7DragonCastleAttack_ERP_3840x2048_fps24_qp42_2084k_preprocess	46.729	2048	3840	38.345	0.990	65.862
G7PressConference_4096x2048_fps30_preprocess	G7PressConference_ERP_4096x2048_fps30_qp42_509k_preprocess	22.707	2048	4096	38.719	0.991	67.071
G8AlpsParagliding_3840x1920_fps25_preprocess	G8AlpsParagliding_ERP_3840x1920_fps25_qp42_964k_preprocess	37.148	1920	3840	37.502	0.990	63.081
G8ANewEmpire_3840x2048_fps29.97_preprocess	G8ANewEmpire_ERP_3840x2048_fps29.97_qp42_5571k_preprocess	50.376	2048	3840	34.362	0.986	65.558
G9ConcertLive_4096x2048_fps30_preprocess	G9ConcertLive_ERP_4096x2048_fps30_qp42_1246k_preprocess	43.130	2048	4096	42.100	0.992	66.679
G10BoatInPark_4096x2048_fps30_preprocess	G10BoatInPark_ERP_4096x2048_fps30_qp42_1507k_preprocess	40.372	2048	4096	39.018	0.992	70.791
G10BuddhaCave_4096x2048_fps30_preprocess	G10BuddhaCave_ERP_4096x2048_fps30_qp42_170k_preprocess	20.475	2048	4096	40.465	0.990	50.016
G10XiaoGuang_4096x2048_fps30_preprocess	G10XiaoGuang_ERP_4096x2048_fps30_qp42_393k_preprocess	25.278	2048	4096	35.335	0.984	68.685

Table 6.4: Full information table for Cubepadding with saliency inclusion.

Reference Video	Distorted Video	Mos	height	width	psnr	msssim	vmaf
G10BoatInPark_4096x2048_fps30_preprocess	G10BoatInPark_ERP_4096x2048_fps30_qp27_14547k_preprocess	68.308	2048	4096	43.745	0.999	92.061
G10BuddhaCave_4096x2048_fps30_preprocess	G10BuddhaCave_ERP_4096x2048_fps30_qp27_1289k_preprocess	65.164	2048	4096	47.405	0.999	89.868
G10XiaoGuang_4096x2048_fps30_preprocess	G10XiaoGuang_ERP_4096x2048_fps30_qp27_2502k_preprocess	70.666	2048	4096	44.133	0.999	93.337
G1AbandonedKingdom_7680x3840_fps30_preprocess	G1AbandonedKingdom_ERP_7680x3840_fps30_qp27_45406k_preprocess	68.452	3840	7680	45.368	0.999	94.901
G1Aerial_7680x3840_fps25_preprocess	G1Aerial_ERP_7680x3840_fps25_qp27_18646k_preprocess	80.186	3840	7680	46.111	0.999	91.746
G2ForgottenBook_7680x3840_fps30_preprocess	G2ForgottenBook_ERP_7680x3840_fps30_qp27_41804k_preprocess	59.033	3840	7680	45.029	0.999	94.342
G2FormationPace_7680x3840_fps29.97_preprocess	G2FormationPace_ERP_7680x3840_fps29.97_qp27_15300k_preprocess	76.729	3840	7680	48.271	0.999	91.968
G3BackcountrySkiing_3840x1920_fps25_preprocess	G3BackcountrySkiing_ERP_3840x1920_fps25_qp27_19131k_preprocess	64.916	1920	3840	43.974	0.999	96.179
G3GetYoGurl_3840x1920_fps29.97_preprocess	G3GetYoGurl_ERP_3840x1920_fps29.97_qp27_14177k_preprocess	60.037	1920	3840	44.606	0.999	96.389
G4CliffsideMansion_7680x3840_fps30_preprocess	G4CliffsideMansion_ERP_7680x3840_fps30_qp27_21044k_preprocess	72.270	3840	7680	46.924	0.999	95.918
G5Neighborhood_3840x1920_fps23.976_preprocess	G5Neighborhood_ERP_3840x1920_fps23.976_qp27_24718k_preprocess	62.925	1920	3840	43.364	0.999	98.906
G5ResistMarch_3840x1920_fps29.97_preprocess	G5ResistMarch_ERP_3840x1920_fps29.97_qp27_19898k_preprocess	72.314	1920	3840	41.311	1.000	95.005
G6AngelFallsClimbing_7680x3840_fps29.97_preprocess	G6AngelFallsClimbing_ERP_7680x3840_fps29.97_qp27_54581k_preprocess	88.617	3840	7680	42.339	0.999	92.676
G7OrchestraOfSpheres_7680x3840_fps24_preprocess	G7OrchestraOfSpheres_ERP_7680x3840_fps24_qp27_4824k_preprocess	81.154	3840	7680	50.276	0.999	92.282
G8DivingWithSharks_7680x3840_fps29.97_preprocess	G8DivingWithSharks_ERP_7680x3840_fps29.97_qp27_42160k_preprocess	72.322	3840	7680	40.173	1.000	92.132
G8YourMan_7680x3840_fps29.97_preprocess	G8YourMan_ERP_7680x3840_fps29.97_qp27_10412k_preprocess	74.758	3840	7680	45.699	0.999	92.310
G9ConcertLive_4096x2048_fps30_preprocess	G9ConcertLive_ERP_4096x2048_fps30_qp27_7102k_preprocess	74.963	2048	4096	50.630	0.999	93.878
G9DrivingInCity_3840x1920_fps30_preprocess	G9DrivingInCity_ERP_3840x1920_fps30_qp27_11315k_preprocess	70.135	1920	3840	43.034	0.999	93.680
G4HahaWaterfall_3840x1920_fps29.97_preprocess	G4HahaWaterfall_ERP_3840x1920_fps29.97_qp27_27203k_preprocess	71.501	1920	3840	43.055	0.999	94.052
G7UcainaWaterfall_3840x1920_fps29.97_preprocess	G7UcainaWaterfall_ERP_3840x1920_fps29.97_qp27_46307k_preprocess	72.326	1920	3840	41.330	0.999	93.752
G10BoatInPark_4096x2048_fps30_preprocess	G10BoatInPark_ERP_4096x2048_fps30_qp37_3270k_preprocess	58.305	2048	4096	39.034	0.994	78.515
G10BuddhaCave_4096x2048_fps30_preprocess	G10BuddhaCave_ERP_4096x2048_fps30_qp37_236k_preprocess	34.698	2048	4096	42.662	0.996	71.620
G10XiaoGuang_4096x2048_fps30_preprocess	G10XiaoGuang_ERP_4096x2048_fps30_qp37_644k_preprocess	45.178	2048	4096	39.501	0.997	84.635
G1AbandonedKingdom_7680x3840_fps30_preprocess	G1AbandonedKingdom_ERP_7680x3840_fps30_qp37_9283k_preprocess	58.828	3840	7680	40.232	0.996	78.298
G1Aerial_7680x3840_fps25_preprocess	G1Aerial_ERP_7680x3840_fps25_qp37_3307k_preprocess	61.639	3840	7680	40.983	0.996	74.717
G2ForgottenBook_7680x3840_fps30_preprocess	G2ForgottenBook_ERP_7680x3840_fps30_qp37_9445k_preprocess	56.442	3840	7680	39.296	0.993	75.075
G2FormationPace_7680x3840_fps29.97_preprocess	G2FormationPace_ERP_7680x3840_fps29.97_qp37_3823k_preprocess	59.280	3840	7680	43.444	0.996	81.879
G3BackcountrySkiing_3840x1920_fps25_preprocess	G3BackcountrySkiing_ERP_3840x1920_fps25_qp37_4153k_preprocess	60.379	1920	3840	38.475	0.997	82.820
G3GetYoGurl_3840x1920_fps29.97_preprocess	G3GetYoGurl_ERP_3840x1920_fps29.97_qp37_3395k_preprocess	54.064	1920	3840	39.521	0.995	86.257
G4CliffsideMansion_7680x3840_fps30_preprocess	G4CliffsideMansion_ERP_7680x3840_fps30_qp37_4698k_preprocess	61.179	3840	7680	41.834	0.994	83.656
G5Neighborhood_3840x1920_fps23.976_preprocess	G5Neighborhood_ERP_3840x1920_fps23.976_qp37_7721k_preprocess	59.554	1920	3840	37.632	0.993	83.100
G5ResistMarch_3840x1920_fps29.97_preprocess	G5ResistMarch_ERP_3840x1920_fps29.97_qp37_3644k_preprocess	59.341	1920	3840	35.020	0.997	83.600
G6AngelFallsClimbing_7680x3840_fps29.97_preprocess	G6AngelFallsClimbing_ERP_7680x3840_fps29.97_qp37_5981k_preprocess	77.647	3840	7680	36.970	0.997	77.744
G7OrchestraOfSpheres_7680x3840_fps24_preprocess	G7OrchestraOfSpheres_ERP_7680x3840_fps24_qp37_1330k_preprocess	50.493	3840	7680	45.562	0.996	81.545
G8DivingWithSharks_7680x3840_fps29.97_preprocess	G8DivingWithSharks_ERP_7680x3840_fps29.97_qp27_42160k_preprocess	72.322	3840	7680	40.173	1.000	92.132
G8YourMan_7680x3840_fps29.97_preprocess	G8YourMan_ERP_7680x3840_fps29.97_qp37_2519k_preprocess	57.932	3840	7680	41.816	0.997	81.446
G9ConcertLive_4096x2048_fps30_preprocess	G9ConcertLive_ERP_4096x2048_fps30_qp37_2256k_preprocess	61.501	2048	4096	44.997	0.996	80.652
G9DrivingInCity_3840x1920_fps30_preprocess	G9DrivingInCity_ERP_3840x1920_fps30_qp37_2350k_preprocess	47.478	1920	3840	30.083	0.996	82.337
G4HahaWaterfall_3840x1920_fps29.97_preprocess	G4HahaWaterfall_ERP_3840x1920_fps29.97_qp37_6897k_preprocess	62.788	1920	3840	37.380	0.996	81.033
G7UcainaWaterfall_3840x1920_fps29.97_preprocess	G7UcainaWaterfall_ERP_3840x1920_fps29.97_qp37_13551k_preprocess	67.724	1920	3840	34.586	0.995	75.401
G10BoatInPark_4096x2048_fps30_preprocess	G10BoatInPark_ERP_4096x2048_fps30_qp42_1507k_preprocess	40.372	2048	4096	36.489	0.986	65.113
G10BuddhaCave_4096x2048_fps30_preprocess	G10BuddhaCave_ERP_4096x2048_fps30_qp42_170k_preprocess	20.475	2048	4096	38.673	0.981	44.700
G10XiaoGuang_4096x2048_fps30_preprocess	G10XiaoGuang_ERP_4096x2048_fps30_qp42_393k_preprocess	25.278	2048	4096	35.444	0.984	68.648
G1AbandonedKingdom_7680x3840_fps30_preprocess	G1AbandonedKingdom_ERP_7680x3840_fps30_qp42_4140k_preprocess	48.628	3840	7680	37.388	0.987	59.178
G1Aerial_7680x3840_fps25_preprocess	G1Aerial_ERP_7680x3840_fps25_qp42_1216k_preprocess	54.221	3840	7680	38.269	0.985	56.183
G2ForgottenBook_7680x3840_fps30_preprocess	G2ForgottenBook_ERP_7680x3840_fps30_qp42_4374k_preprocess	41.642	3840	7680	37.131	0.981	55.634
G2FormationPace_7680x3840_fps29.97_preprocess	G2FormationPace_ERP_7680x3840_fps29.97_qp42_1749k_preprocess	36.226	3840	7680	41.442	0.992	73.385
G3BackcountrySkiing_3840x1920_fps25_preprocess	G3BackcountrySkiing_ERP_3840x1920_fps25_qp42_1608k_preprocess	42.188	1920	3840	35.565	0.993	67.882
G3GetYoGurl_3840x1920_fps29.97_preprocess	G3GetYoGurl_ERP_3840x1920_fps29.97_qp42_1688k_preprocess	35.850	1920	3840	36.949	0.989	67.385
G4CliffsideMansion_7680x3840_fps30_preprocess	G4CliffsideMansion_ERP_7680x3840_fps30_qp42_2230k_preprocess	41.513	3840	7680	39.041	0.984	70.357
G5Neighborhood_3840x1920_fps23.976_preprocess	G5Neighborhood_ERP_3840x1920_fps23.976_qp42_3894k_preprocess	50.164	1920	3840	34.560	0.983	66.375
G5ResistMarch_3840x1920_fps29.97_preprocess	G5ResistMarch_ERP_3840x1920_fps29.97_qp42_1484k_preprocess	44.988	1920	3840	32.077	0.994	71.382
G6AngelFallsClimbing_7680x3840_fps29.97_preprocess	G6AngelFallsClimbing_ERP_7680x3840_fps29.97_qp42_1663k_preprocess	34.241	3840	7680	30.861	0.975	41.264
G7OrchestraOfSpheres_7680x3840_fps24_preprocess	G7OrchestraOfSpheres_ERP_7680x3840_fps24_qp42_683k_preprocess	22.367	3840	7680	42.243	0.988	70.820
G8DivingWithSharks_7680x3840_fps29.97_preprocess	G8DivingWithSharks_ERP_7680x3840_fps29.97_qp42_5645k_preprocess	49.405	3840	7680	41.447	0.994	63.060
G8YourMan_7680x3840_fps29.97_preprocess	G8YourMan_ERP_7680x3840_fps29.97_qp42_1193k_preprocess	34.826	3840	7680	38.585	0.989	69.119
G9ConcertLive_4096x2048_fps30_preprocess	G9ConcertLive_ERP_4096x2048_fps30_qp42_1246k_preprocess	43.130	2048	4096	41.843	0.989	66.591
G9DrivingInCity_3840x1920_fps30_preprocess	G9DrivingInCity_ERP_3840x1920_fps30_qp42_1069k_preprocess	33.589	1920	3840	36.503	0.991	69.482
G4HahaWaterfall_3840x1920_fps29.97_preprocess	G4HahaWaterfall_ERP_3840x1920_fps29.97_qp42_2776k_preprocess	60.077	1920	3840	34.859	0.990	68.554
G7UcainaWaterfall_3840x1920_fps29.97_preprocess	G7UcainaWaterfall_ERP_3840x1920_fps29.97_qp42_5233k_preprocess	56.358	1920	3840	31.686	0.987	57.757