Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Dissertation

Presented to the University of Dublin, Trinity College

in fulfilment of the requirements for the Degree of

Doctor of Philosohpy in Computer Science

July 2022

Algorithms for Quality Optimization in Omnidirectional Video

Simone Maurizio Croci

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Simone Maurizio Croci

July 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Simone Maurizio Croci

July 19, 2022

To my parents.

Acknowledgments

First of all, I need to thank Prof. Aljosa Smolic, who gave me the opportunity to work in the V-SENSE team. Next, I must thank all the colleagues with whom I collaborated in my research here listed in alphabetical order: Julian Cabrera, Roman Dudek, Mairead Grogan, Sebastian Knorr, Cagri Ozcinar, and Emin Zerman. I also need to thank Gail Weadick for her help sorting out the numerous bureaucratic issues. Special thanks also to Alberto Foglia. And finally, I thank my parents for their encouragement.

Simone Maurizio Croci

University of Dublin, Trinity College
July 2022

Algorithms for Quality Optimization in Omnidirectional Video

Publication No. ____________________

Simone Maurizio Croci, PhD
University of Dublin, Trinity College, 2022

Supervisor: Smolic, Aljosa

Omnidirectional video (ODV) is a recent imaging technology, which is currently getting increasingly popular thanks to its ability to create an immersive and interactive viewing experience. Nowadays, viewing ODV is becoming affordable and easy, thanks to relatively cheap head-mounted displays and the possibility to view ODV with widespread devices like smartphones and tablets. Moreover, the ODV technology has reached a level that is mature enough to attract content producers.

Even if the ODV technology has reached an adequate level of maturity, it is still developing. There are several challenges during ODV production that can introduce artifacts in ODV even with the current technology. Moreover, for the coding and transmission of ODV, the current solutions could be further developed and improved.

Therefore, the objective of this thesis is to develop technologies that can help to improve the quality of ODV and be applied during ODV production and for the development of better coding and transmission methods. Three research areas characterize this thesis:

artifact detection, artifact correction, and quality assessment.

Regarding artifact detection, we propose a general framework that extends artifact detection methods for stereoscopic 3D (S3D) standard images to S3D omnidirectional images (ODI), which can be used by artists in the post-production workflow in order to optimize the quality. Moreover, methods for the detection of two common artifacts in S3D ODIs, namely color and sharpness mismatch, are also proposed. For the evaluation of the artifact detection framework, a new dataset of S3D ODIs with visual attention data was created.

For the second research area, *i.e.*, artifact correction, the thesis proposes three different solutions for the correction of color mismatch. One solution is based on traditional visual computing techniques, and the other two solutions are based on deep learning. The evaluation of these methods shows their effectiveness.

Finally, for quality assessment, a framework that extends full-reference quality metrics for standard video to monoscopic ODV was developed. For the development and evaluation of the framework, a dataset of monoscopic ODVs with subjective quality scores and visual attention data was created. The evaluation of the framework shows that it has a better quality assessment performance than the commonly used quality metrics for monoscopic ODV.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Historical Context

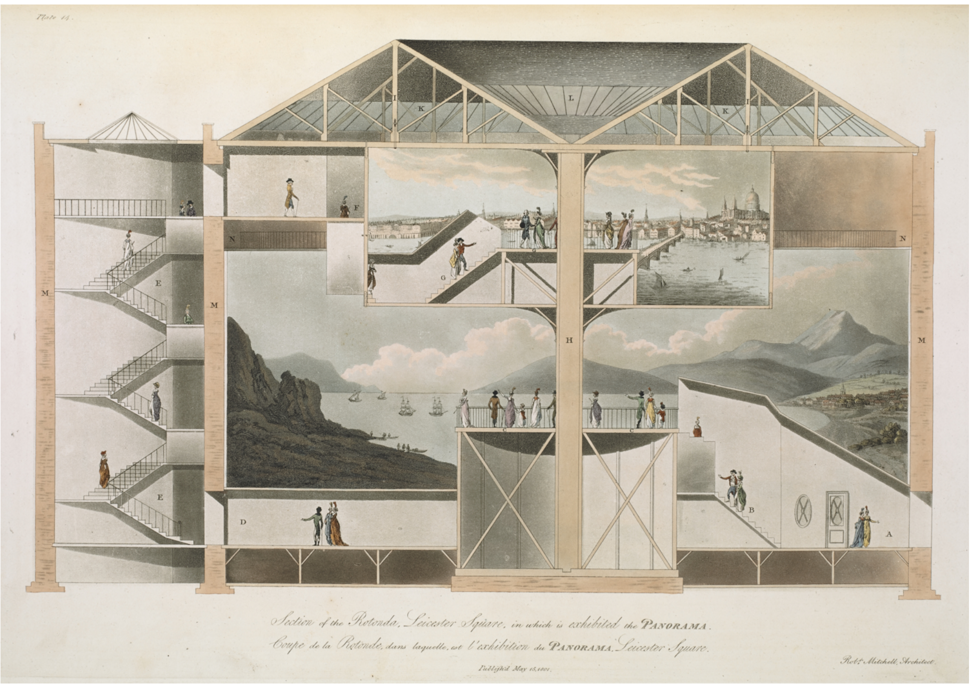


Plate 14.
K
L
K
N
F
G
N
M
E
M
H
M
E
D
B
A
Section of the Rotunda, Leicester Square, in which is exhibited the PANORAMA.
Coupe de la Rotonde, dans laquelle, est l'exhibition du PANORAMA, Leicester Square.
Publish'd May 15, 1801.
Robt. Mitchell, Architect.

Figure 1.1: Section of the Rotunda building in London's Leicester Square built for the exhibition of panoramas by Robert Barker. Figure taken from [4].

One of the first attempts to create a panoramic view dates back to 1519, when the artist Baldassarre Peruzzi painted the walls of the Sala delle Prospettive in the Villa Farnesina depicting a virtual space with colonnades and distant views of Rome [10]. The word *panorama*, from the Greek words *pan* meaning "all", and *horama* meaning "view", was invented by Robert Barker in 1787 [4]. He also became famous for his panorama paintings that he exhibited in the Rotunda in Leicester Square in London (Figure 1.1) [11, 12], which was a purpose-built building for displaying panoramas. Panoramic photography quickly followed the invention of photography in 1839 [13]. In 1843, Joseph Puchberger patented a 150-degree panoramic camera, while the first mass-produced panoramic camera, Al-Vista, was released in 1898. The introduction of digital photography in the 1980s made the creation of panoramic photographs easier. But, only in the late 1990s, panoramic photography started to become popular thanks to the success of digital cameras and improved editing software. In the early 2000s, the first rudimentary ODVs were made and their popularity slowly increased throughout the decade. Since about 2015, ODV has seen a marked increase in popularity [14]. In this year, YouTube started to allow ODV playback, and in 2016 ODV uploads. Also Facebook launched the possibility to view ODV on its social platform in late 2015. Moreover, in 2016, several important news outlets like the New York Times, the Guardian, and BBC, started making ODV content regularly for their news coverage. In 2016, even the Olympics in Rio de Janeiro were live-streamed in ODV [4].

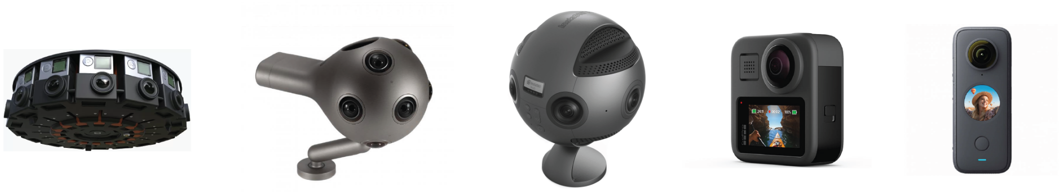


Figure 1.2: Omnidirectional video cameras.

Figure 1.3: Head-mounted displays.

The time of ODV seems to have come. This claim is based on two points: affordable and capable production technology and ease of viewing [14]. Regarding the production, in comparison to the early days of ODV, the cameras and editing tools are now much cheaper. For the acquisition of ODV, in addition to professional cameras like Kandao Obsidian and Insta360 Titan, there are also cheap consumer cameras like GoPro Max and Insta360 One X2, just to name a few (Figure 1.2). It is important to mention, that the production technology is still developing, but it has already reached a level of maturity stable enough to attract content producers. Regarding the ease of viewing, as it was mentioned before, content sharing platforms like YouTube and Facebook together with different media outlets have made consuming ODV commonplace and accessible from different devices. Moreover, the viewing of ODV no longer requires costly headsets. The so-called head-mounted displays (HMD) are becoming increasingly popular and cheaper (Figure 1.3). Examples of commercially successful HMDs are Oculus Quest 2, Sony PlayStation VR, HTC Vive Cosmos, etc.

1.2 Omnidirectional Video

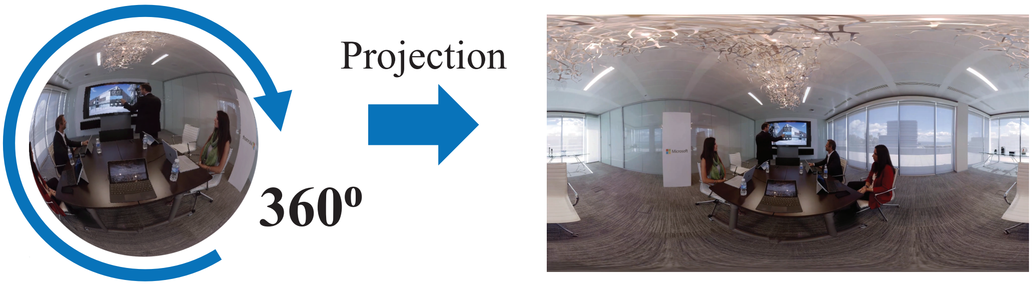


Figure 1.4: Example of omnidirectional video.

Omnidirectional video, also referred to as 360-degree video (Figure 1.4), can be conceived as a spherical video where the viewers are placed at its center allowing them to look in every direction, differently from standard video where just a small portion of the 360-degree view is presented. ODV is intended to be viewed with an HMD that shows only the content in the direction where the viewer is looking at. In contrast to traditional video, ODV provides a higher immersive and interactive viewing experience. Different from virtual reality (VR), which presents mainly computer-generated environments to the viewer, ODV is obtained by capturing live action in the real world.

Thanks to its immersive nature, ODV can be exploited for different applications. For example, ODV can be used in entertainment [15, 16], like for movies [17], the live streaming of sports and cultural events [18], and also computer games [19]. Other applications include communication [20], *i.e.*, ODV conferencing, health care [21], and education [22]. Regarding the latter, it has been noticed that students learn better and recall more information when they are provided with rich immersive media [4].

1.3 Challenges and Problems

The production of ODV obtained by stitching and blending together multiple views is not an easy task, as there are many technical challenges to overcome, especially when capturing and post-processing in stereoscopic 3D (S3D) [23]. Some challenges are the physical limitations of the chosen capture system, improper camera alignment, errors in post-production or compositing, etc [24, 25]. These challenges usually introduce artifacts that can create visual discomfort and consequently degrade the quality of experience (QoE) [26] when watching ODV with an HMD [27, 28, 29]. Therefore, the detection and correction of these artifacts are very important. There are already different detection and correction methods for traditional video [30, 31, 32], which usually need to be adapted to ODV, but there are not so many methods specifically designed for ODV.

Besides the production, also the coding and transmission of ODV represent challenges that can be a source of quality degradation. For example, due to the large field of view of ODV [33], higher video resolution is necessary, and consequently, also higher memory requirements are demanded. For the development and evaluation of new solutions to these technical challenges, like new compression and streaming approaches [20], subjective and in particular objective quality assessment methods are necessary to ensure a high QoE. Even though there are already quality metrics for ODV like [34, 35, 36, 37, 38, 39], these metrics have a limited correlation with subjective quality perception.

In general, methods for traditional video cannot be directly applied to ODV. There are two unique aspects of ODV not present in standard video that must be taken into account when dealing with ODV. First, ODV is a spherical signal, but it is usually stored in planar 2D formats. The mapping between spherical and planar representation inevitably introduces spherical projection distortions not present in standard video. Furthermore, planar representations have discontinuities/borders not present in the spherical representation. Second, the field of view of ODV is much larger than traditional video, but only a part of ODV can be viewed with an HMD. In [37], it was found that less than 65% of the ODV area is watched by the viewers. Therefore, it is important to consider the viewers' behavior while exploring ODV with an HMD [40, 41, 42], and to identify the ODV regions that are most likely viewed [43, 44, 45, 46, 15], especially when assessing quality.

1.4 Research Question

Considering the challenges and problems related to ODV, and the state-of-the-art methods for the analysis and processing of ODV, we identified three research areas of relevance for the development and success of ODV: detection of artifacts that are introduced during acquisition and post-production, correction of these artifacts, and quality evaluation. For each research area, we defined the following research objectives:

Artifact Detection

Develop methods for the detection and localization of artifacts that can be used by artists.

Artifact Correction

Develop methods for the correction of common artifacts that can improve the quality and reduce time and efforts in the production workflow.

Quality Assessment

Develop quality metrics that can be used for the development and optimization of coding and transmission solutions.

The research question of the thesis, which embraces all the research objectives, is formulated as follows:

Research Question

How to optimize the quality of omnidirectional video?

1.5 Contributions

The thesis contributions that realize the research objectives are listed here:

Artifact Detection

1. **A general artifact detection framework for stereoscopic 3D omnidirectional images** that extends artifact detection methods for S3D standard images to S3D omnidirectional images (ODI) based on the subdivision of the ODIs into planar Voronoi patches, the integration of visual attention, and the artifact visualization.

2. **Two methods for the detection of sharpness mismatch in stereoscopic 3D images:** the first is based on edge contrast and width histograms, and the second on psychophysical data from an experiment about the perception of sharpness mismatch.

3. **A method for the detection of color mismatch in stereoscopic 3D images** based on the difference between color statistics extracted from the two stereoscopic views.

4. **A dataset with stereoscopic 3D omnidirectional images and visual attention data** gathered during a subjective experiment used for the evaluation of the artifact detection framework.

Artifact Correction

1. **A color mismatch correction approach for stereoscopic 3D omnidirectional images based on planar Voronoi patches and a robust color transfer method**.

2. **Two deep learning-based approaches for the correction of color mismatch** developed for S3D standard images and extended to S3D ODIs.

Quality Assessment

1. **A general framework for full-reference objective quality assessment of monoscopic omnidirectional video** that extends full-reference quality metrics for monoscopic standard video to monoscopic ODV based on planar Voronoi patches and visual attention.

2. **A dataset with monoscopic omnidirectional videos together with subjective quality scores and visual attention data** gathered during two subjective experiments used for the evaluation of the quality framework.

1.6 Thesis Outline

After this introduction, the background Chapter 2 introduces the foundations of omnidirectional video and its processing pipeline. In this chapter, also two components of some solutions proposed in this thesis are presented, namely, planar Voronoi patches and visual attention. In Chapter 3, the contributions in the area of artifact detection are described. The general framework for the detection of artifacts in S3D ODIs is introduced together with the methods for the detection of sharpness and color mismatch. Also the new S3D ODI dataset used to evaluate the framework is described. Chapter 4 presents the contributions in the area of artifact correction, *i.e.*, methods for the correction of color mismatch. The contributions in the area of quality assessment are presented in Chapter 5. Precisely, the full-reference quality assessment framework for monoscopic ODV is introduced together with a new ODV dataset for the framework evaluation. This thesis finishes with Chapter 6 where the conclusions are presented and the future work is discussed.

1.7 Publications

This is the list of the conference and journal papers where most of the work carried out in this thesis was published:

1. S. Knorr, **Croci, Simone**, and A. Smolic, "A modular scheme for artifact detection in stereoscopic omni-directional images," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, (Maynooth, Ireland), 2017

2. **Simone Croci**, S. Knorr, and A. Smolic, "Saliency-based sharpness mismatch detection for stereoscopic omnidirectional images," in *14th European Conference on Visual Media Production (CVMP)*, (London, UK), 2017

3. **Simone Croci**, S. Knorr, L. Goldmann, and A. Smolic, "A framework for quality control in cinematic VR based on Voronoi patches and saliency," in *IEEE International Conference on 3D Immersion (IC3D)*, (Brussels, Belgium), 2017

4. **Simone Croci**, S. Knorr, and A. Smolic, "Sharpness mismatch detection in stereoscopic content with 360-degree capability," in *IEEE International Conference on Image Processing (ICIP)*, (Athens, Greece), 2018

5. **Simone Croci**, M. Grogan, S. Knorr, and A. Smolic, "Colour correction for stereoscopic omnidirectional images," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, (Belfast, UK), 2018

6. R. Dudek, **Simone Croci**, A. Smolic, and S. Knorr, "Robust global and local color matching in stereoscopic omnidirectional content," *Elsevier Signal Processing: Image Communication*, vol. 74, pp. 231–241, 2019

7. **Simone Croci**, S. Knorr, and A. Smolic, "Study on the perception of sharpness mismatch in stereoscopic video," in *11th International Conference on Quality of Multimedia Experience (QoMEX)*, (Berlin, Germany), 2019

8. **Simone Croci**, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Voronoi-based objective quality metrics for omnidirectional video," in *11th International Conference on Quality of Multimedia Experience (QoMEX)*, (Berlin, Germany), 2019

9. **Simone Croci**, C. Ozcinar, E. Zerman, S. Knorr, J. Cabrera, and A. Smolic, "Visual attention-aware quality estimation framework for omnidirectional video using spherical Voronoi diagram," *Springer Quality and User Experience (QUX)*, vol. 5, 2020

10. **Croci, Simone**, E. Zerman, and A. Smolic, "VIVA-Q: Omnidirectional video quality assessment based on Voronoi patches and visual attention." ISO/IEC JTC1/ SC29/AG2 M56165, 2021

11. **Simone Croci**, C. Ozcinar, E. Zerman, R. Dudek, S. Knorr, and A. Smolic, "Deep color mismatch correction in stereoscopic 3D images," in *IEEE International Conference on ImageProcessing (ICIP)*, (Anchorage, USA), 2021

Chapter 2

Background

This background chapter begins with an introduction to ODV and its processing pipeline. This part is followed by a section about a new planar subdivision of ODV called planar Voronoi patches, which is used in different analysis and processing methods proposed in this thesis. In the end, there is a section about visual attention, also used in some of the proposed solutions.

2.1 Omnidirectional Video

2.1.1 Introduction

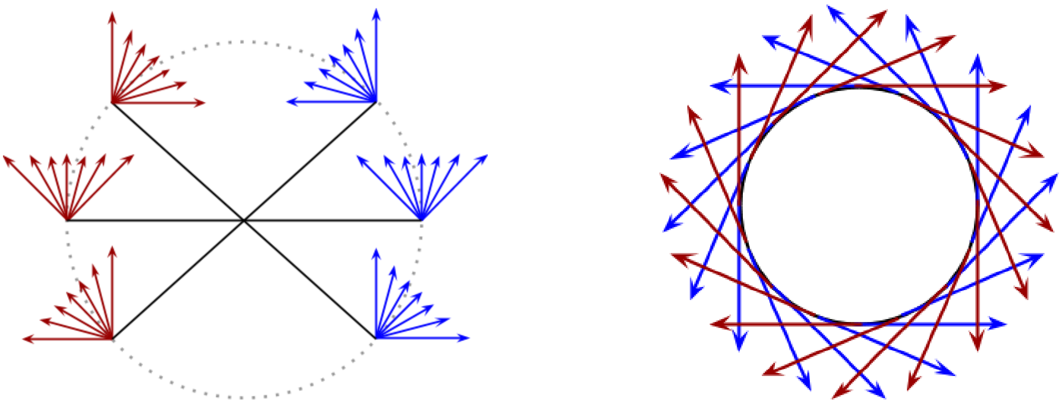
Omnidirectional video can be explained based on the seven-dimensional plenoptic function $\mathcal{P}$ [58]. This function is defined as the intensity of the light rays passing through every location (x, y, z), at every angles (θ, φ), for every wavelength λ, and at every time t:

$$\mathcal{P}(x, y, z, \theta, \varphi, \lambda, t). \tag{2.1}$$

Essentially, the plenoptic function represents all visual information available to an observer at any point in space and time. In the case of ODV, we have a single virtual camera that is moving through the 3D world capturing the light rays in every direction from the viewpoint of the camera. ODV can then be defined as the following function:

$$\mathcal{P}_{\mathrm{ODV}}(\theta, \varphi, \lambda, t) = \mathcal{P}(V_x(t), V_y(t), V_z(t), \theta, \varphi, \lambda, t), \tag{2.2}$$

where $(V_x(t), V_y(t), V_z(t))$ is the position of the camera through time. For simplicity, the wavelength λ can be dropped from the equation since the color information can be represented with a 3D color space. Therefore, ODV can be defined by a three dimensional



(a) Correct version.

(b) Approximation.

Figure 2.1: Stereoscopic 3D omnidirectional video. Red and blue rays correspond to the rays of the left and right eye, respectively. The circles represent the position of the eyes when the head rotates around the fixed midpoint between the eyes. Figures taken from [5].

function $\mathcal{P}_{\mathrm{ODV}}(\theta, \varphi, t)$. From this function, novel views of the world from the constrained camera viewpoint can be rendered in any direction.

So far, we have introduced monoscopic ODV, but there is also a stereoscopic 3D (S3D) version. In theory, in order to have a correct S3D ODV, we should capture an S3D standard video for every possible head orientation as illustrated in Figure 2.1a. This is necessary since the eyes move on the so-called viewing circle when the head rotates around the fixed midpoint between the eyes. In reality, an approximation is used as shown in Figure 2.1b [59, 60, 61, 62, 6]. Instead of capturing all the light rays inside the eye field of view at every position on the viewing circle, the approximation captures only the rays tangent to the viewing circle. Mathematically speaking, let's assume that θ and φ are spherical coordinates as shown in Figure 2.2, and the midpoint between the eyes is $V(t) = (V_x(t), V_y(t), V_z(t))$. Then, with the head oriented in the direction $(cos(\varphi), sin(\varphi), 0)$, the positions of the left and right eye on the viewing circle are defined as follows:

$$V^L(\varphi, t) \quad = \quad V(t) + (-sin(\varphi), cos(\varphi), 0)\ D_{pupil}/2, \tag{2.3}$$

$$V^R(\varphi, t) \quad = \quad V(t) + (sin(\varphi), -cos(\varphi), 0)\ D_{pupil}/2, \tag{2.4}$$

where D_{pupil} is the interpupillary distance. Assuming that at $V^L(\varphi, t)$ and $V^R(\varphi, t)$ the rays tangent on the viewing circle are directed according to the angles (θ, φ), the S3D

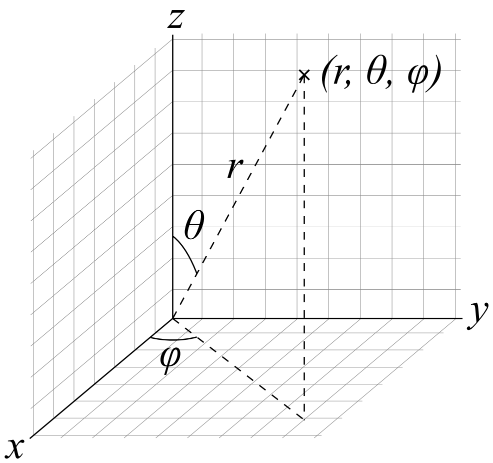


Figure 2.2: Spherical coordinates (r, θ, φ).

ODV consists of two monoscopic ODVs, one for each eye, defined as follows:

$$\mathcal{P}^{L}_{S3D-ODV}(\theta,\varphi,\lambda,t) = \mathcal{P}(V^{L}_{x}(\varphi,t),V^{L}_{y}(\varphi,t),V^{L}_{z}(\varphi,t),\theta,\varphi,\lambda,t), \tag{2.5}$$

$$\mathcal{P}^{R}_{S3D-ODV}(\theta,\varphi,\lambda,t) = \mathcal{P}(V^{R}_{x}(\varphi,t),V^{R}_{y}(\varphi,t),V^{R}_{z}(\varphi,t),\theta,\varphi,\lambda,t). \tag{2.6}$$

S3D ODV creates a correct stereoscopic experience at the center of the observer's view, but increasingly incorrect to the left and right. This is not a big problem because humans have binocular vision at the basis of stereopsis only in the central 114 degrees of the horizontal visual field, and they usually orient the head in the direction they want to look at. From this point, mainly monoscopic ODV is explained, but what presented can be easily adapted to S3D ODV.

Until now, the theoretical description of ODV has been presented. Concretely, ODV is stored as discrete samples of $\mathcal{P}_{\text{ODV}}(\theta,\varphi,t)$. At a particular time sample t_0, we have a 2D function $\mathcal{P}^{t_0}_{\text{ODV}}(\theta,\varphi) = \mathcal{P}_{\text{ODV}}(\theta,\varphi,t_0)$ whose 2D discrete samples are usually stored in a 2D array, *i.e.*, in an image. $\mathcal{P}^{t_0}_{\text{ODV}}(\theta,\varphi)$ can be conceived as a spherical signal, and in order to store it on a planar 2D representation, a mapping from the sphere to the 2D plane, *i.e.*, a spherical projection [63], must be defined. In cartography, this problem applied to the terrestrial globe has been studied for a long time [64, 65]. There are different spherical projections like equirectangular project (ERP), cubemap projection (CMP), octahedral project, icosahedral projection, and custom projections like the planar Voronoi patches presented in Section 2.2. For example, in the case of ERP, the polar angle θ and the azimuthal angle φ are mapped to the vertical and horizontal axis of the 2D image plane, respectively, as illustrated in Figure 2.3. The drawbacks of the spherical projections are the distortion of the spherical signal, as can be seen in Figure 2.3 in the pole regions, and the introduction of discontinuities/borders. These are important aspects that must be

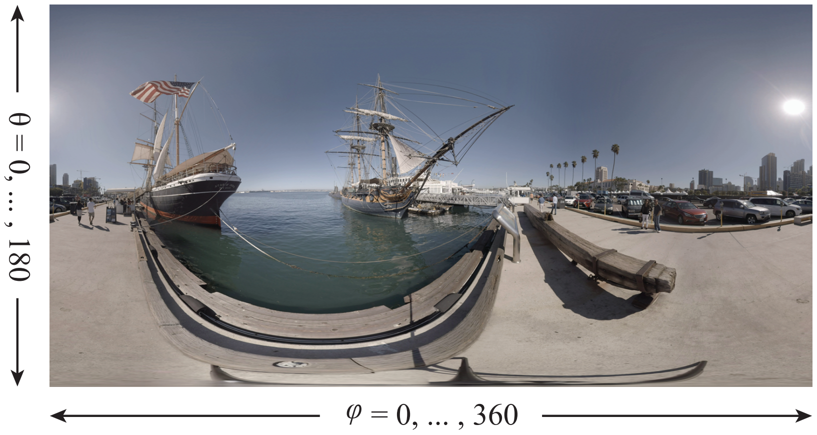


Figure 2.3: Omnidirectional image in equirectangular projection format.

taken into consideration when processing ODVs.

2.1.2 Processing Pipeline

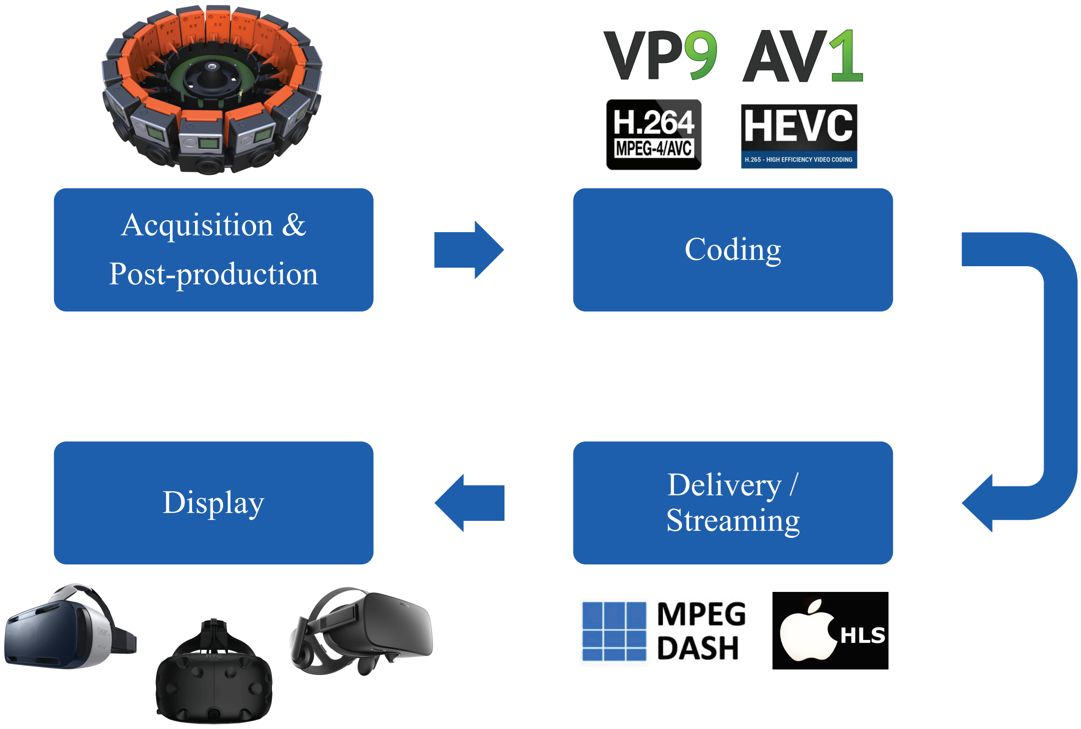


Figure 2.4: Omnidirectional video processing pipeline.

Figure 2.4 shows the typical ODV processing pipeline, from acquisition and post-production to display. ODV acquisition consists of sampling the monoscopic ODV function $\mathcal{P}_{\text{ODV}}$, or the two stereoscopic ODV functions $\mathcal{P}^{L}_{\text{S3D-ODV}}$ and $\mathcal{P}^{R}_{\text{S3D-ODV}}$, using a single

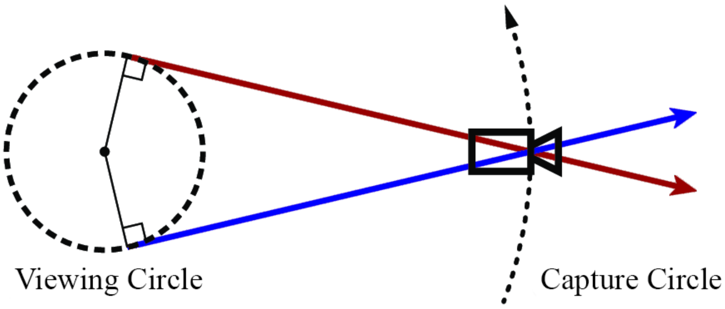


Figure 2.5: Single camera capturing a stereoscopic 3D omnidirectional image. Figure taken from [6].

or multiple cameras. In the case of a static scene, a single camera rotating around its center of projection can be used to capture monoscopic ODIs. A single camera can also capture S3D ODIs when it is rotated outside the viewing circle as shown in Figure 2.5. In the case of dynamic scenes, a single rotating camera cannot be used. In this case, a possible solution consists of a curved mirror that redirects the light rays from surrounding directions to the camera sensor [66]. Possible shapes of the mirror are hyperbola, parabola, or ellipse. This solution works only for monoscopic ODV, and it does not capture rays coming from every direction, even if its field of view is usually superior to 180 degrees. Another solution based on redirecting light rays uses a fisheye lens instead of curved mirrors [67]. The field of view of this solution is usually around 180 degrees, but by combining two cameras with a fisheye lens, like in the camera GoPro Max, the 360-degree field of view of a monoscopic ODV could be covered. There exists also solutions that are based on more than two cameras. In these solutions, the cameras could be arranged in a radial/circular setup [61, 62], like in Google Jump [61], or in a spherical setup, like in Nokia Ozo. A problem of multi-camera solutions capturing monoscopic ODV is the physical impossibility of having different cameras sharing the same center of projection. A consequence of this physical limitation is the presence of artifacts in ODV, like stitching artifacts.

During post-production, the source videos are stitched together in the best possible way, and the artifacts caused by the technical limitations of the capture system are detected and possibly corrected. Specifically, the post-production workflow consists of six steps: 1) data ingest, 2) automatic rough stitching of camera views, 3) manual fine stitching with the removal of stitching and blending artifacts, 4) color grading, 5) editing, and 6) finishing (rendering).

After ODV acquisition and post-production, there is coding and delivery/streaming. Compared to traditional video, ODV introduces new technical challenges especially for storage and transmission [20], such as higher memory and data rate requirements due to the large field of view of ODV [33]. Existing solutions for standard video could also be applied to ODV. For instance, Advanced Video Coding (AVC/H.264) [68] and High Efficiency Video Coding (HEVC/H.265) [69] standards could be used for compression, while Dynamic Adaptive Streaming over HTTP (MPEG-DASH) [70] could be used for transmission. To be compatible with these solutions for standard video, ODV must be stored and transmitted in planar 2D formats based on different spherical projections, which introduce geometric distortions and artificial discontinuities/borders, as previously explained. Ad-hoc solutions for ODV have been also developed. These solutions are viewport-aware, *i.e.*, they take into account the prediction of the viewport trajectory. There are two main variations of viewport-aware solutions: viewport-dependent projection approaches [71], and tile-based approaches [72, 20]. It is worth mentioning that MPEG-I coding standardization [73] for immersive technologies, including ODV, is currently underway.

The last stage in the processing pipeline is display. ODV is ideally viewed with an HMD that shows the portion of the ODV according to the head orientation. Alternatively, it can be viewed with flat-screen devices, like smartphones, tablets, and computers, where the viewpoint is controlled with the finger or the mouse. The region of the ODV that is viewed is called viewport. As illustrated in Figure 2.6, the viewer controls the viewport with three degrees of freedom corresponding to the rotations around three axes, namely, pitch, yaw, and roll axis. The rendering of the viewport consists of projecting it onto the 2D planar format where the ODV is stored and computing the viewport pixels by sampling the ODV. For the viewport projection, the relationship between the viewport and ODV must be defined. For that, the viewport can be conceived as a plane tangent to the spherical representation of ODV, where the points of the viewport plane and the ODV sphere are related by gnomonic projection [74].

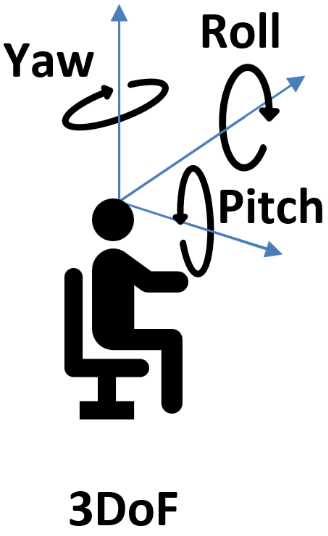


Figure 2.6: Degrees of freedom of omnidirectional video.

2.2 Planar Voronoi Patches

A component of some solutions presented in this thesis is the subdivision of the ODV into planar Voronoi patches. They can be conceived as a new spherical projection technique characterized by low projection distortions that subdivides the ODV in a uniform manner. The planar Voronoi patches were developed in order to efficiently extend methods for traditional video to ODV by applying these methods to each patch. Alternative solutions were considered like equirectangular projection (ERP) and the faces of cubemap projection (CMP). ERP was excluded because it is characterized by large projection distortions and processing it could require too many computational resources (especially memory) due to its large field of view (FoV). The faces of CMP were also excluded even if they are characterized by less projection distortions than ERP, since they have a relatively large FoV. Applying methods for traditional video to regions with a large FoV is not ideal in quality assessment and artifact detection when trying to localize the regions that have a low quality. Planar Voronoi patches are ideal to extend methods for traditional video to ODV since they are characterized by low projection distortions, and they allow to freely choose their number and consequently their size, which is useful in quality assessment and artifact detection for a better localization of the low quality regions. Moreover, they also provide the flexibility to choose their angular resolution, which is convenient in quality assessment to set equal to the angular resolution of the viewer's HMD in order to reproduce the viewing conditions.

For the extraction of M planar Voronoi patches from a given ODV, the spherical Voronoi diagram [75] of M evenly distributed points on the sphere is computed as illus-

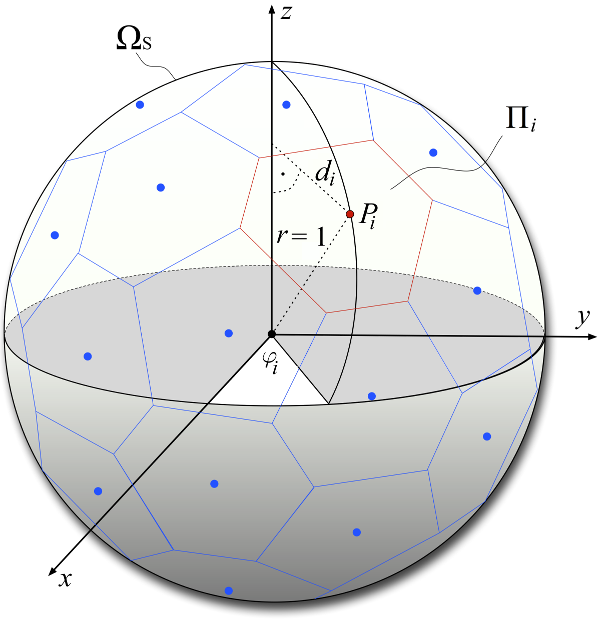


Figure 2.7: Spherical Voronoi diagram of evenly distributed points on the sphere.

trated in Figure 2.7. M evenly distributed points $P_i = (x_i, y_i, z_i)$ on the sphere, where $i \in [1, M]$, can be obtained according to the following equations:

$$\begin{aligned} \varphi_i &= (i-1)\,\pi\left(3-\sqrt{5}\right), && (2.7)\\ z_i &= \left(1-\frac{1}{M}\right)\left(1-\frac{2\,(i-1)}{M-1}\right), && (2.8)\\ d_i &= \sqrt{1-z_i^2}, && (2.9)\\ x_i &= d_i\,cos(\varphi_i) \quad \text{and} && (2.10)\\ y_i &= d_i\,sin(\varphi_i), && (2.11) \end{aligned}$$

where φ_i is the azimuthal angle and d_i is the distance of the point from the z-axis.

The spherical Voronoi diagram defines for each input point P_i the spherical patch Π_i on the surface of the sphere Ω_S that contains all the points that are closer to P_i than to any of the other input points P_l:

$$\Pi_i = \{P \in \Omega_S \mid d_S(P, P_i) \leq d_S(P, P_l)\ \forall l \neq i\}, \tag{2.12}$$

where $d_S(P, P_i)$ is the spherical distance between the point P and the point P_i, *i.e.*, the length of the shortest path on the surface of the sphere connecting these two points. Notice that by using evenly distributed points P_i on the sphere, we guarantee that the spherical Voronoi patches Π_i have approximately equal size.

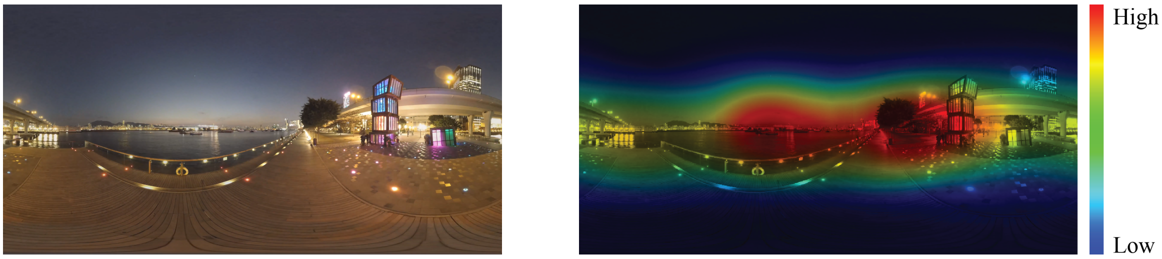
For the computation of the spherical Voronoi diagram, the current solutions in the field of computational geometry [76, 77] provide two options. The first option consists of applying algorithms specifically designed for this problem, like Hyeon-Suk *et al.*'s algorithm, [78]. The second option exploits the duality between the spherial Voronoi diagram and the spherical Delaunay triangulation, and consists of solving the spherical Delaunay triangulation [79, 80] and converting the triangulation into a Voronoi diagram. In our case, we chose the second option [80].

After the computation of the spherical Voronoi diagram, for each spherical Voronoi patch Π_i a planar Voronoi patch Π'_i is extracted from the ODV. This operation is obtained by first positioning the plane of the planar patch Π'_i on the centroid of the spherical patch Π_i tangent to the sphere. The points on the sphere and the planar patch Π'_i are related by the gnomonic projection [74], and the pixels of Π'_i are computed by sampling the ODV in ERP using bilinear interpolation. The angular resolution of each planar Voronoi patch Π'_i is defined by the pixels per visual angle, a parameter that is kept constant for each patch.

2.3 Visual Attention for Omnidirectional Images and Video

Human visual attention is a perceptual mechanism of the human visual system (HVS) that represents nature's answer to the problem of visual information overload [81]. It allows us to selectively process the most relevant portions of the vast amounts of visual information available to the HVS [82]. The concept of saliency map was introduced in [83] and it corresponds to a two dimensional topographic map that denotes the saliency of each pixel, *i.e.*, the higher the scalar value stored in the map, the more salient the pixel is (Figure 2.8). When analyzing and processing ODV, like in ODV quality assessment or coding, it is useful to take visual attention into account in order to give more weight to the regions where end-users are actually looking at, that is, regions that should have high quality.

Visual attention modeling and saliency prediction for standard images and video is an ongoing research topic. Good overviews can be found in [84, 85, 86, 87]. Many computational models that focus on different visual features that motivate visual attention towards a particular target location have been proposed for traditional 2D content in the last 30 years. The visual features can be classified as low level features (color, intensity, orientation) [88, 89], and high level features (face [90] and object [91] detection, image-center prior [89]). The extracted features can be linearly combined [88] or integrated using



(a) Omnidirectional image.

(b) Visual attention/saliency map.

Figure 2.8: Example of visual attention/saliency map.

learned weights [90, 92]. More recently, models using deep neural networks have shown impressive performance in predicting visual attention [93, 94, 95]. A recent overview of deep learning-based methods for standard images and video is given by Borji in [87].

Visual attention for omnidirectional contents, however, is a relatively new research area. A testbed suitable for subjective evaluations of omnidirectional content, including the recording of eye-tracking information, was introduced in [96], while the authors of [97] created a dataset of head movements of users watching ODVs with an HMD. In [98], a simple approach was proposed to obtain visual attention maps by treating raw experimental head direction trajectories in omnidirectional content. The approach excludes parts of a trajectory where the head motion is too fast to fixate the viewer's attention and fuses together the fixation of different viewers. Finally, Gaussian filtering is performed to produce the final visual attention map. According to a study of the eye fixation patterns inside the viewport presented in [99], a more correct alternative to the Gaussian filter for the weighting of the viewport trajectories is represented by a distribution with a donut shape positioned at the viewport center. The authors of [100] estimated visual attention maps for ODIs viewed with HMDs, when the use of an eye tracker device is not possible. They collected viewport data and proposed a new method to transform the gathered data into visual attention maps. They also proposed a method called Fused Saliency Maps that adapts saliency models for traditional images to ODIs. In [45], four different methods for the computation of visual attention maps from eye-tracking data are compared.

In [101], a visual attention model for cylindrical ODIs was proposed for visual robot navigation. The authors of [102] presented a spherical saliency model to compute saliency maps by fusing together static features (intensity, chromatic, and spherical orientation), that are themselves obtained through multiscale analysis on the sphere. In [103], the authors extended their work and presented a computational model of dynamic visual attention on the sphere which combines static and motion features in order to detect salient locations in omnidirectional image sequences while working directly in spherical coordinates.

More recently, a first attempt to attract attention to the problem of creating and predicting saliency maps for ODIs was presented in the Salient360! Grand Challenge at the ICME 2017 conference [104], where head- and eye-tracking ground truth data was given in the form of saliency maps [99]. In [105], the authors fine-tuned traditional image saliency prediction to ODIs by subdividing the ODI into undistorted patches and providing a convolutional neural network (CNN) with the patches together with spherical coordinates for each pixel in the patches. Recently, SalGAN [95], a generative adversarial network for saliency estimation in traditional images, was extended to ODIs obtaining SalGAN360 [106]. Another deep learning-based solution called CubePadding360 was proposed by Cheng *et al.* [107]. Differently from the previous solutions, CubePadding360 was developed for ODV instead of ODIs, and it was trained in a weakly supervised manner. Moreover, Xu *et al.* [108] created a large-scale eye-tracking ODV dataset and developed a model that predicts where the viewer will look at an upcoming time based on ODV content and the past gaze positions.

Chapter 3

Voronoi-based Artifact Detection in Stereoscopic 3D Omnidirectional Images

3.1 Introduction

The production of omnidirectional video is not a simple task due to technical challenges that can introduce artifacts and degrade the quality. Therefore, the detection of artifacts is an important aspect of quality control. This chapter deals with artifact detection, and our ultimate goal is to provide algorithms and tools for automatic detection and visualization/highlight of artifacts in order to give automatic feedback to artists and reduce time and efforts in the post-production process. To our knowledge, there is no scientific publication in this area except for ours, and thus this is an open research field that we believe of high importance.

In this chapter, a general framework for the detection of artifacts in S3D ODIs is proposed, together with two methods for the detection of sharpness mismatch (SM) and one method for the detection of color mismatch (CM). A dataset of S3D ODIs that was created for the evaluation of the artifact detection methods is also presented in this chapter. Most of these contributions were published in [48, 49, 50, 53].

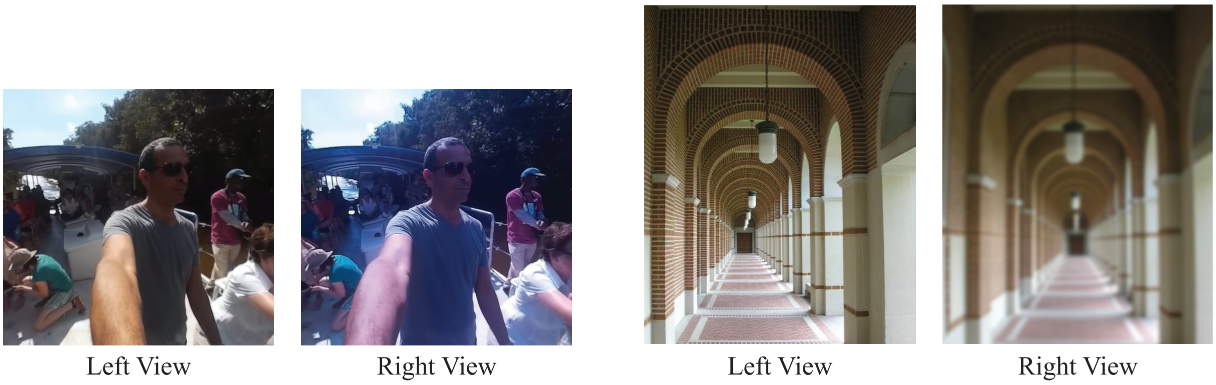
3.2 Background

The common artifacts introduced during the production of S3D ODV can be organized into three categories: binocular rivalry issues, conflicts of depth cues, and artifacts that occur in both monoscopic and stereoscopic 360-degree content production [25, 47]. Binoc-

Table 3.1: Binocular rivalry issues in stereoscopic 3D omnidirectional video.

Artifact/ Issue	**Characteristics**	**Caused by**
Geometrical misalignment	Improper (vertical) alignment of left and right images	Cameras or lenses not properly aligned Tilting head or changing yaw while looking at the pole caps with an HMD
Luminance/ color mismatch	Difference in hue, saturation and/or intensity between left and right image	Cameras not properly matched (*e.g.* different aperture) Varying lighting conditions at different camera locations
Visual mismatch	Reflections, lens flares, polarization Contamination Missing or different objects in one of the views	Varying lighting conditions at different camera locations Contamination due to environmental conditions (*e.g.* rain, dust, etc.) Compositing errors in post
Depth of field/ sharpness mismatch	Difference in sharpness or depth of field	Different aperture settings of cameras Focal length of cameras not properly matched
Synchronization	Left and right image sequences are not synchronized	Cameras are not synchronized/ gen-locked Editing errors in post
Hyperconvergence/ hyperdivergence	Objects are too close to or too far from the viewer's eyes	Too much negative or positive parallax between left and right image
Pseudo-3D	Left and right images are swapped	Swapped images in HMDs Editing error in post
Ghosting	Double edges of objects	Stitching and blending artifacts in post

ular rivalry issues are present when there is a misalignment between the left and right stereo images (Table 3.1). An example of a common binocular rivalry issue is color mismatch (CM), which occurs when the color of an object in the left image is different from the color of the same object in the right image (Figure 3.1a). Another example is sharpness mismatch (SM), which occurs when the sharpness of a region in the left image is different from the sharpness of the same region in the right image (Figure 3.1b). On the other hand, depth conflicts are present when different depth cues used by the human visual system are conflicting (Table 3.2). The depth cues are categorized in monocular when they require the input from just one eye, such as perspective, motion parallax,



(a) Color mismatch.

(b) Sharpness mismatch.

Figure 3.1: Examples of artifacts present in stereoscopic 3D omnidirectional video.

Table 3.2: Depth conflicts in stereoscopic 3D omnidirectional video.

Depth conflict	**Characteristics**	**Caused by**
Vergence vs. accommodation	Eyes accommodate on screen plane but converge or diverge on objects in front or behind the screen plane	Parallax between objects in the left and right view
Stereopsis vs. interposition	Foreground objects are occluded by background objects	3D compositing errors in post
Accommodation vs. depth of field	Eyes accommodate on screen plane but scene or part of scene is out of focus	Wide aperture of cameras
Stereopsis vs. (aerial) perspective	Monocular depth cue ”perspective” or ”aerial perspective” does not match with binocular depth cue ”stereopsis”	3D compositing errors in post
Stereopsis vs. motion parallax	Motion of objects does not match with their distance	3D compositing errors in post
Stereopsis vs. size	Relative or familiar size of objects does not match with their distance	3D compositing errors in post
Stereopsis vs. light and shading	Distance or shape of objects does not match with their shadings	3D compositing errors in post
Stereopsis vs. texture gradient	Texture gradients are not in line with the descending of depth in the scene	3D compositing errors in post

Table 3.3: Artifacts in both monoscopic and stereoscopic 3D omnidirectional video.

Artifact/ Issue	**Characteristics**	**Caused by**
Stitching artifacts	Visible seams and misaligned/ broken edges	Improper camera arrangement Registration and alignment errors in post
Blending artifacts	Visible color and luminance mismatches of regions within an ODI	Varying lighting conditions at different camera locations Compositing errors in post
Warping artifacts	Visible deformations of objects	Improper camera arrangement Registration and alignment errors in post
Wobbling artifacts	Unsteady scene appearance over time	Temporal inconsistent stitching of camera views (non-stabilized image sequences)

interposition, etc., and binocular when they require the input from both eyes, such as convergence, stereopsis or retinal disparity. An example of depth conflicts is stereopsis vs. interposition, which occurs when foreground objects are occluded by background objects due to 3D compositing errors in post-production. The third category of issues includes the artifacts that are present only in ODV, like stitching or blending artifacts (Table 3.3). These artifacts occur only in multi-camera systems used for panorama capture.

Over the last years, binocular rivalry issues and conflicts of depth cues have been investigated in detail for standard S3D content, *e.g.* for cinema screens [25, 28] and 3D-TV [27, 109], and more recently for omnidirectional S3D content for HMDs [110]. Many publications focused on the assessment of 3D quality in terms of subjective and objective quality evaluation. In [111], the authors investigated with subjective tests how viewer annoyance depends on various technical parameters such as vertical disparity, rotation and field-of-view mismatches, as well as color and luminance mismatches between the views. In [112], a stereo camera distortion detecting method based on statistical models was presented in order to detect vertical misalignment, camera rotation, unsynchronized zooming, and color mismatch in native S3D content. The authors of [113] introduced a full-reference quality assessment metric for stereoscopic images based on the perceptual binocular characteristics. The proposed metric handles asymmetric distortions of stereoscopic images by incorporating human visual system characteristics. Moreover, in [114] another full-reference metric was presented that evaluates a large variety of measures and that takes 2D picture quality, binocular rivalry, and disparity map degradation into account. The authors maximized the correlation with the mean opinion score (MOS) by using linear regression.

In this chapter, however, the goal is to support the artist by giving direct feedback regarding S3D quality during post-production. Thus, full-reference quality metrics can not be applied in this context. In [115], the authors explored the relationship between the perceptual quality of stereoscopic images and visual information, and they introduced a model for binocular quality perception. Based on this model, a no-reference quality metric for stereoscopic images was proposed. The proposed metric is a top-down method modeling the binocular quality perception of the human visual system (HVS) in the context of blurriness and blockiness.

A large variety of artifact detection methods, including sharpness mismatch (SM) and color mismatch (CM) detection methods, were introduced in [30, 116, 117]. For SM, the three papers proposed approaches that first apply dense disparity estimation, and then analyze high-frequency differences between both views [30], or estimate and compare a simple blur model between corresponding patches [116], or analyze differences of edges using a gradient-based approach [117]. Liu *et al.* [31] presented an automatic no-reference approach for measuring the probability of sharpness mismatch (PSM). This probability is estimated by measuring width deviations of edge pairs in different depth planes in both views. They demonstrated that the proposed metric outperforms the state-of-the-art S3D quality metrics that analyze SM between stereoscopic views.

For measuring in-picture sharpness, different metrics have been developed. In [118], a new perceptual no-reference image sharpness metric based on the notion of just noticeable blur (JNB) was introduced. The proposed metric is able to predict the relative amount of blurriness in images with different content. An ideal metric is the cumulative probability of blur detection (CPBD) metric [119], as it outperforms most other no-reference sharpness metrics on Gaussian blur. It was developed based on human blur perception at different contrasts.

For CM, the authors of [30] use the results of the disparity estimation to reconstruct one view from the other and compare the colors from the original and the reconstructed view based on the mean square error (MSE) in the RGB color space. To eliminate the influence of matching errors and occlusions, the pixels with the highest mismatch are ignored. The authors of [120] proposed several objective metrics for luminance mismatch and evaluated their correlation with the results of subjective experiments.

None of the related work presented in this section focused on S3D artifact detection in ODIs or ODV. To our knowledge only our work published in [47, 48, 49, 50, 52] focuses on S3D quality assessment methods that deal with ODIs and can be easily extended to ODV.

3.3 Voronoi-based Framework

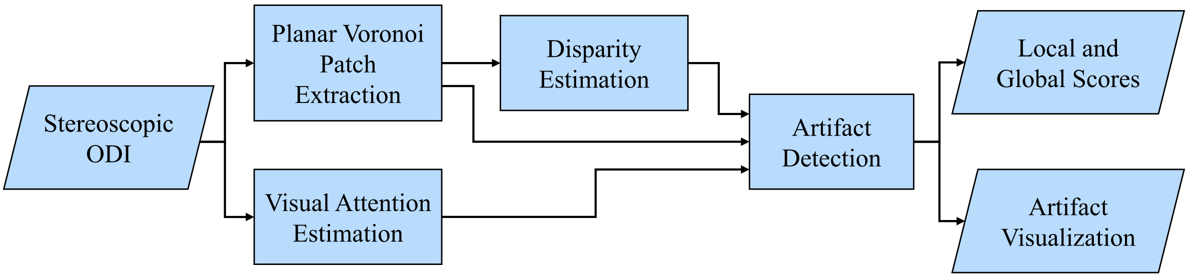


Figure 3.2: Voronoi-based framework for artifact detection in stereoscopic 3D omnidirectional images.

For the detection of artifacts in S3D ODIs, we developed a general framework published in [48, 49, 50] and illustrated in Figure 3.2. In the problem of artifact detection in S3D ODIs with a large field of view, it is important not only to detect artifacts but also to localize them. For this reason, we propose a framework that uniformly subdivides the S3D ODIs into patches, and then it analyzes each patch for the presence of artifacts. First, the framework extracts approximately equally sized planar Voronoi patches from the left and right view of the ODI as described in Section 2.2. In parallel, visual attention is estimated. Then, for each pair of corresponding S3D patches, a disparity map is computed. Afterward, the pairs of corresponding S3D patches are processed independently by an artifact detection method for standard S3D content, taking visual attention into account in order to give more weight to artifacts that are present in regions with high visual attention. From the processing of each patch, a local patch score is computed, which is large if artifacts are detected, and small otherwise. Next, the patch scores are combined together in order to obtain global scores, and finally, the patch scores are visualized in the ODI using a color-coded representation. All the components of the framework are presented in the next sections.

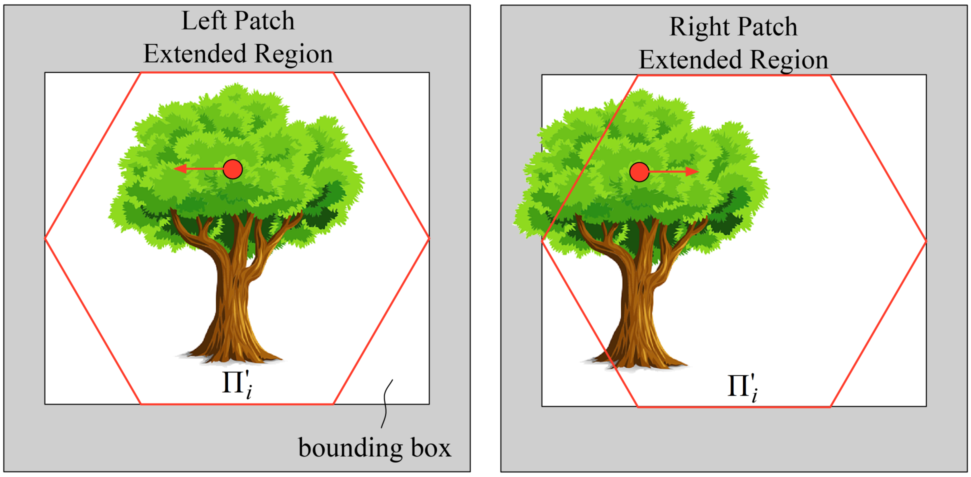
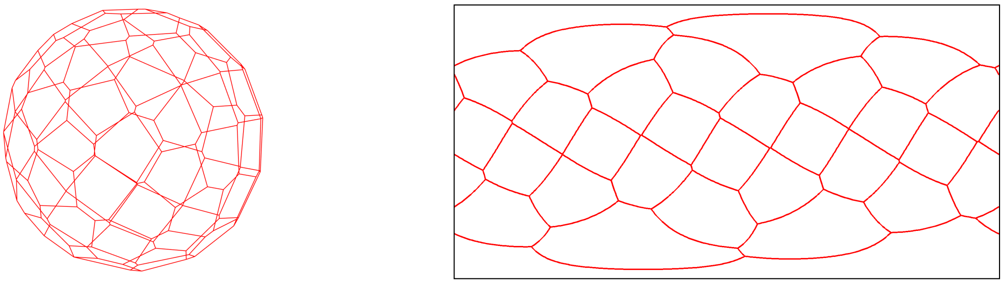


Figure 3.3: Disparity compensation for corresponding planar Voronoi patches Π'_i in the left and right images, respectively.



(a) Spherical Voronoi diagram with 30 patches.

(b) Spherical Voronoi diagram mapped into the equirectangular projection format.

Figure 3.4: Voronoi patches.

3.3.1 Planar Voronoi Patches

The planar Voronoi patches Π'_i are extracted from the ODI according to the method described in Section 2.2, which is based on the computation of a spherical Voronoi diagram of evenly distributed points on the sphere. In the presence of disparity, it can occur that a region inside a planar Voronoi patch in one view is outside the corresponding planar Voronoi patch in the other view. In order to cope with the disparity, we add a border around the Voronoi patch when it is extracted, as shown in Figure 3.3.

The number of patches and thus the size of each patch have an impact on the localization of the artifacts. The larger the patch size is, the more difficult it is to detect and localize the artifacts if they only appear in small areas of the ODI. We empirically found out that 30 patches are a good number for the localization of the artifacts. Figure 3.4 shows the spherical Voronoi diagram computed from 30 evenly distributed points on the sphere and its projection into the equirectangular projection format.

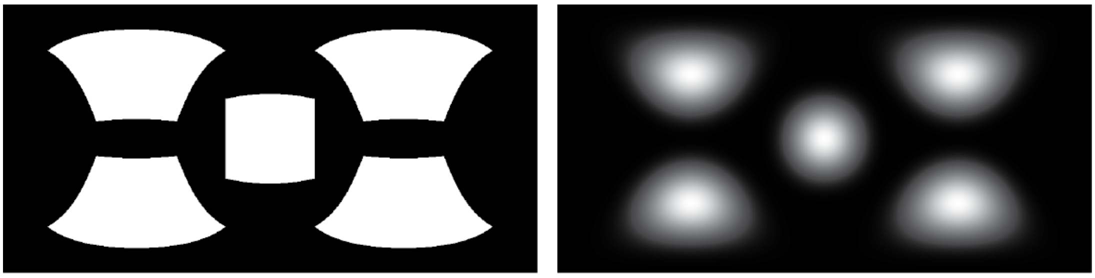


Figure 3.5: Projections of the viewport (left) and the Gaussian filter defined in the viewport (right) into the equirectangular projection format.

3.3.2 Visual Attention Estimation

The visual attention map is useful in order to identify regions that should have a high visual quality, *i.e.*, regions that should be free from artifacts (*e.g.* SM) where end-users are actually looking at. For visual attention estimation, we developed a new method that was published in [48]. This method computes visual attention maps in ERP format, and it was inspired by De Abreu *et al.*'s method [100]. In our approach, the visual attention map is computed from a sequence of HMD viewport positions recorded while a viewer is freely looking at an ODI. For each viewport position, a filter kernel centered on the viewport and defined by its dimension is projected onto the ERP image (see Figure 3.5). The projections of the filter kernels are then added in order to obtain the final visual attention map (see Figure 3.6).

In our approach, we use the Gaussian filter centered on the viewport according to the assumption proposed in [100] that the viewer tends to look at the center of the viewport rather than at the borders. This assumption is supported by two facts, namely, that the visual acuity is at its maximum at the center of the human visual field (fovea), and the head tends to follow the eye movements to preserve the eye resting position (eyes looking straight ahead). According to the paper [99] published after our research, this assumption was corrected based on a new study. According to this study, the eye fixation patterns inside the viewport can be better modelled with a donut-shaped distribution centered on the viewport rather than with a Gaussian distribution.

The Gaussian filter is defined as follows:

$$h\left(u,v\right) = e^{-\frac{1}{2}\left(\frac{u^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right)}, \tag{3.1}$$

where (u, v) are pixel coordinates centered on the HMD viewport, while σ_u and σ_v control the horizontal and vertical filter size and are related to the field of view and the resolution of the HMD viewport. Figure 3.5 shows five projections of the viewport and the Gaussian filter defined on the viewport into the ODI in ERP format. In contrast to our method,

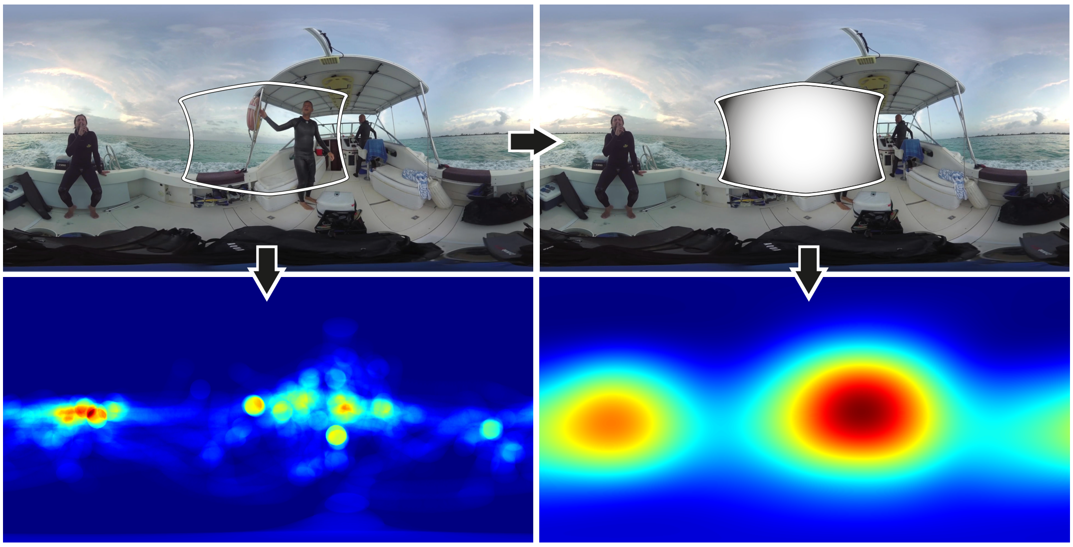


Figure 3.6: Computation of visual attention maps: original ODI in equirectangular projection format with overlaid viewport and center gaze (top left), with overlaid Gaussian filter defined in the viewport (top right), and the resulting visual attention maps with (bottom right) and without (bottom left) Gaussian filtering. Blue: low visual attention, red: high visual attention.

the one by De Abreu *et al.* [100] needs to model the pixel deformation in the ERP format since it applies a Gaussian filter in this format after the projection of the viewport centers. In our approach, we apply the filter on the viewport, and then we project the filter on the equirectangular format without the need to model the pixel deformation.

3.3.3 Disparity Estimation

As per pixel disparity information is usually required for the detection of the artifacts in S3D content, dense disparity estimation is the third step of our approach. Disparity estimation, also called stereo matching, is a well-studied problem in photogrammetry [121, 122], and different solutions have been developed [123, 124, 125, 126, 127]. To estimate disparity maps between the corresponding left and right planar Voronoi patches, we apply the semi-global block matching approach described in [127] which delivers good results at reasonable computational costs.

Since the disparity estimation can be noisy and inaccurate, we apply a consistency check for the disparity values, and only disparity values that are consistent are used for further computations. Assuming that $D_{L\to R}$ and $D_{R\to L}$ are the disparity maps from left to right view and from right to left view, respectively, then the disparity at pixel (x, y) in $D_{L\to R}$ is valid if

$$|D_{L\to R}(x,y) + D_{R\to L}(x - D_{L\to R}(x,y), y)| \leq \delta, \tag{3.2}$$

where δ is a predefined threshold.

3.3.4 Artifact Detection and Analysis

Inspired by the work in [128], where visual attention is used for weighting the spherical PSNR in the context of coding, we incorporate visual attention in the detection of artifacts to weight the artifacts depending on the likelihood of appearance in the HMD viewport of the end-users. In this way, more weight is given to the artifacts present in regions where the viewer is looking with high probability.

Visual attention can be integrated at two levels: pixel and patch level. At the pixel level, the pixel visual attention $\psi(p)$ can be used in order to weight each pixel p that is processed using a weight equal to $g'(\psi(p))$, where g' is a function that can be freely chosen. The visual attention at the pixel level can be used by the artifact detection methods for standard S3D images in order to compute the local patch scores. At the patch level, the patch visual attention Ψ_i, which is equal to the average of the pixel visual attention inside the patch i, is used to weight the local patch scores using a weight equal to $g''(\Psi_i)$, where g'' is also a function that can be freely chosen. The visual attention at the patch level is used to compute the global scores of the ODI.

We propose two global scores, the visual attention-based weighted sum of local patch scores and the number of patches with artifacts. On the one hand, the visual attention-based weighted sum of local patch scores is defined by the following equation:

$$S_{global} = \frac{\sum_{i=1}^{M} g''(\Psi_i)\ S_i}{\sum_{i=1}^{M} g''(\Psi_i)}, \tag{3.3}$$

where S_i is the local patch score, and M is the number of patches. On the other hand, the number of patches with artifacts can be computed as follows:

$$\sum_{i=1}^{M} \mathbb{1}_{g''(\Psi_i)\ S_i \geq \gamma}, \tag{3.4}$$

where γ is a user-defined threshold and $\mathbb{1}$ is an indicator function, which is equal to one if the condition $g''(\Psi_i)\ S_i \geq \gamma$ is true, and zero otherwise.

As mentioned before, the visual attention weight functions g' and g'' can be freely chosen and controlled by the artist. In this way, the artist can decide whether to ignore the visual attention by choosing a constant g', or to completely ignore the pixels with visual attention lower than a threshold, by setting the weight of these pixels to zero. An example of function g' and g'' used in some of the publications is the piece-wise linear function defined as

$$g'(x) = g''(x) = \begin{cases} 5 \cdot x, & x \leq 0.2, \\ 1, & x > 0.2, \end{cases} \tag{3.5}$$

i.e., pixels and patches with visual attention larger than 20% have maximal weight when calculating the scores.

3.3.5 Artifact Visualization

For the artist assessing the quality of the ODI, it is useful to visualize the patch scores directly on the ERP image. Different colormaps can be applied in order to assign a color to the patch score. In the results presented in this thesis, we used the jet colormap, which assigns blue to 0, red to the maximum possible score, and green to a middle score. Additionally, the patch scores and the patch visual attention can be displayed directly within each patch as text to further substantiate the analysis. Patch visual attention can be displayed for practical reasons within post-production workflows, as artists should first get visual feedback if the artifact exists, and then decide, dependent on the visual attention in a patch, if a correction is required. Figure 3.7 illustrates an example of visualization of SM patch scores in an ODI with SM localized in the center using the jet colormap.

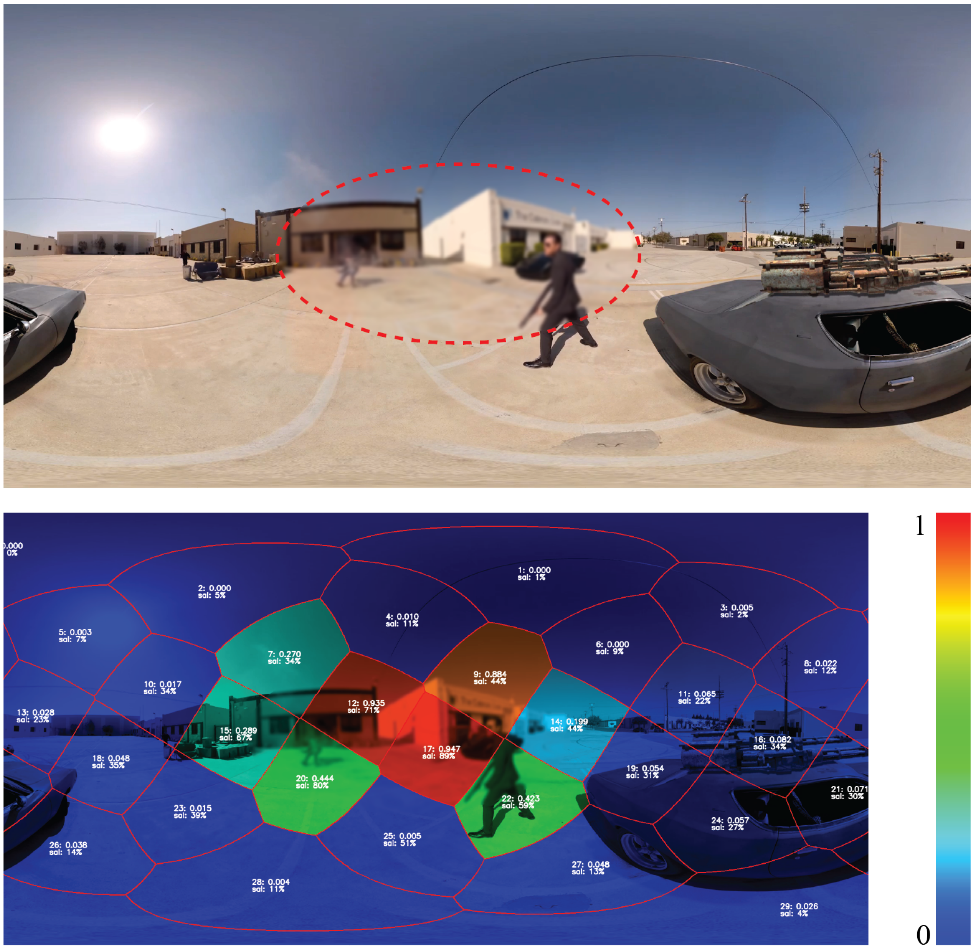


Figure 3.7: Example of sharpness mismatch score visualization: left ODI with blur inside the red ellipse (top) and sharpness mismatch visualization (bottom) including a text overlay with the sharpness mismatch patch score and the visual attention.

3.4 Omnidirectional Image Artifact Dataset

This section introduces the dataset of S3D ODIs together with visual attention maps that was created for the evaluation of the artifact detection methods. The dataset was made public in [50], and to our knowledge, it is the first and only dataset with S3D ODIs currently publicly available.

3.4.1 Image Selection

The dataset consists of 96 S3D ODIs collected from different public sources. The resolution of the ODIs ranges from 1920×960 to 4640×2320 pixels per view. In order to have a large variety, the dataset has the following characteristics: 32 indoor scenes, 51 landscape scenes, 48 scenes containing humans, 47 ODIs with both pole caps covered, 19 ODIs with only the top pole cap covered, 30 ODIs without pole caps, 90 ODIs captured in native

3D while 6 post-converted to 3D.

The dataset was captured with a range of different 360°-rigs. These are Google Odyssey (7), Jaunt rig prototype I and II (7), Panocam POD 3D (9), VUZE VR (5), Nokia OZO (4), customized rig by INVR (3), customized rig by Jumpgate (4), Omnicam 3D (1), customized rig with Mobius cameras (1), unknown 3D rigs (49) and post-converted (6).

3.4.2 Subjective Experiment

To compute the visual attention maps of the dataset using the method presented in Section 3.3.2, we organized a subjective experiment similar to the one described in [100]. During the experiment, the participants were asked to freely look at S3D ODIs while wearing an HMD and sitting on a turn-chair. The HMD used in the experiment is an Oculus Rift DK2 with a resolution of 960×1080 pixels per view and a vertical and horizontal field of view equal to 100°. While the subjects were looking at the images, we recorded the viewport center locations on each of the ODIs, assuming that the center of the viewport corresponds to the visual target location of the user. As previously explained, this assumption was proposed in [100], but later corrected in [99] where a donut-shaped distribution of the eye fixations was proposed.

The experiment was divided into a training and test session. During the training session, the subjects got familiar with the experiment, while a demo image was displayed. During the test session, the dataset of 96 ODIs was displayed in random order using the software application introduced in [100] which was modified to display S3D ODIs. For cross-platform compatibility reasons, the application was implemented using the WebVR and the ThreeJS APIs, so that it can run with different HMDs on different web browsers. The application is able to collect viewport information from the gyroscopic sensors at the refresh rate of the HMD. For Oculus Rift DK2 the maximum refresh rate is 75 Hz, meaning 13.33 ms per frame.

Each image was displayed for 15 seconds, and according to [100], the data captured during the first second was discarded as it adds trivial information on the starting viewing direction. A break of one minute was introduced in the middle of the experiment. A total of 17 subjects (3 females and 14 males) between 20 and 56 years and with normal stereo vision took part in the experiment. In order to keep anonymity, we assigned an identifier to each of them.

Figure 3.8 shows some visual attention maps computed with our method. These maps show that the test subjects tended to look at the equator of the ODIs rather than at the pole caps. Moreover, high-level features like the bear in ODI 54 and the gunshot in ODI

23 attracted the visual attention of the subjects.

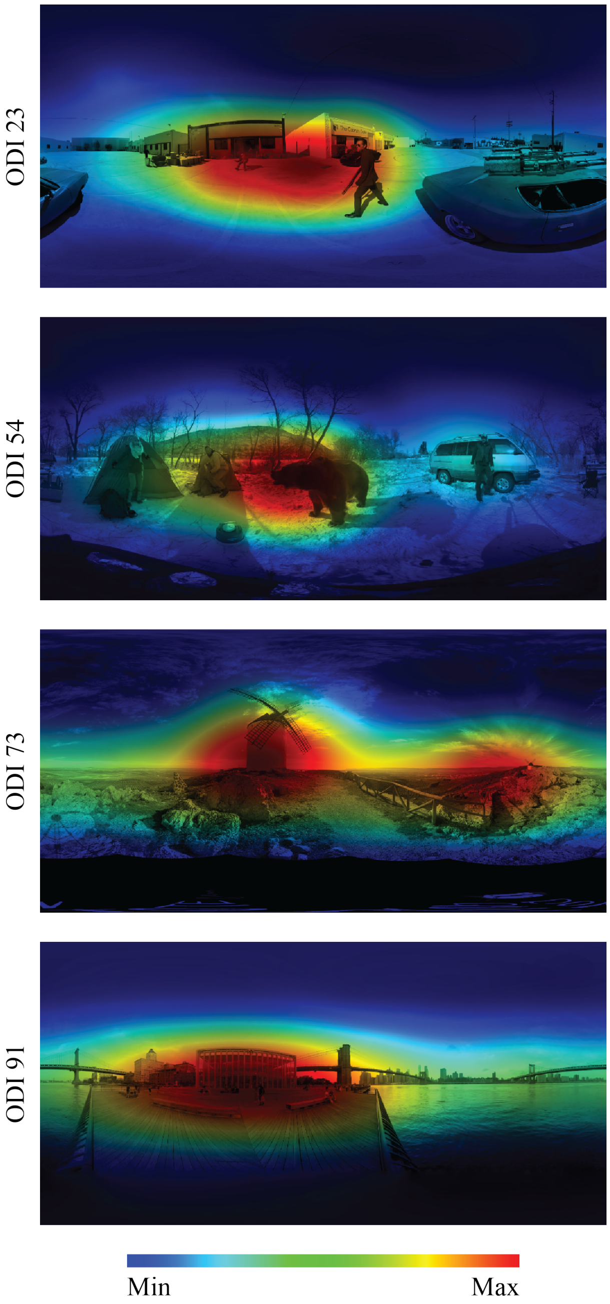


Figure 3.8: Visual attention maps.

3.5 Sharpness Mismatch Detection

Sharpness mismatch (SM) can be introduced in S3D video during shooting with two cameras having different focal lengths or aperture settings, or by asymmetric compression. In this section, two different SM detection methods for S3D standard images are presented. These methods are then extended to S3D ODIs based on our general framework (Section 3.3).

3.5.1 Histogram-based Method

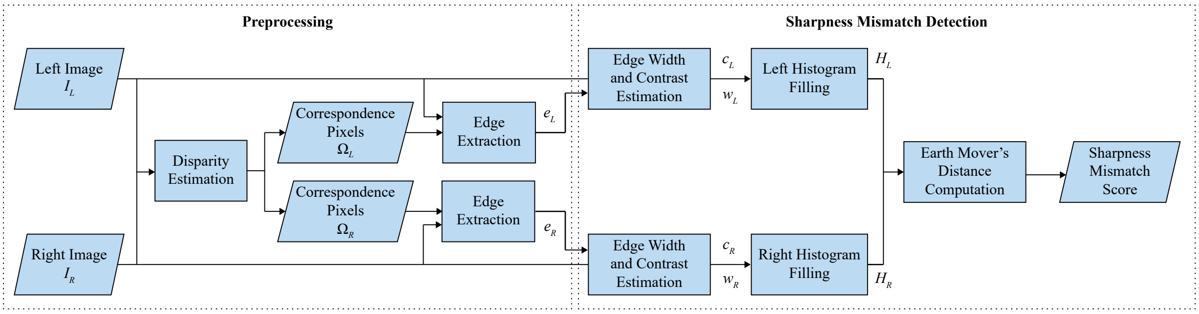


Figure 3.9: Overview of the processing steps of the histogram-based sharpness mismatch detection method.

Our histogram-based sharpness mismatch detection method (HSMD), which was published in [50], is based on the observation that when blur is applied to an image, the widths of most of the edges increase. Essentially, the proposed method computes the edge width distributions in the form of histograms in the left and right view, and then it estimates their difference. The method consists of a preprocessing and an actual SM detection step illustrated in Figure 3.9. In the preprocessing step, the disparity maps $D_{L\to R}$ from the left to the right view and $D_{R\to L}$ in the other direction are estimated and a consistency check is applied as described in Section 3.3.3. The validated disparity map $D_{L\to R}$ defines the valid correspondences between the left image I_L and the right image I_R. Then, the sets of pixels $\Omega_L \subseteq I_L$ and $\Omega_R \subseteq I_R$ with valid correspondences are extracted, *i.e.*, each pixel $(x, y) \in \Omega_L$ has a valid correspondence $(x', y) \in \Omega_R$ with $x' = x - D_{L\to R}(x, y)$.

Then, edge pixels $e_L \in \Omega_L$ and $e_R \in \Omega_R$ are extracted in both images using the Canny edge detector [129]. For each edge pixel, the edge width and contrast are estimated using the method described in [130]. Based on the edge pixels, two 2D histograms $H_L(c_i, w_j)$ and $H_R(c_i, w_j)$ with edge contrast bins c_i and edge width bins w_j are filled for the left and right view, respectively. Finally, the SM score is obtained by computing the distance between the two histograms. First, 1D edge width histograms $H_L^i(w_j) = H_L(c_i, w_j)$ and

$H^i_R(w_j) = H_R(c_i, w_j)$ are extracted from the original 2D histograms for each edge contrast bin c_i. In order to obtain an histogram distance independent of the amount of edge pixels, i.e. the total area of the histograms, we normalize the 1D edge width histograms with:

$$\hat{H}^i_{[L,R]} = \frac{H^i_{[L,R]}}{A^i}, \tag{3.6}$$

where $A^i = max(A^i_L, A^i_R)$, and A^i_L and A^i_R are the areas of the left and right histograms H^i_L and H^i_R: $A^i_{[L,R]} = \sum_j H^i_{[L,R]}(w_j)$.

A well-established metric to measure differences between two histograms H_0 and H_1 is the earth mover's distance $EMD(H_0, H_1)$ [131]. More precisely, EMD computes the flow f_{ij} which represents the amount that is transferred from bin i in H_0 to bin j in H_1. Formally, $EMD(H_0, H_1)$ is defined as follows:

$$\begin{aligned} EMD(H_0, H_1) \;&=\; \min_{\{f_{ij}\}}(\sum_i \sum_j f_{ij} d_{ij}) \;+ \\ &\quad\; \alpha|\sum_i H_0(i) - \sum_j H_1(j)|, \end{aligned} \tag{3.7}$$

subject to the following constraints:

$$f_{ij} \geq 0, \sum_j f_{ij} \leq H_0(i), \sum_i f_{ij} \leq H_1(j), \tag{3.8}$$

$$\sum_i \sum_j f_{ij} = min(\sum_i H_0(i), \sum_j H_1(j)), \tag{3.9}$$

where α is a user-defined parameter, and d_{ij} is the distance between the bins i and j. We define $d_{ij} = |i - j|/N$ with N equal to the number of bins.

The final SM score S is then obtained by summing the distances between 1D histograms weighted by the number of edge pixels as follows:

$$S = \sum_i EMD(\hat{H}^i_L, \hat{H}^i_R)\ \nu_i, \tag{3.10}$$

with

$$\nu_i = \frac{A^i_L + A^i_R}{\sum_k \left(A^k_L + A^k_R\right)}. \tag{3.11}$$

3.5.2 Just Noticeable Sharpness Mismatch-Based Method

This section presents another method for the detection of SM published in [53] and based on the just noticeable sharpness mismatch (JNSM), *i.e.*, the minimal level of SM that is perceived by the human visual system and creates discomfort. When an S3D image has a low level of SM, it can happen that the viewer does not perceive it, that is, she/he perceives the image as sharp as the sharpest of the two stereoscopic views. The mechanism of the human visual system behind this behavior is called interocular blur suppression [132]. For the development of our new method for the detection of SM, we are interested in the limits of interocular blur suppression, *i.e.*, the threshold where the human brain is no more able to suppress SM. We call this threshold JNSM. In other words, given an S3D image without SM, the JNSM is the minimal amount of blur applied to one of the two views, so that the viewer perceives a difference with respect to the original S3D image.

In the scientific literature, there are different publications about binocular suppression [133, 134, 135, 136, 132], which is a mechanism related to interocular blur suppression. Binocular suppression occurs when the stereoscopic views are of different quality (i.e. not only SM), and the higher quality view dominates the perceived quality. Binocular suppression was studied by Julesz [133] based on experiments with random dot stereograms. It was also investigated in studies related to monovision correction [134, 135], and asymmetric compression [136, 132]. The JNSM was studied in [132] with a psychophysical experiment using wave gratings with vertical and horizontal orientations, and with different contrasts and spatial frequencies. It was observed that orientation, contrast, and spatial frequency do not have a large influence on the JNSM.

In this section, we first present a psychophysical experiment that extends the study in [132] by exploring how the JNSM is influenced by other two dimensions of the wave grating stimulus: the symmetric blur, *i.e.*, blur equally applied to the two views, and the disparity. Then, we present the SM detection method based on the JNSM.

Psychophysical Experiment

The goal of the experiment was to measure the JNSM for the development of a new method for SM detection. In particular, we selected stimuli necessary to develop the core of our method, that is, a criterion for the evaluation of the perceived SM at the edges in an S3D image. In order to measure the JNSM in the experiment, two stimuli identical except for SM were shown at the same time. One stimulus was the reference stimulus without SM, and the other one was the test stimulus with SM. The task of the subjects was to see whether these two stimuli were perceived as different.

Stimuli: The stimuli used in the experiment were squared S3D wave gratings with a side length equal to 6 degrees of visual angle. The wave gratings of the two views were obtained by applying the Gaussian filter to a sequence of 12 equal-sized vertical stripes of two alternating gray intensities. Here, we assume that when the standard deviation σ

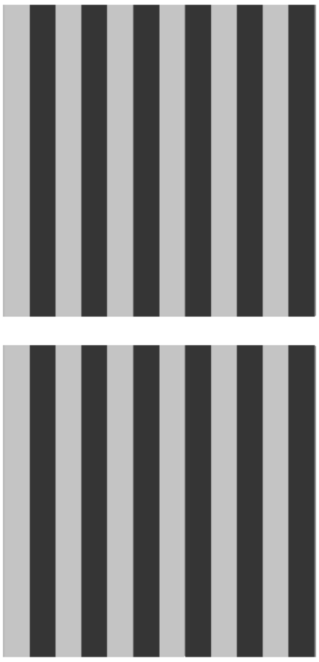
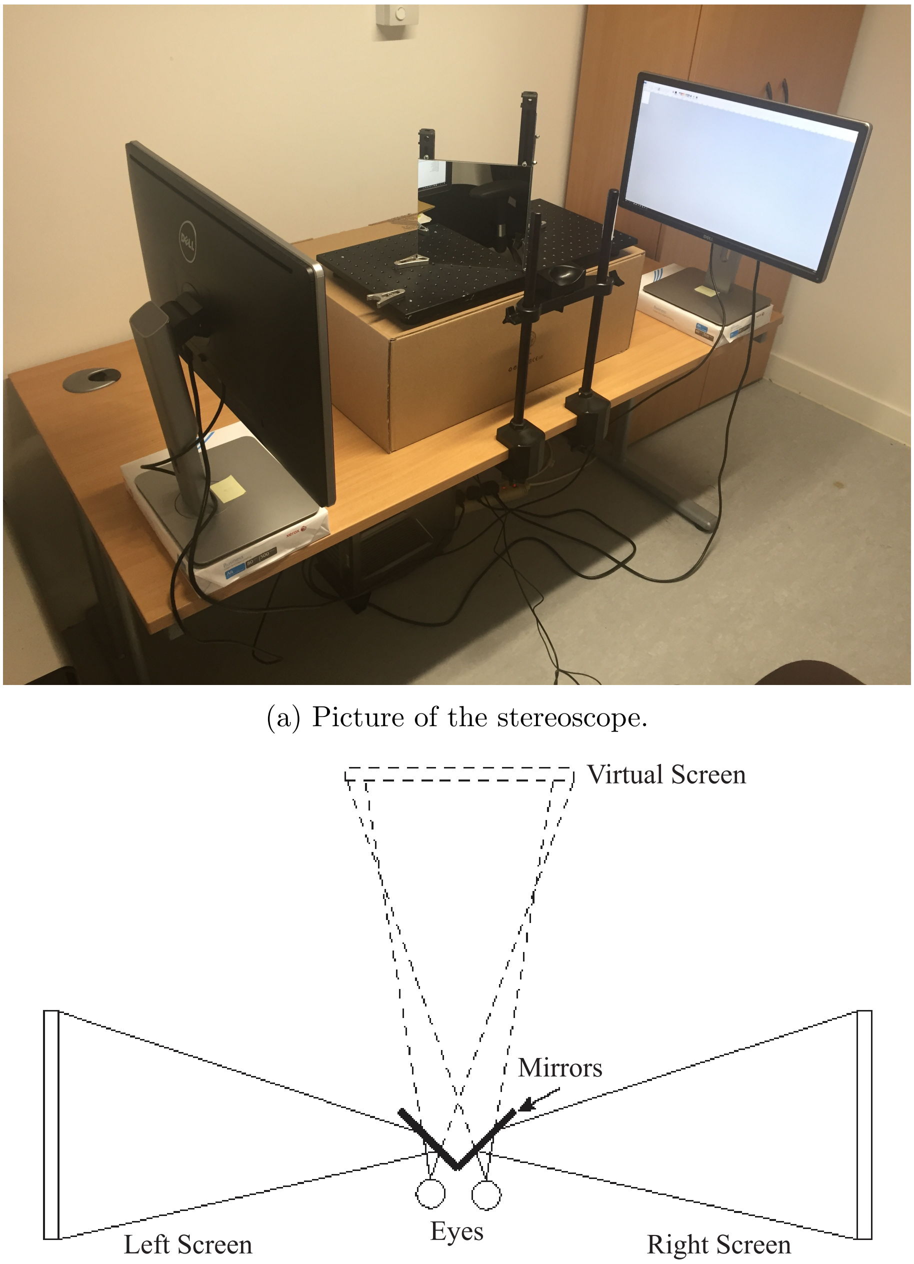


Figure 3.10: Vertical arrangement of the reference (top) and test (bottom) stimuli used in the experiment.

of the Gaussian filter is zero, the Gaussian kernel is a Dirac delta function, and no blur is introduced. The Gaussian filter was applied, because according to [137], defocus-based effects of lens aberrations in images can be modeled with Gaussian blur. Reference stimuli without SM, *i.e.*, with symmetric blur applied to the two views, were obtained by using the same standard deviation σ of the Gaussian filter in both views. From a given reference stimulus, test stimuli with SM were generated by adding $\Delta\sigma$ to the standard deviation σ of one view, and other test stimuli were generated with $\Delta\sigma$ added to the σ of the other view. The reference and test stimuli were vertically arranged as shown in Figure 3.10.

A total of 34 reference stimuli were shown in the experiment. 30 reference stimuli had symmetric blur applied to the two views defined by the standard deviations σ $\{0, 0.5, 1, 2, 5\}$ arcmin, Michelson contrasts $\{0.20, 0.50, 0.98\}$, and disparity equal to zero. The remaining four reference stimuli had no blur ($\sigma = 0$), Michelson contrast 0.5, and disparities $\{-67.4, 67.4\}$ arcmin. We selected the σ values similar to [138], and we also checked the σ histograms of image datasets [2, 1] to be sure to cover most of the σ values of these datasets. Moreover, we intentionally chose a low, medium, and high contrast. Regarding the disparities, we selected a positive and negative disparity large enough to cover most of the possible disparity range of S3D images.

Procedure: In order to measure the JNSM, we used the method of limits [132]. In the experiment, at the moment when each reference stimulus was initially shown, the test



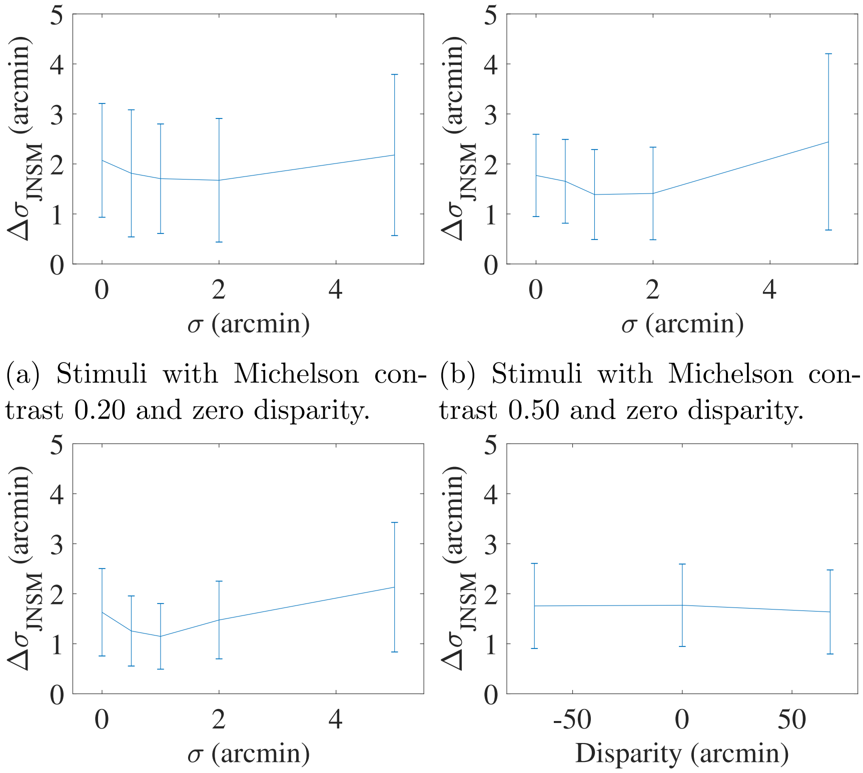
(a) Picture of the stereoscope.

(b) Technical illustration of the stereoscope. Figure taken from [139].

Figure 3.11: Wheatstone stereoscope used in the experiment.

stimulus was identical to it ($\Delta\sigma = 0$). At each second, the SM of the test stimulus was automatically increased by adding 0.07 arcmin to $\Delta\sigma$. The task of the subject was to indicate when she/he started to see a difference between the two stimuli.

Apparatus: For the experiment, we built a Wheatstone stereoscope [140] shown in Figure 3.11 in order to avoid crosstalk. Our stereoscope has two mirrors at 45 degrees fixed on an optical breadboard, two Dell P2415Q monitors, and a chin rest. The effective monitor size is 29.6cm × 52.7cm, the monitor resolution is 3840×2160 pixels, the viewing



(a) Stimuli with Michelson contrast 0.20 and zero disparity.

(b) Stimuli with Michelson contrast 0.50 and zero disparity.

(c) Stimuli with Michelson contrast 0.98 and zero disparity.

(d) Stimuli with Michelson contrast 0.50 and $\sigma = 0$.

Figure 3.12: Plots of the average $\Delta\sigma_{JNSM}$ with the standard deviation of the subjects' $\Delta\sigma_{JNSM}$ illustrated with a vertical bar. σ is the standard deviation of the symmetric Gaussian blur applied to the reference stimulus.

distance from the monitors is 0.7m, and the visual resolution is 89 pixels/degree. The monitors were carefully calibrated with the X-Rite i1Display Pro colorimeter and the DisplayCAL application. The white point was set to 6500K, the white level to 200 cd/m2, and the gamma to 2.2.

Subjects: In total, 23 subjects, 19 males and four females, took part in our experiment. The subjects were aged between 22 and 52, with an average of 32 years. The subjects had a normal or corrected-to-normal vision.

Data Analysis: The JNSM is expressed here as $\Delta\sigma_{JNSM}$, which is equal to the smallest $\Delta\sigma$ that generates a test stimulus perceived differently than the corresponding reference stimulus. The final $\Delta\sigma_{JNSM}$ value is obtained by averaging the subjects' $\Delta\sigma_{JNSM}$ values. Figure 3.12 shows the plots of the final $\Delta\sigma_{JNSM}$. First, as already observed in [132], the $\Delta\sigma_{JNSM}$ of the gratings with different contrasts are similar. Second, interestingly symmetric blur has an influence on the JNSM: starting from the grating without symmetric Gaussian blur (σ equal to 0) the $\Delta\sigma_{JNSM}$ initially decreases, and around σ 1 arcmin the $\Delta\sigma_{JNSM}$ begins to increase. Third, it can also be observed that the $\Delta\sigma_{JNSM}$ remains nearly constant across different disparities. For this reason, disparity is not considered in our SM detection method. Based on the large stimuli parameter ranges considered in the experiment, we can conclude that the studied stimuli characteristics, *i.e.*, contrast, symmetric blur, and disparity, do not have a large influence on the JNSM in general.

Proposed Method

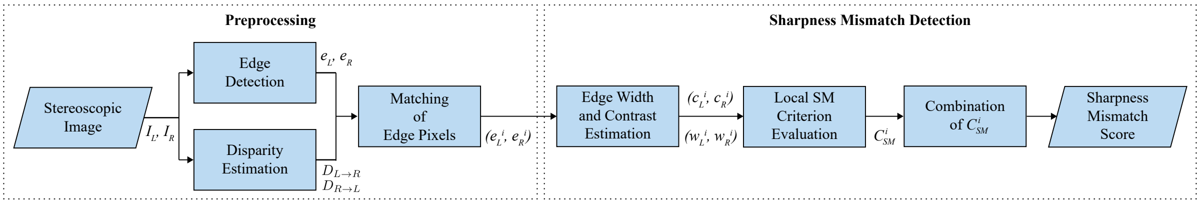


Figure 3.13: Overview of the processing steps of the proposed method for sharpness mismatch detection.

This section presents the new method for the detection of SM in S3D images based on the JNSM, which was inspired by [31]. Since when the SM is below the JNSM, it is suppressed by the interocular blur suppression, the proposed method analyzes corresponding edge pixels between the two views and checks whether their SM is above the JNSM. As shown in Figure 3.13, the proposed method is divided into a preprocessing step and an SM detection step.

In the preprocessing step, the disparity is estimated and a consistency check is computed as described in Section 3.3.3. In parallel to the disparity estimation, edge pixels $e_L \in I_L$ and $e_R \in I_R$ are extracted in both images using the Canny edge detector [129]. Then, the edge pixels between the two views are matched, obtaining edge pixel pairs (e_L^i, e_R^i) with $i = 1 \ldots N$.

In the actual SM step, for each edge pixel $e^i_{[L,R]}$ the edge width $w^i_{[L,R]}$ and contrast $c^i_{[L,R]}$ are estimated using the method in [130]. The edge width $w^i_{[L,R]}$ is then converted into the standard deviation $\sigma^i_{[L,R]}$ of the Gaussian filter that, when applied to a step edge, generates an edge with the same width.

Next, for each matched edge pixel pair (e_L^i, e_R^i) a local SM criterion C^i_{SM} is evaluated to check if the SM of the edge pixel pair is larger than the JNSM. In particular, the criterion checks whether the difference $|\sigma_L^i - \sigma_R^i|$ is larger than the $\Delta\sigma^i_{JNSM}$ of an edge with contrast $(c_L^i + c_R^i)/2$ and Gaussian blur standard deviation $min(\sigma_L^i, \sigma_R^i)$. In our method, $\Delta\sigma^i_{JNSM}$ is obtained by bilinear interpolation of the experiment data. The local SM criterion C^i_{SM} is formally expressed as follows:

$$C^i_{SM} = \mathbb{1}_{\Delta\sigma^i_{JNSM} \leq |\sigma^i_L - \sigma^i_R|}, \tag{3.12}$$

where $\mathbb{1}$ is an indicator function, which is equal to one if the condition $\Delta\sigma^i_{JNSM} \leq |\sigma^i_L - \sigma^i_R|$ is true, and zero otherwise. Finally, the results of the N local SM criteria are averaged to obtain the final score:

$$S = \frac{1}{N}\sum_{i=1}^{N} C^i_{SM}. \tag{3.13}$$

3.5.3 Evaluation

Standard S3D Images

In this section, the SM detection methods are evaluated based on standard S3D images. In this evaluation, we compare, the histogram-based SM detection method (HSMD) from Section 3.5.1, the just noticeable SM-based detection method (JNSMD) from Section 3.5.2, the method introduced in Narvekar *et al.* [119] (cumulative probability of blur detection, CPBD), and the state-of-the-art Liu *et al.*'s method [31] (probability of sharpness mismatch, PSM).

The performance comparison was evaluated based on two datasets: LIVE 3D Phase II [2]. and Ningbo 3D Phase I [1]. These two datasets were obtained by introducing different degrees of distortions to some artifact-free stereoscopic reference images.

The LIVE 3D Phase II dataset consists of 8 reference S3D images and 360 images obtained by symmetrically and asymmetrically distorting the reference images with 5 different types of distortions (compression using the JPEG and JPEG2000 compression standards, additive white Gaussian noise, Gaussian blur and a fast-fading model based on the Rayleigh fading channel). For the evaluation of SM, we only took the subset of 24 asymmetrically distorted images with Gaussian blur, since defocus-based effects of lens aberrations can be modeled with Gaussian blur [137].

The Ningbo 3D Phase I dataset consists of 10 reference S3D images and 370 images obtained by distorting the right view of the reference images with 4 types of distortions (compression using the JPEG and JPEG2000 compression standards, additive white Gaussian noise, and Gaussian blur). For the evaluation of SM, we again only took a subset of 100 asymmetrically distorted images with Gaussian blur.

For each image, the datasets provide a subjective difference mean opinion score (DMOS) in the range 0-100 that was obtained through subjective experiments.

For the method comparison, we evaluated the correlation between the subjectively obtained DMOS and the SM scores of the methods by fitting a logistic function to transform the SM scores to DMOS. A well-suited logistic function was proposed by the Video

Table 3.4: Ningbo 3D Phase I dataset [1].

	PLCC	SROCC	RMSE	MAE
CPBD [119]	0.8359	0.587	2.054	1.739
PSM [31]	0.8542	0.5426	1.945	1.599
HSMD [50]	**0.8708**	**0.6296**	**1.839**	**1.455**
JNSMD [53]	0.8604	0.5496	1.906	1.551

Table 3.5: LIVE 3D Phase II dataset [2].

	PLCC	SROCC	RMSE	MAE	OR
CPBD [119]	0.7069	0.4307	5.091	4.192	0.02
PSM [31]	0.9276	0.7572	2.913	2.094	0
HSMD [50]	**0.9548**	**0.8205**	**2.152**	**1.563**	**0**
JNSMD [53]	0.9217	0.7769	2.944	2.092	0

Quality Expert Group in [141] and is defined by

$$\mathrm{DMOS}_p(S) = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{S-\beta_3}{\|\beta_4\|}}} + \beta_2, \tag{3.14}$$

where DMOS_p is the predicted DMOS of the SM score S, and β_{1-4} are parameters that are computed during the fitting.

After the fitting of the logistic function, the following performance metrics were applied in order to evaluate how well the logistic function predicts the subjective DMOS: Pearson's linear correlation coefficient (PLCC), Spearman's rank ordered correlation coefficient (SROCC), root mean squared prediction error (RMSE), mean absolute prediction error (MAE), and outlier ratio (OR). PLCC and SROCC measure the prediction accuracy and the monotonicity, respectively. The larger these two metrics are, the more accurate and monotonic the prediction is. For RMSE, MAE, and OR, the smaller the metric, the better the performance of the prediction is. Note that the LIVE 3D Phase II dataset does not provide the standard deviation of the DMOS, which is necessary to compute the OR.

Table 3.4 and 3.5 show the performance metrics for the two datasets. The best values are marked in bold. As can be seen, HSMD outperforms CPBD, PSM, and JNSMD for all metrics. A weakness of HSMD, PSM, and JNSMD compared to CPBD is the need for disparity maps. For this reason, geometrical misalignment may negatively influence the analysis, but this does not compromise the state-of-the-art performance of HSMD.

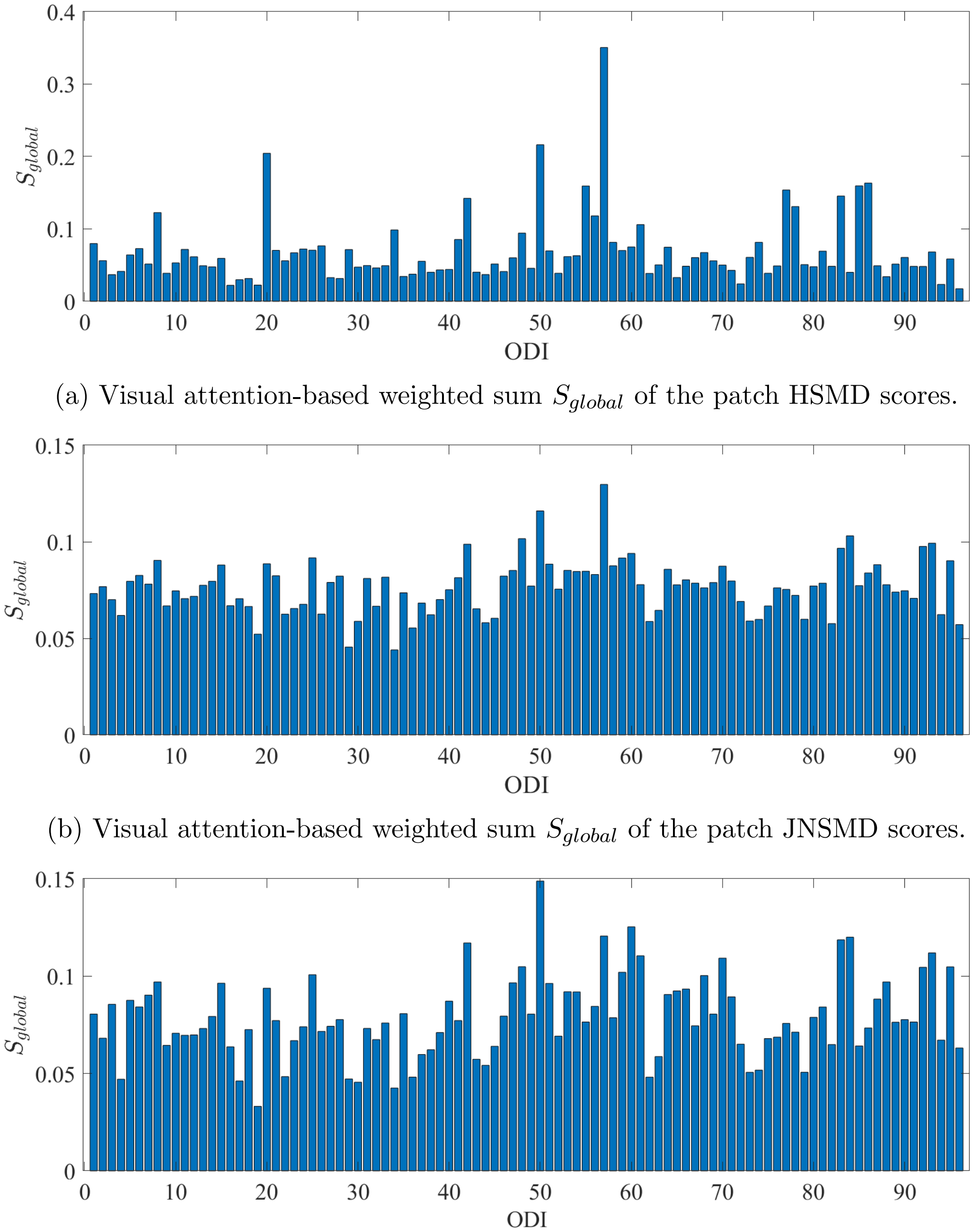
S3D Omnidirectional Images

In this chapter, our two SM detection methods together with Liu *et al.*'s PSM method [31] are evaluated based on the ODI dataset introduced in Section 3.4. For the computation of the global SM scores presented in Section 3.3.4, we define $g''(x) = x$. Figure 3.14 shows the visual attention-based weighted sum S_{global} of the patch scores defined by Equation 3.3.

The ODIs with the smallest global scores are ODI 19, ODI 36, and ODI 44. Among these ODIs, ODI 19 has a low score because it was converted from 2D to 3D in post-production. SM is unlikely in post-converted images as they are generated using depth-image-based rendering. On the other hand, ODI 44 was captured with Jaunt, an off-centered slit camera.

The ODIs with the largest global scores are ODI 42, ODI 48, ODI 50, ODI, 57, and ODI 83. ODI 42 was captured with the VUZE camera that has 4 stereo camera pairs, while ODI 50 and ODI 57 were captured with Panocam's POD 3D that consists of 9 stereo camera pairs capturing the left and right view of the ODI independently.

Figure 3.15-3.16-3.17 shows the analysis of some ODIs with visual attention maps, the visualization of patch scores, and close-ups. These figures show how SM is correctly detected in the ODIs. In addition to SM, the SM detection methods are sometimes also able to detect asymmetric distortions that can be confused with SM, like the mismatch between corresponding glares (ODI 21 and ODI 51), the presence of stitching and blending artifacts (ODI 45 and ODI 48), and contamination (ODI 73).



(a) Visual attention-based weighted sum S_{global} of the patch HSMD scores.

(b) Visual attention-based weighted sum S_{global} of the patch JNSMD scores.

(c) Visual attention-based weighted sum S_{global} of the patch PSM scores.

Figure 3.14: Visual attention-based weighted sum S_{global} of the patch scores obtained with different sharpness mismatch detection methods.

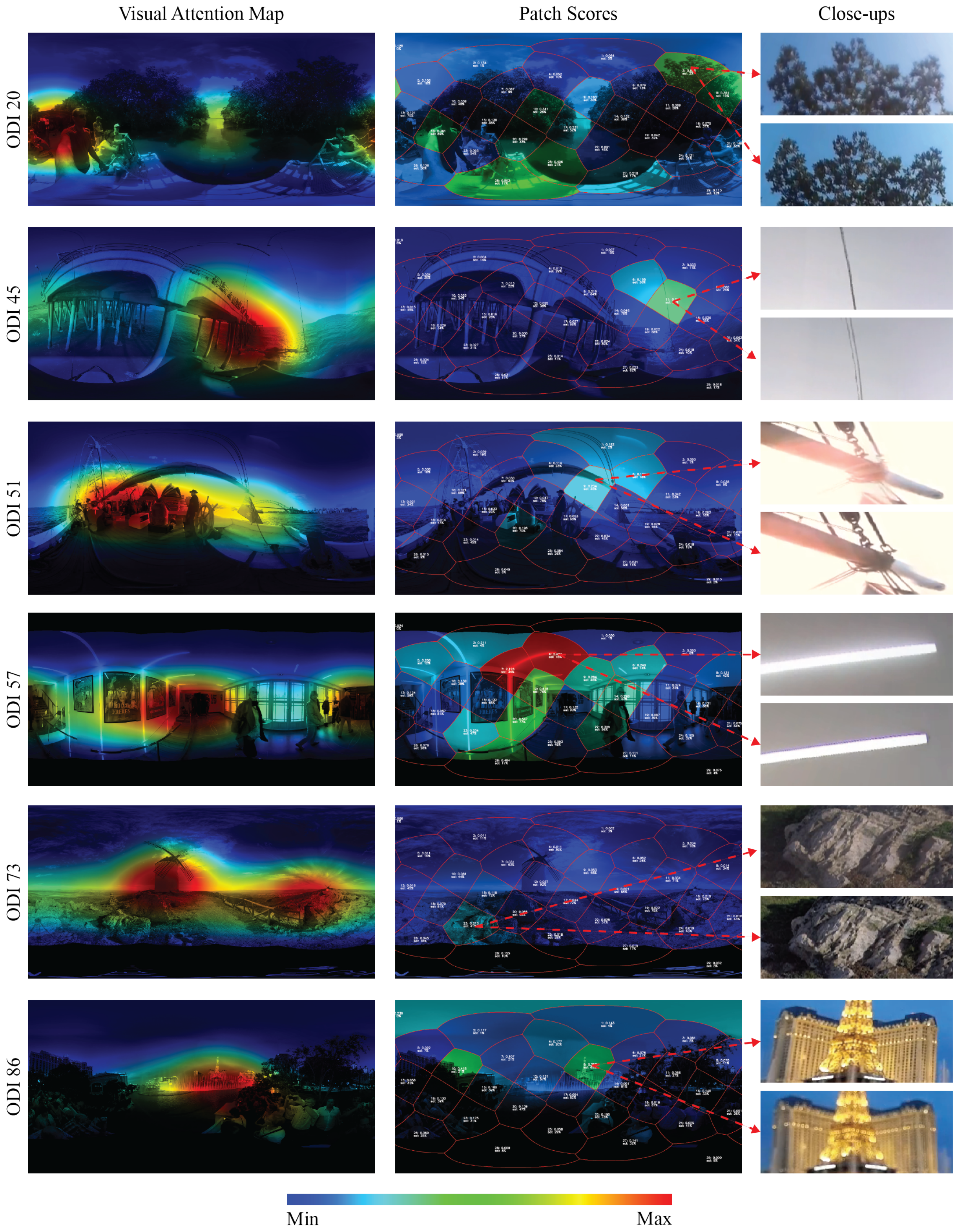


Figure 3.15: Sharpness mismatch analysis based on HSMD.

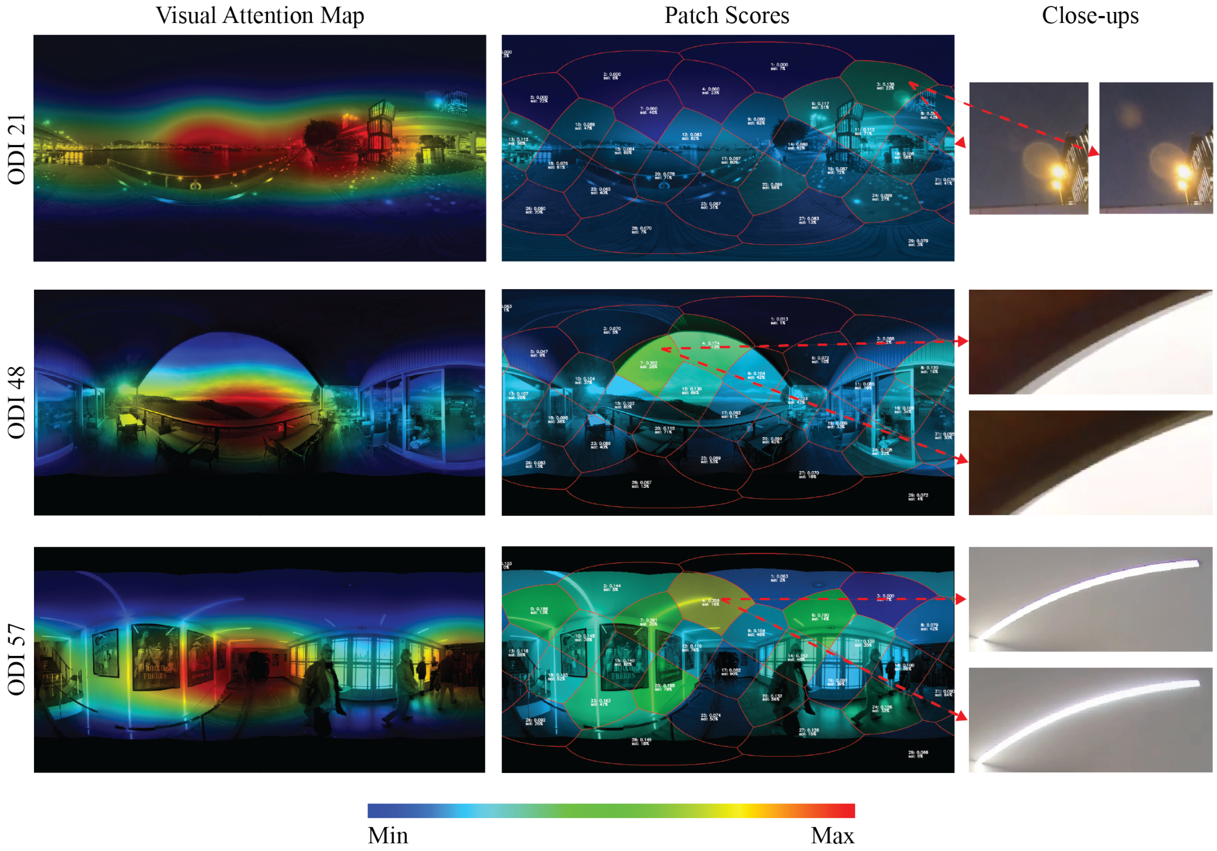


Figure 3.16: Sharpness mismatch analysis based on JNSMD.

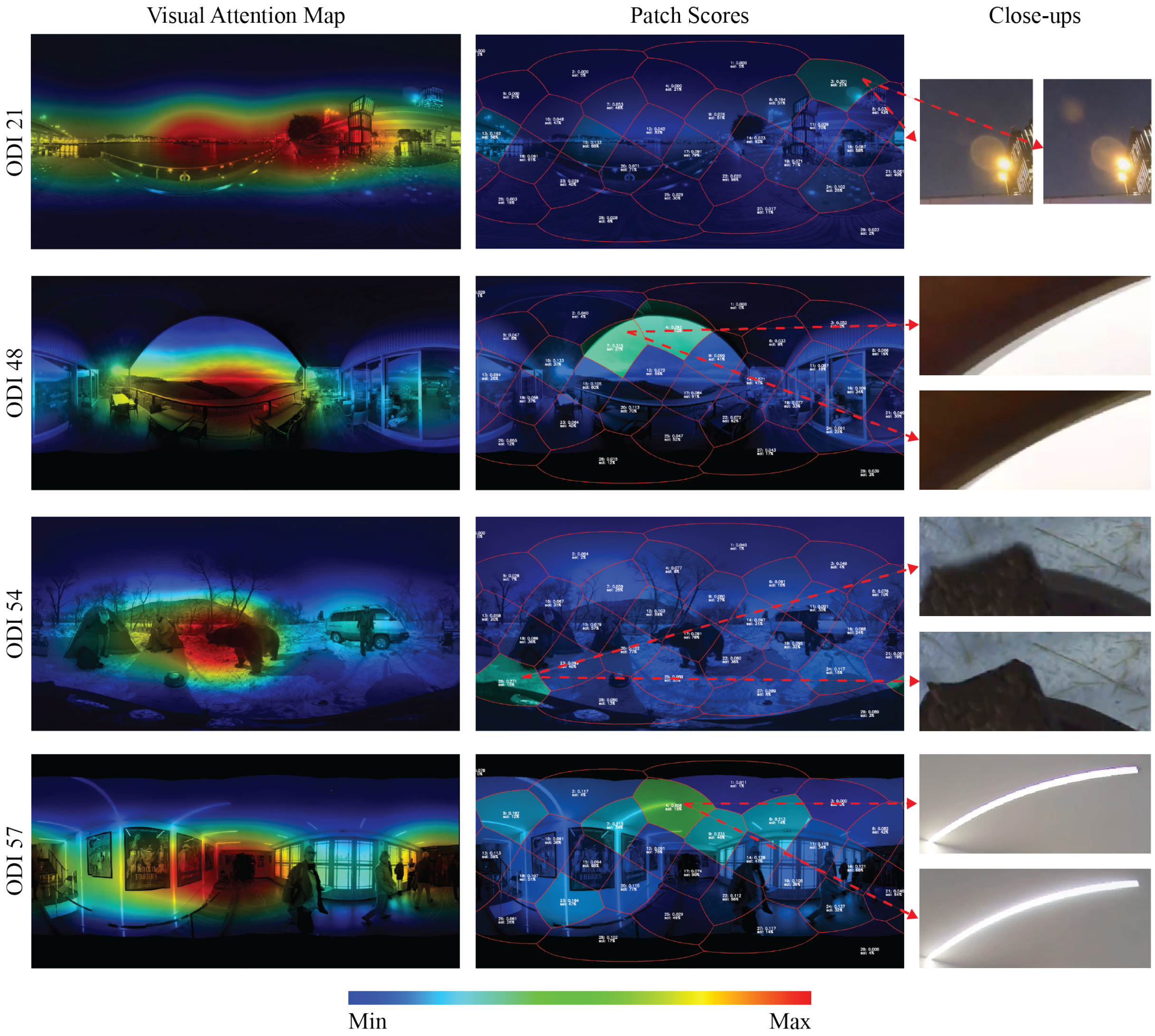


Figure 3.17: Sharpness mismatch analysis based on PSM.

3.6 Color Mismatch Detection

This section introduces a new method for the detection of color mismatch (CM) in standard S3D images published in [49] and extended to S3D ODIs based on our framework (Section 3.3).

3.6.1 Proposed Method

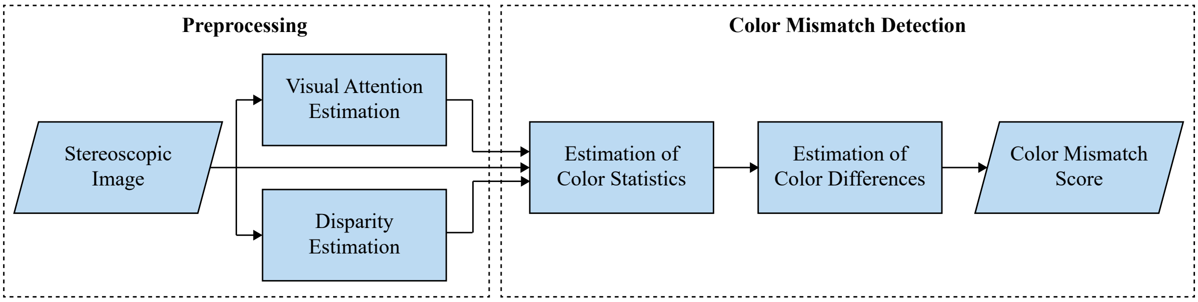


Figure 3.18: Overview of the processing steps of the proposed color mismatch detection method.

If there is no CM in S3D images, the color statistics of common regions between the two views are very similar. The more CM is present, the more different these color statistics are. For this reason, the proposed method compares color statistics of common regions between the two views similar to the statistics used in the color transfer method by Reinhard *et al.* [142]. The proposed CM detection method is divided into a preprocessing and an actual CM detection phase. An overview of the processing steps is illustrated in Figure 3.18.

In the preprocessing step, first the visual attention $\psi(p)$ is computed at each pixel p (Section 3.3.2). Then, disparity is estimated between the left and right images, and the consistency check described in Section 3.3.3 is applied to exclude inaccurate correspondences. Based on the validated disparity map $D_{L\rightarrow R}$, regions that are present in both the left and right images are detected. A pixel (x, y) in the left image and a pixel (x', y) in the right image belong to the same region if $x' = x - D_{L\rightarrow R}(x, y)$. We define the sets of pixels in the left and right image belonging to the common regions as Ω_L and Ω_R, respectively.

Inspired by Reinhard *et al.* [142], for all pixels belonging to the common regions Ω_L and Ω_R, the color statistics mean and standard deviation of the color channels are extracted. Reinhard *et al.* extract color statistics that differ from ours in two aspects. The first is the color space, *i.e.*, instead of using the $l\alpha\beta$ color space, we extract the statistics from the Lab color space since it is perceptually uniform. The second aspect is the integration of visual attention, which is used in our approach to weight the pixels.

Let's define $I_L(p)$ and $I_R(p)$ as the colors at the pixel p defined in the Lab color space in the left image I_L and right image I_R, respectively. Then, we first extract the visual attention-based mean for each color channel as follows

$$\mu_X = \frac{1}{\sum_{p \in \Omega_X} g'(\psi(p))} \sum_{p \in \Omega_X} I_X(p) \cdot g'(\psi(p)), \tag{3.15}$$

where $X \in \{L, R\}$, and $g'(\psi(p))$ is the user-defined weighting function that controls the influence of the visual attention $\psi(p)$. For the generation of the results, g' is a piece-wise linear function, which is defined as

$$g'(x) = \begin{cases} 5 \cdot x, & x \leq 0.2, \\ 1, & x > 0.2, \end{cases} \tag{3.16}$$

i.e., pixels with visual attention larger than 20% have a maximal weight when calculating the scores. We also extract the visual attention-based standard deviation for each color channel defined as

$$\sigma_X = \frac{1}{\sum_{p \in \Omega_X} g'(\psi(p))} \sum_{p \in \Omega_X} (I_X(p) - \mu_X)^2 g'(\psi(p)). \tag{3.17}$$

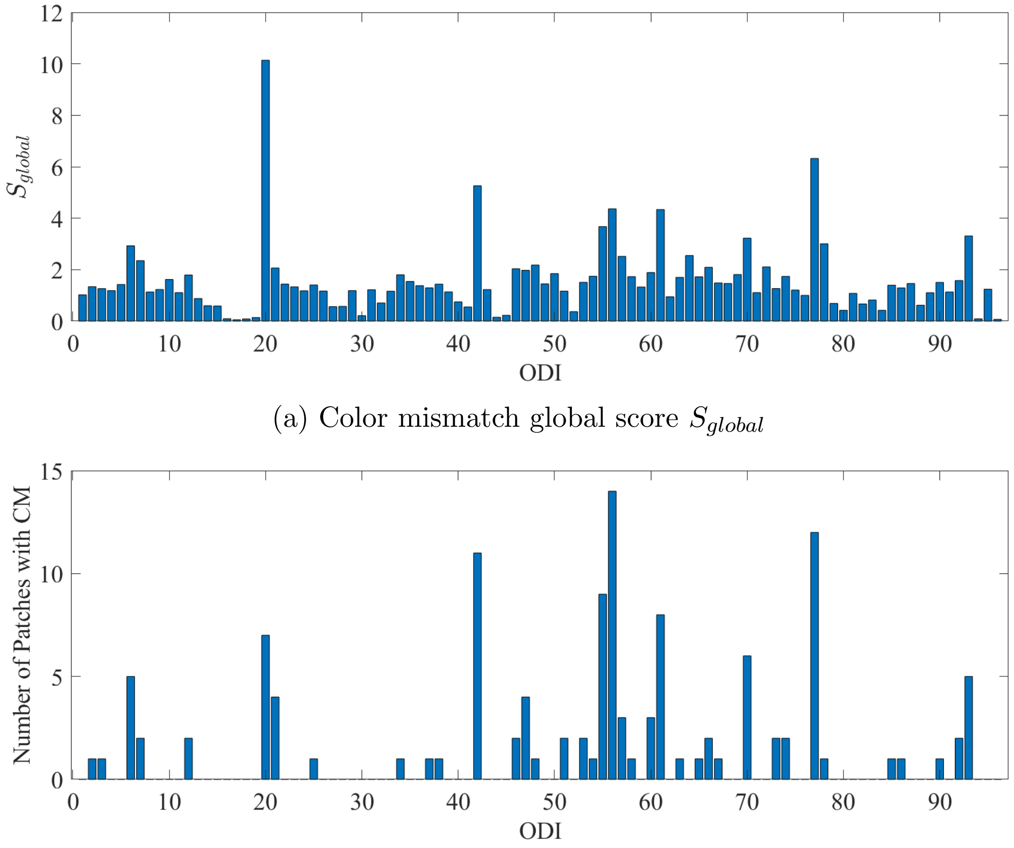
Finally, the CM score is computed according to:

$$S = \sqrt{\|\mu_L - \mu_R\|^2 + \lambda \|\sigma_L - \sigma_R\|^2}, \tag{3.18}$$

where λ is a tuning parameter that was set to one for the generation of the results.

3.6.2 Evaluation

In order to demonstrate the performance and usability of our proposed method, we evaluated the quality of 96 ODIs from our dataset. For each ODI, we computed the two global scores proposed in Section 3.3.4, *i.e.*, the global score S_{global} and the number of patches with detected CM defined by Equation 3.3 and 3.4, respectively, where the visual attention weight function g'' is the identity function and γ is equal to 0.2. Figure 3.19 shows these global scores. ODI 16 and ODI 17, which were converted from 2D to 3D in post-production, have the lowest score S_{global}, and no patch with CM was detected. CM is very unlikely for post-converted images as the stereoscopic views are generated using depth-image-based rendering. ODI 56 and ODI 61 have a high score S_{global}. These ODIs were captured with Panocam POD 3D which uses 9 stereo camera pairs capturing the left and right view of the ODI independently. ODI 20 was captured with the Vuze



(a) Color mismatch global score S_{global}

(b) Number of patches with color mismatch.

Figure 3.19: Color mismatch analysis of the ODI dataset.

VR camera which uses 4 stereo camera pairs. Here the score S_{global} is the highest of all ODIs under evaluation (see Figure 3.20). ODI 42, ODI 56, and ODI 77 are ODIs that are characterized by more than 10 patches with CM.

Figure 3.20 exemplary shows six ODIs with detected CM, together with their visual attention maps, the visualization of the patch scores, and close-ups of detected regions. As seen in the figure, our proposed method detects and highlights patches with CM correctly. It can also be observed, that the CM detection method is able to detect artifacts like contamination (ODI 37) and different glares (ODI 55).

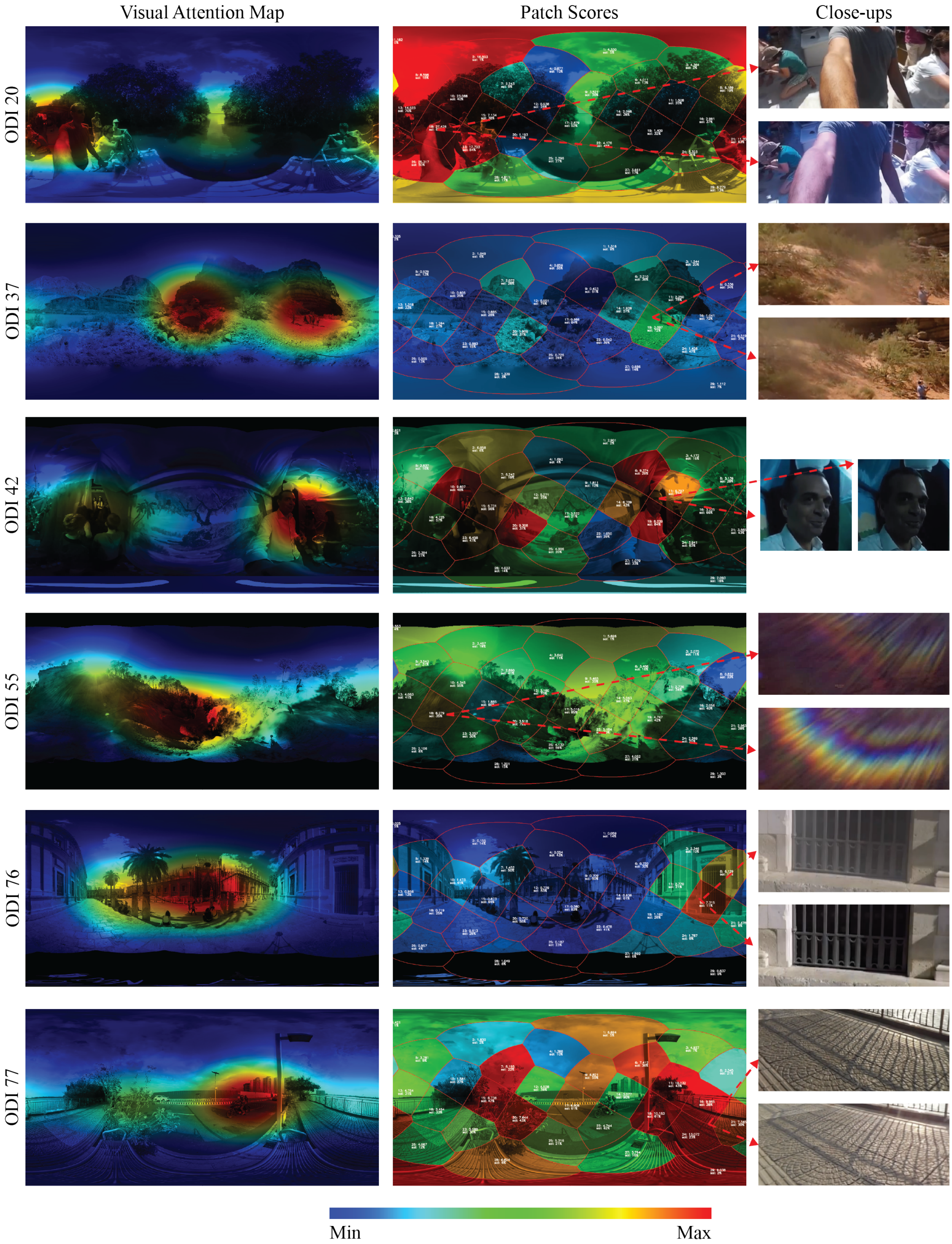


Figure 3.20: Color mismatch analysis.

3.7 Conclusions

The core of this chapter is a general framework for artifact detection in S3D ODIs based on planar Voronoi patches and visual attention. The framework takes artifact detection methods for S3D standard images and extends them to S3D ODIs. The framework was applied for the detection of two artifacts, namely, SM and CM. For SM, two detection methods were also proposed. The first is based on edge contrast and width, and the second is based on the just noticeable SM measured in a subjective experiment. According to the evaluation of these methods, the first approach has the highest performance among the state-of-the-art methods, while the second approach has a lower performance but still comparable to the performance of the first method. For CM, a detection method based on color statistics was proposed. Finally, for the evaluation of the artifact detection framework, a dataset of S3D ODIs was created with viewport trajectories collected during a subjective experiment.

In the future, the weaknesses of the proposed solutions could be investigated and fixed. For example, even if the SM detection method JNSMD was developed based on data collected during a psychophysical experiment, it did not reach the performance of the other proposed method HSMD. To improve JNSMD, a more accurate estimation method of the edge standard deviation σ could be used, visual attention could be integrated to weight regions according to their relevance, and the JNSM threshold could be replaced with the probability of SM perception, which would require a new subjective experiment to measure this probability. Another point left for future research is the estimation of the thresholds necessary to decide whether an artifact is perceived by the viewer, since the proposed artifact detection methods only compute a score. Subjective tests with potential end-users could be organized in order to find the detection thresholds and fine-tune other hyperparameters. Moreover, the proposed artifact detection methods are based on disparity estimation, which could be inaccurate in the presence of strong vertical disparity (*e.g.* at the pole caps), or due to homogeneous regions (*e.g.* sky). In future research, the influence of inaccurate disparity could be investigated, and if needed more robust disparity estimation methods could be applied. Another point is the use of visual attention estimated by algorithms instead of visual attention estimated from viewers. We evaluated our framework based on visual attention estimated from viewers, which is usually difficult to obtain. In practice, automatic methods for visual attention estimation should be applied. Furthermore, in this chapter, we studied two common artifacts in particular, namely, CM and SM. Nevertheless, other common artifacts that characterize ODV, like stitching and blending artifacts, could be studied.

Chapter 4

Color Mismatch Correction for Stereoscopic 3D Omnidirectional Images

4.1 Introduction

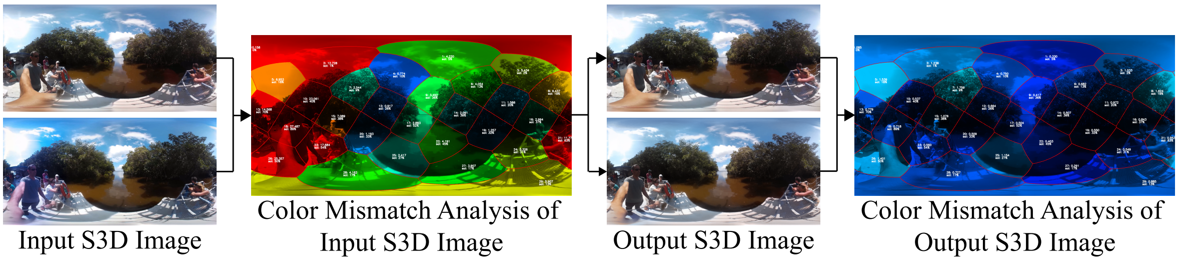

Input S3D Image
Color Mismatch Analysis of Input S3D Image
Output S3D Image
Color Mismatch Analysis of Output S3D Image

Figure 4.1: Color mismatch correction.

Stereoscopic 3D images can have color mismatch (CM) between the left and right image due to reasons like different camera and lens characteristics, different reflections resulting from different camera positions and orientations, polarized light, etc. The presence of CM can reduce the quality of experience (QoE) and cause problems when processing the S3D images [25], *e.g.*, for depth estimation. For this reason, it is important to remove this artifact that occurs frequently.

Color correction of S3D images (Figure 4.1) consists of selecting either the left or right image as the reference image and correcting the other image, which is called the target image. The reference image contains the color information that must be mapped to the target image while preserving the structure information, *i.e.*, the edges, present in the target image.

For color correction, there are two categories of approaches, namely, global [143, 144] and local [145] methods. Global methods estimate a single color transformation, while local methods estimate different local transformations. Global methods usually analyze the color distributions of the reference and target images, but they fail in the presence of local CM. Local methods can fix local CM using correspondences between the target and reference image, but they are usually sensitive to the quality of the estimated correspondences.

For CM correction, there are already several methods based on traditional visual computing techniques [143, 144, 145, 52, 32, 3, 146, 147, 51], but only two published methods based on deep learning [148, 57] including our approach [57]. In this chapter, we propose an approach based on traditional visual computing techniques, *i.e.*, , planar Voronoi patches and a color transfer method, and we also present new deep learning-based solutions. In particular, we propose local methods for ODIs that are robust to inaccuracies in the correspondence estimation. Some of the contributions presented in this chapter were published in [51, 57].

4.2 Background

This section first presents the color correction approaches based on traditional visual computing techniques. Then, the field of deep learning is briefly introduced together with the single deep learning-based color correction method that had been published before ours.

4.2.1 Color Correction Based on Traditional Visual Computing Techniques

In the computer vision and multi-view video processing communities, the initial efforts to solve CM between multiple views used exposure compensation (or gain compensation) [143]. This approach adjusts the gain level of images to compensate for appearance differences caused by different exposure levels. However, this approach may fail in the case of local differences.

The authors of [144] propose a simple method to compute 3D lookup tables with a non-linear process that minimizes the colorimetric properties of the source images. Wang *et al.* [145] proposed a robust algorithm to correct the color discrepancy between images, which neither requires a color calibration chart/object nor explicitly compensates for the image as a whole. Instead, they correct the image region by region using local feature correspondences.

Dudek *et al.* [149] proposed a combination of global and local color correction. While the global color correction is performed using a classic histogram-based method applied to the entire image, for local color correction a region-based approach based on optical flow estimation is used. More recently, Dudek *et al.* [52] extended their previous method to S3D ODVs. The global color correction was substituted by an iterative method, while the local color correction consists of applying the local color correction approach from [149] to two patches extracted from the ODV. In [32], a method is proposed that combines global and local color information to correct color discrepancies between stereoscopic image pairs. In the first step, the algorithm uses dense stereo matching and global color correction to initialize color values, and in the second step, it improves the local color smoothness and global color consistency of the corrected image while maintaining the colors of the first step as much as possible.

For large baseline multi-view video, Ye *et al.* [150] introduced a robust color correction method that enforces spatio-temporal color consistencies and gradient preservation by solving a global optimization problem. The authors of [151] proposed an effective color correction method for multi-view image stitching which first finds coherent content regions in inter-image overlaps, where reliable color correspondences are extracted, and then parameterizes a color remapping curve as a transform model, and expresses the constraints of color consistency, contrast, and gradient in a uniform energy function.

While many methods have been proposed for stereo and multiview color correction, not so many have explicitly considered the color correction of S3D ODIs. Distortions introduced when ODIs are stored in ERP format can cause errors during correspondence estimation, limiting the quality of color correction results generated by applying the methods not designed for S3D ODIs directly to ERP format.

The image processing and computer graphics communities were developing similar color manipulation methods, called color transfer techniques. These methods transfer the color feel from a palette image to a target image and assume that the content of the images is different. The earliest work in this area was by Reinhard *et al.* [142], who proposed transforming the mean and standard deviation of each color channel in the target image to match that of the palette image. Since then, more complex techniques have been used to model the color distributions of the images more accurately, including histograms and Gaussian mixture models [152, 153]. While global color transfer functions are often used, including affine, radial basis, and optimal transport functions [154, 155, 156], local techniques have also been proposed to allow for more flexibility in the recoloring [157, 158]. An efficient method was developed by Pitie *et al.* [3]. It first estimates a global color function that converts the color distribution of one image into another, and then it reduces possible grain artifacts generated by the color function. Recently, Grogan and Dahyot

[146, 159] proposed a color transfer technique that could also be extended to take into account color correspondences between the target and palette images, so that the method could be used to color correct images of the same scene. They showed that this method performs as well as other state-of-the-art color corrections techniques, with the advantage of being more robust to correspondence outliers. In this chapter, this method is extended to ODIs in order to reduce local CM.

4.2.2 Deep Learning

Deep learning [160] is a subfield of machine learning that studies deep artificial neural networks, that is, neural networks (NN) with many layers. A NN is a group of neurons connected together, where each neuron performs a simple operation, usually a weighted sum of its inputs plus a non-linearity. Despite this simple operation, when the neurons are combined together in a NN, they can perform complex operations. Each NN can model a large variety of functions parameterized based on the internal parameters of the NN, which can be learned through efficient optimization methods based on the backpropagation technique [161]. In the field of visual computing, a special type of NN is usually used, the so-called convolutional neural network (CNN) [160, 162]. This type of network allows a drastic reduction of parameters compared to traditional fully connected networks, and it works similarly to the human brain. Like the region of the human brain that processes visual information, the so-called visual cortex [163], CNNs are organized in layers. Moreover, nearby neurons in a layer represent nearby regions in the input image, and the deeper a neuron is in the CNN, the more complex are the stimuli the neuron responds to. One of the first CNNs was invented in 1998 by LeCun *et al.* [164] for the recognition of hand-written digits. CNNs became famous in 2012 when the CNN called AlexNet developed by Krizhevsky *et al.* [165] won the ImageNet Challenge [166]. In the following years, this challenge was won by more sophisticated and deeper CNNs. Among them, it is worth mentioning VGG [167], GoogLeNet [168], and ResNet [169]. Currently, CNNs has reached state-of-the-art performance in different visual computing tasks like object detection [170], optical flow estimation [171], image super-resolution [172], etc.

4.2.3 Color Correction Based on Deep Learning

To our knowledge, except ours [57], only Yuanyuan *et al.* [148] applies deep learning to the problem of CM correction in S3D images. This approach first applies the color correction method by Zheng *et al.* [32] obtaining an intermediate result that is then processed by SRCNN [7], which is a convolutional neural network that was developed for super-resolution. The results show the state-of-the-art performance of this method.

4.3 Voronoi-Based Approach for Stereoscopic 3D Omnidirectional Images

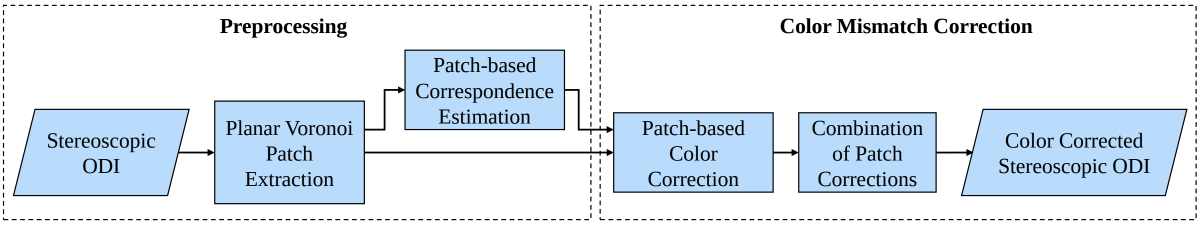

Preprocessing
Color Mismatch Correction
Stereoscopic ODI
Planar Voronoi Patch Extraction
Patch-based Correspondence Estimation
Patch-based Color Correction
Combination of Patch Corrections
Color Corrected Stereoscopic ODI

Figure 4.2: Voronoi-based color mismatch correction method (Voro-CMC).

This section presents a novel method for the correction of CM based on planar Voronoi patches, which we call Voronoi-based color mismatch correction method (Voro-CMC) and we published in [51]. In the case of S3D ODIs, the presence of local CM is more likely than in traditional S3D images, since S3D ODIs are often obtained by combining images captured by different cameras that can have characteristic and setting mismatches or different positions and orientations that can result in different reflections. Therefore, our method subdivides the S3D ODI into planar Voronoi patches, where global color correction transformations are locally fitted and then globally combined. The evaluation of the method shows that it is able to considerably reduce the CM. As illustrated in Figure 4.2, the method is divided into a preprocessing step and a CM correction step.

In the preprocessing step, the ODI is subdivided into planar Voronoi patches as described in Section 2.2. In the presence of disparity, it can occur that a region inside a planar Voronoi patch in one view is outside the corresponding planar Voronoi patch in the other view. In order to cope with the disparity, we add a border around the planar Voronoi patch when the patch is extracted, as shown in Figure 3.3. The number of patches and thus the size of each patch influence the reduction of the CM. If the CM is localized in a small region and the patch is large, then the proposed method could have difficulty in matching the colors between the two views. We empirically found that 30 patches are a good number for most of the ODIs that we processed.

Following the patch extraction, color correspondences between corresponding planar Voronoi patches of the target and reference view are estimated. We investigated two methods for the estimation of correspondences: the semi-global block matching approach [127] and the coarse-to-fine PatchMatch approach [173], but we found no significant difference between the color correction results generated using these approaches.

In the actual CM correction step, for each patch, we use the correspondences to estimate a color transformation that recolors the patch of the target view so that it is more similar to the reference view, using the method proposed in [146]. For a given patch, let's assume that we have the color correspondences $\{c_T^k, c_R^k\}_{k=1..N}$ defined in the RGB color space. Then, we fit two Gaussian mixture models GMM_T and GMM_R to the target colors c_T^k and reference colors c_R^k of the correspondences, respectively:

$$GMM_Y(x) = \frac{1}{\mathrm{N}} \sum_{k=1}^{\mathrm{N}} \mathcal{N}(x;\, c_Y^k,\, h\mathbf{I}), \quad \text{with } Y \in \{T, R\}, \tag{4.1}$$

where $x \in \mathbb{R}^3$ are color values, and each Gaussian $\mathcal{N}$ is associated with an identical isotropic covariance matrix $h\mathbf{I}$. The goal is to align the two Gaussian mixture models by warping the target one as follows:

$$GMM_T'(x|\theta) = \frac{1}{\mathrm{N}} \sum_{k=1}^{\mathrm{N}} \mathcal{N}(x;\, \zeta(c_T^k, \theta),\, h\mathbf{I}), \tag{4.2}$$

where ζ represents a parametric thin plate spline (TPS) transformation controlled by the parameter θ. Technically, the alignment between GMM_R and the warped GMM_T' is obtained by minimizing the $\mathcal{L}_2$ distance between them. This $\mathcal{L}_2$ technique was shown to be robust to correspondence outliers, and the smooth TPS function ensures that similar colors in the patch remain similar after recoloring, eliminating artifacts in the gradient of the image which can appear when using other re-coloring methods [152].

Once the transformations ζ_i have been estimated for each patch Π_i', they have to be combined to recolor the entire ODI of the target view. To ensure that there are no harsh color changes between patches in the recolored ODI, we use weight masks to blend the transformations. For each transformation ζ_i, a corresponding weight mask G_i is computed in ERP format. To compute the value of a pixel in the weight mask G_i, the spherical distance between this pixel and the centroid of the patch Π_i' in the spherical representation of the ODI is computed, and a Gaussian function is applied to it. In this way, in G_i, pixels that lie close to the patch centroid will have higher weights than those further away. Then, when recoloring the ODI I_T of the target view in ERP format to its corrected version I_T', the color of the pixel at location (j, k) is given by:

$$I_T'(j,k) = \frac{\sum_{i=1}^{M} G_i(j,k) \cdot \zeta_i(I_T(j,k), \theta_i)}{\sum_{i=1}^{M} G_i(j,k)}, \tag{4.3}$$

where M is the number of patches. In this manner, each local color transformation has the most influence in the area from which it is estimated, and the color transformations are smoothly blended without creating any artifacts at the patch borders.

4.4 Deep Learning-Based Approaches for Stereoscopic 3D Standard Images

Two general deep learning-based color correction approaches for S3D standard images are proposed in this section. In the first approach, the target and reference images are preprocessed and then fed to a CNN, while in the second approach, the target and reference images are fed directly to a CNN. These solutions are extended from standard to omnidirectional S3D images in Section 4.5.

4.4.1 Approach with Preprocessing

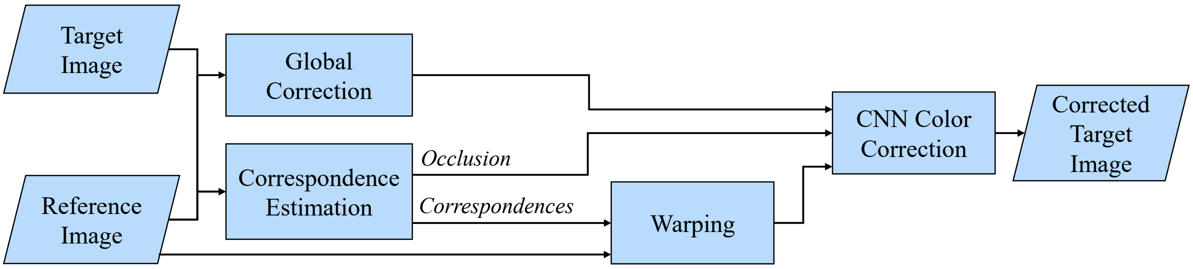


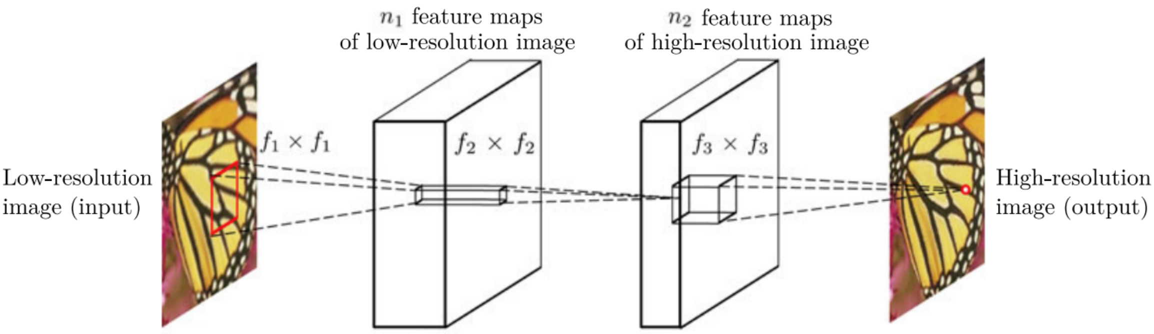
Figure 4.3: Approach with preprocessing (PreProcNet).

The first approach is called PreProcNet and consists of a preprocessing phase followed by a color correction phase illustrated in Figure 4.3. The idea behind PreProcNet is to combine the advantages of local and global color correction, that is, the ability of local correction to fix local CM, and the robustness of global correction in regions where correspondences are difficult to compute. More precisely, in the preprocessing phase, a globally and a locally corrected target image are computed. Next, these two corrected target images are fed to a CNN that computes the final corrected target image.

For the global color correction in the preprocessing phase, we apply the color transfer method by Pitie *et al.* [3]. On the other hand, the local color correction is obtained by warping the reference image into the target image based on correspondences computed by SIFT-Flow [174], which is an optical flow estimation method robust to CM and already used in another CM correction approach [32] and in a CM evaluation metric [175]. SIFT-Flow cannot estimate correspondences in the occluded regions, and in order to deal with them, an occlusion mask is computed and fed to the CNN.

Two CNNs were considered for the second phase, SRCNN [7] and U-Net [8]. PreProcNet with the first CNN is called SRCNN-PreProcNet and with the second CNN U-Net-PreProcNet.

SRCNN-PreProcNet



n_1 feature maps of low-resolution image
n_2 feature maps of high-resolution image
$f_1 \times f_1$
$f_2 \times f_2$
$f_3 \times f_3$
Low-resolution image (input)
High-resolution image (output)

Figure 4.4: SRCNN [7]. Figure taken from [7].

SRCNN [7], illustrated in Figure 4.4, is a simple network developed for image super-resolution, which consists of three convolutional layers. This network was already used for CM correction in [148].

U-Net-PreProcNet

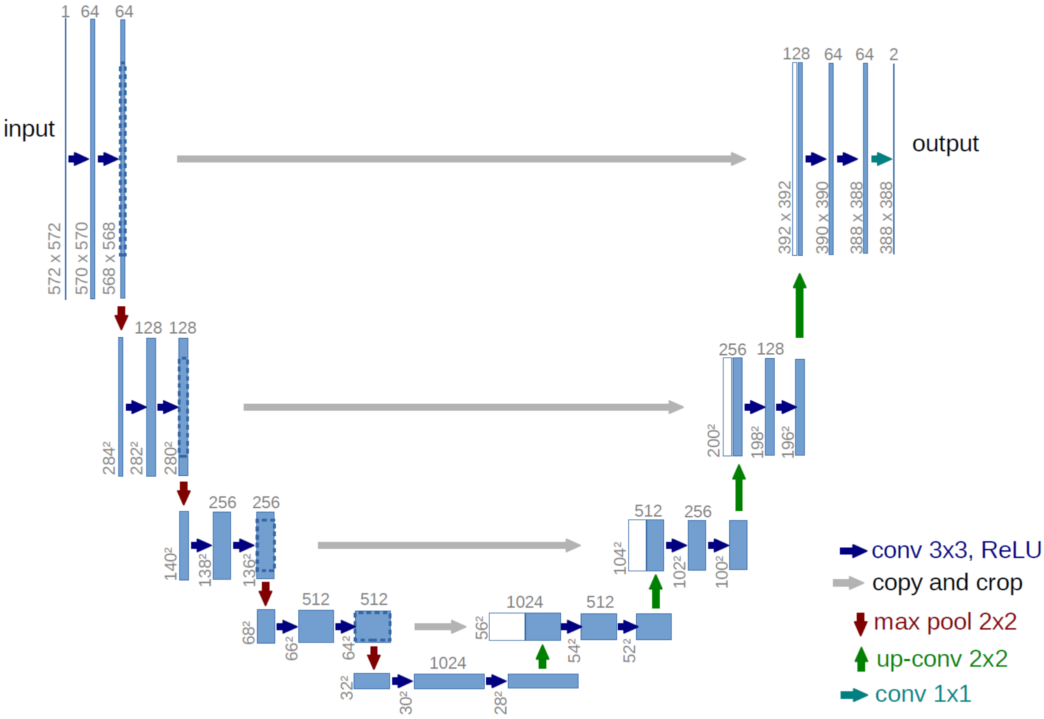


Figure 4.5: U-Net [8]. Figure taken from [8].

U-Net [8], illustrated in Figure 4.5, is a deep network originally developed for biomedical image segmentation and later applied to different image-to-image translation problems [176]. It consists of a contracting encoder and an expanding decoder linked by skip connections. The contracting encoder has a sequence of layers, where at regular intervals, the resolution of the feature maps is halved using max pooling, while the number of channels is doubled. Similarly, in the expanding decoder at regular intervals, the resolution of the feature maps is doubled using transposed convolution, and the number of channels is halved.

4.4.2 End-to-End Approach

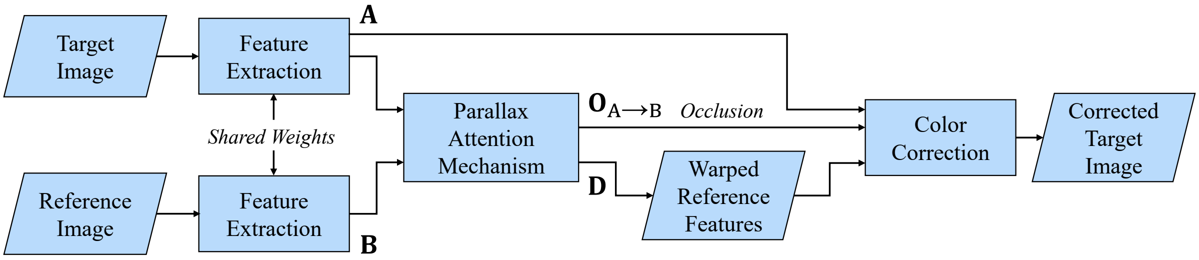


Figure 4.6: End-to-end approach (E2ENet).

The second approach is a local color correction solution able to correct local CM, inspired by [177], and illustrated in Figure 4.6. It consists of an end-to-end network called E2ENet that takes as input the target and reference image without any preprocessings and computes the corrected target image. Besides color correction, E2ENet also learns to estimate correspondences in the presence of CM and uses them in order to warp the feature map of the reference image into the target image. Therefore, E2ENet can be categorized as a multi-task network.

The proposed network has three main components: feature extraction, parallax-attention mechanism (PAM) [178], and color correction. The feature extraction component extracts the feature maps **A** and **B** from the target and reference image, respectively, which are necessary for the color correction. The second component, *i.e.*, PAM [178], estimates correspondences along horizontal epipolar lines assuming that the input S3D image is rectified. If the S3D image is not rectified, different image rectification methods could be applied [179, 180]. PAM also computes an occlusion map $\mathbf{O}_{\mathbf{A}\rightarrow\mathbf{B}}$, and it warps the feature map **B** of the reference image into the target image obtaining the warped feature map **D**. The last component takes as input the feature map **A** of the target image, the warped feature map **D** of the reference image together with the occlusion map $\mathbf{O}_{\mathbf{A}\rightarrow\mathbf{B}}$, and it computes the color corrected target image.

Here, two variants of E2ENet are considered. The first variant implements the feature extraction and the color correction as a sequence of residual blocks. This variant is called

ResBSeq-E2ENet and it was published in [57]. The second variant is similar to U-Net and uses a contracting encoder for the feature extraction and an expanding decoder for the color correction, and it is called EncDec-E2ENet.

Next, PAM is described in detail followed by the presentation of ResBSeq-E2ENet and EncDec-E2ENet.

Parallax-Attention Mechanism

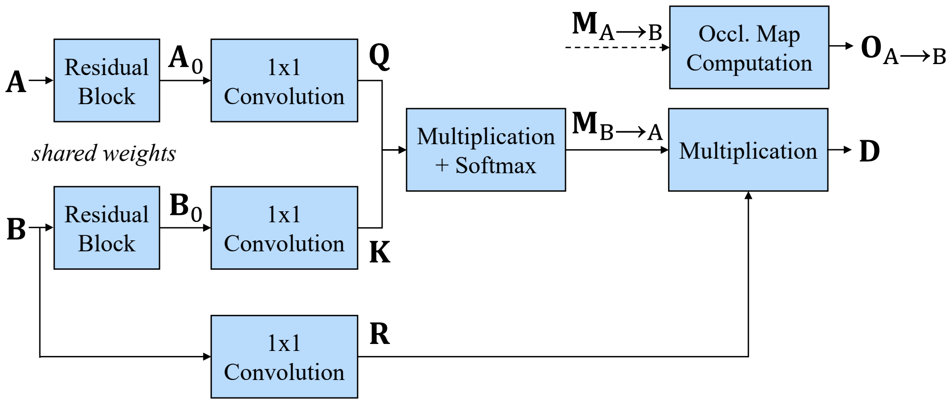


Figure 4.7: Parallax-attention mechanism (PAM).

The parallax-attention mechanism (PAM) [178] estimates correspondences by considering all the pixels along the epipolar lines, and for this reason, it is an example of a non-local network. Correspondence estimation is a long-studied problem in photogrammetry [121, 122], and PAM represents one of the latest solutions.

PAM has already been successfully used in other solutions for disparity estimation [178], S3D image super-resolution [178, 177, 181, 182, 183], binocular image dehazing [184], light field reconstruction [185], and object pose estimation [186]. Differently from the other solutions, we apply PAM to S3D images with CM and we show that it works also in this condition.

PAM is illustrated in Figure 4.7. The inputs of PAM are the feature maps $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{H\times W\times C}$ (H is the height, W is the width, and C are the channels) extracted in the previous component from the target image I_T and the reference image I_R, respectively. In the beginning, there are two residual blocks with shared weights that adapt the input features for the estimation of the correspondences and that generate the feature maps $\mathbf{A}_0$ and $\mathbf{B}_0$. This is important since different tasks require different features, otherwise, the proposed multi-task solution would suffer from training conflicts [187]. Then, a 1×1 convolution layer converts $\mathbf{A}_0$ into a feature map $\mathbf{Q} \in \mathbb{R}^{H\times W\times C}$, and another 1×1 convolution layer converts $\mathbf{B}_0$ into a feature map $\mathbf{K} \in \mathbb{R}^{H\times W\times C}$ that is reshaped to $\mathbb{R}^{H\times C\times W}$. $\mathbf{Q}$ and $\mathbf{K}$ are multiplied and softmax is applied obtaining a parallax attention map $\mathbf{M_{B \to A}} \in \mathbb{R}^{H \times W \times W}$. $\mathbf{M_{B \to A}}$ can be seen as a cost matrix that encodes the correspondences along horizontal epipolar lines. In the next step, $\mathbf{B}$ is processed by a 1×1 convolution layer obtaining $\mathbf{R} \in \mathbb{R}^{H \times W \times C}$, which is multiplied by $\mathbf{M_{B \to A}}$ to generate $\mathbf{D} \in \mathbb{R}^{H \times W \times C}$. $\mathbf{D}$ can be interpreted as the result of the warping of $\mathbf{B}$ into $\mathbf{A}$. PAM also estimates the occlusion map $\mathbf{O_{A \to B}}$. For the occlusion map, a second parallax attention map $\mathbf{M_{A \to B}}$ is estimated by exchanging $\mathbf{A}$ and $\mathbf{B}$. Refer to [178] for the details of the occlusion map computation.

In our case, PAM is trained in an unsupervised way without ground truth correspondences like in [178]. For this component, a correspondence estimation loss $\mathcal{L}_{PAM}$ is minimized. It consists of the sum of three losses: the photometric loss $\mathcal{L}_{pm}$, the smoothness loss $\mathcal{L}_{smooth}$, and the cycle loss $\mathcal{L}_{cycle}$:

$$\mathcal{L}_{PAM} = \mathcal{L}_{pm} + \mathcal{L}_{smooth} + \mathcal{L}_{cycle}. \tag{4.4}$$

The photometric loss $\mathcal{L}_{pm}$ warps the reference image I_R into the target image I_T based on $\mathbf{M_{B \to A}}$, and it computes the difference between the warped reference image and the target image. In addition, it also warps the target image based on $\mathbf{M_{A \to B}}$ and computes the difference between the warped target image and the reference image. More precisely, it is defined as follows:

$$\begin{aligned}\mathcal{L}_{pm} = &\sum_{p \notin \mathbf{O_{A \to B}}} \|I_T(p) - (\mathbf{M_{B \to A}} \otimes I_R)(p)\|_1 + \\ &\sum_{p \notin \mathbf{O_{B \to A}}} \|I_R(p) - (\mathbf{M_{A \to B}} \otimes I_T)(p)\|_1,\end{aligned} \tag{4.5}$$

where $\otimes$ denotes the matrix multiplication.

The smoothness loss $\mathcal{L}_{smooth}$ is applied to the parallax attention maps $\mathbf{M} \in \{\mathbf{M_{A \to B}}, \mathbf{M_{B \to A}}\}$ for the correct handling of textureless regions as follows:

$$\begin{aligned}\mathcal{L}_{smooth} = \sum_{\mathbf{M}} \sum_{i,j,k} (&\|\mathbf{M}(i,j,k) - \mathbf{M}(i+1,j,k)\|_1 + \\ &\|\mathbf{M}(i,j,k) - \mathbf{M}(i,j+1,k+1)\|_1).\end{aligned} \tag{4.6}$$

The cycle loss $\mathcal{L}_{cycle}$ is introduced to achieve cycle consistency according to the following equation:

$$\begin{aligned}\mathcal{L}_{cycle} = &\sum_{p \notin \mathbf{O_{A \to B}}} \|(\mathbf{M_{A \to B}} \otimes \mathbf{M_{B \to A}})(p) - \mathbf{J}(p)\|_1 + \\ &\sum_{p \notin \mathbf{O_{B \to A}}} \|(\mathbf{M_{B \to A}} \otimes \mathbf{M_{A \to B}})(p) - \mathbf{J}(p)\|_1,\end{aligned} \tag{4.7}$$

where $\mathbf{J} \in \mathbb{R}^{HxWxW}$ is a stack of H identity matrices.

ResBSeq-E2ENet

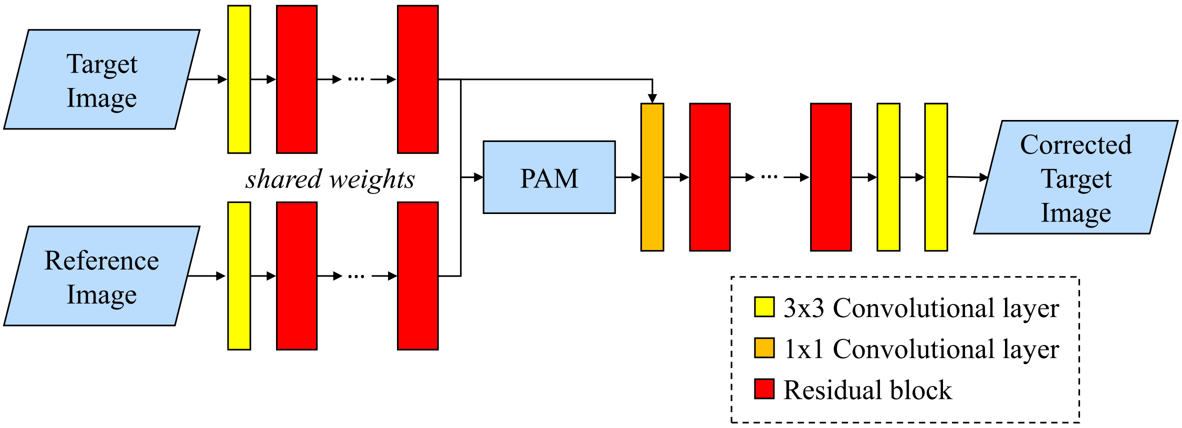


Figure 4.8: ResBSeq-E2ENet.

ResBSeq-E2ENet, which was inspired by [177], is illustrated in Figure 4.8. The idea behind this CNN is to use sequences of residual blocks like in ResNet [169]. In ResBSeq-E2ENet, the feature extraction component consists of a 3×3 convolution layer followed by a sequence of residual blocks. As explained before, the color correction component takes as input the feature map $\mathbf{A}$ extracted from the target image in the feature extraction component, and the outputs of PAM, *i.e.*, the warped feature map $\mathbf{D}$ of the reference image together with the occlusion map $\mathbf{O}_{\mathbf{A}\rightarrow\mathbf{B}}$. First, $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{O}_{\mathbf{A}\rightarrow\mathbf{B}}$ are concatenated and the resulting features are fused by a 1×1 convolution layer. The fused features are then processed by a sequence of residual blocks followed by two 3×3 convolution layers. The number of channels of the feature maps is kept constant through the entire CNN.

EncDec-E2ENet

EncDec-E2ENet is illustrated in Figure 4.9. In this solution, PAM is applied to downscaled feature maps of the input images. In this way, the correspondence estimation by PAM is faster and requires less memory.

For the feature extraction component, a contracting encoder similar to the one of U-Net is used, where at regular intervals, the resolution of the feature maps is halved and the number of channels is doubled. Different from U-Net, the encoder consists of a sequence of blocks containing a convolutional layer followed by four residual blocks. In the contracting encoder, the downscaling of the feature maps is computed by a convolution with stride two.

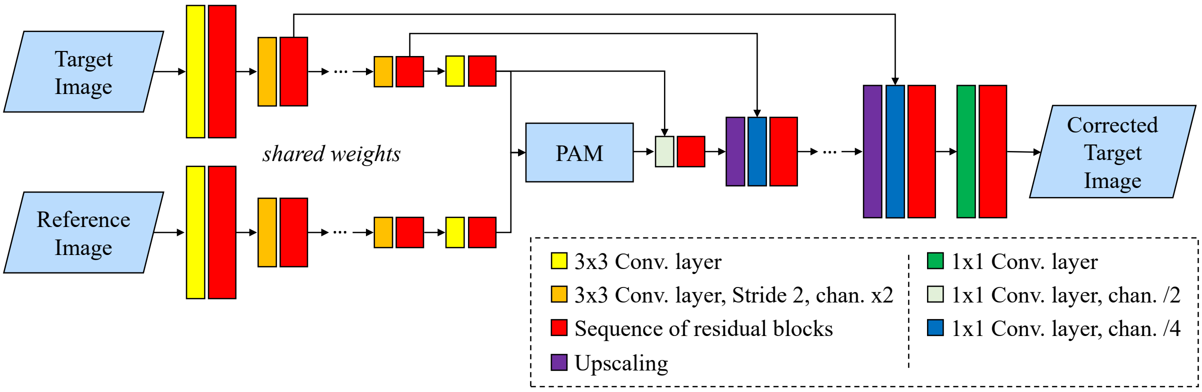

Target Image
Reference Image
shared weights
PAM
Corrected Target Image
3x3 Conv. layer
3x3 Conv. layer, Stride 2, chan. x2
Sequence of residual blocks
Upscaling
1x1 Conv. layer
1x1 Conv. layer, chan. /2
1x1 Conv. layer, chan. /4

Figure 4.9: EncDec-E2ENet.

The color correction component consists of an expanding decoder similar to the one of U-Net, where at regular intervals, the resolution is doubled and the number of channels is halved. Like the encoder, the decoder consists of a sequence of blocks with a convolutional layer followed by four residual blocks. Different from U-Net, in the expanding decoder, the upscaling of the feature maps is computed by bilinear interpolation.

Between the encoder of the target image and the decoder, there are skip connections. These are used in order to better transfer the information of the target image to the color correction component and also avoid the vanishing gradient problem. No skip connections from the encoder of the reference image are used since the reference image is not aligned with the corrected target image.

4.4.3 Losses

The color correction loss $\mathcal{L}_{CC}$ used to train PreProcNet and E2ENet evaluates how different the color corrected target image I'_T is from the ground truth target image I_T^{GT}. Like in [148], for $\mathcal{L}_{CC}$ we use a combination of pixel-based and perceptual losses. More precisely, we use the sum of the mean absolute error (MAE), the mean squared error (MSE), and the negative of the structural similarity index measure (SSIM) [188] as follows:

$$\mathcal{L}_{CC} = \frac{1}{N}\sum_{p} \|I'_T(p) - I_T^{\mathrm{GT}}(p)\|_1 + \frac{1}{N}\sum_{p} \|I'_T(p) - I_T^{\mathrm{GT}}(p)\|_2^2 - \mathrm{SSIM}(I'_T, I_T^{\mathrm{GT}}), \qquad (4.8)$$

where p are the pixels, and N is their number. MSE is commonly used because it makes the network converge fast, but since it computes the quadratic error sum, it is sensitive to the regions with large differences and not so much to the regions with small differences. On the other hand, MAE computes the absolute error sum and it is sensitive to the regions ignored by MSE. In this way, the sum of MSE and MAE improves the robustness of the trained model. SSIM was chosen since it is close to human visual perception.

The multi-task E2ENet also estimates correspondences in addition to color correction. As was mentioned in Section 4.4.2, this secondary task is learned in an unsupervised way by PAM based on the loss $\mathcal{L}_{PAM}$. The loss minimized by E2ENet is defined by the following weighted sum: $\mathcal{L}_{CC} + 0.005\,\mathcal{L}_{PAM}$

4.4.4 Experiments

Dataset

For the training and evaluation of the proposed CNNs, we took undistorted S3D images and introduced CM. The S3D images were taken from three datasets: Flickr1024 [189], InStereo2K [190], and the IVY LAB Stereoscopic 3D image database [191].

Flickr1024 [189] consists of 1024 S3D images with different resolutions collected from albums on the website *Flickr* [192], and are characterized by high quality and rich details covering diverse contents, like landscapes, urban scenes, people, man-made objects, and computer-generated scenes. It has already been used in other studies about S3D image super-resolution [177], denoising [193], stereo matching [194], etc. InStereo2K [190] contains 2050 S3D images with resolution 1080×860. They were taken from different indoor scenes including offices, classrooms, bedrooms, living rooms, and dormitories. The IVY LAB Stereoscopic 3D image database [191] has 120 S3D images with resolution 1920×1080 and captured using a 3D digital camera with dual lenses (Fujifilm FinePix 3D W3). The images were taken from 62 indoor scenes and 58 outdoor scenes, and they are characterized by various contents like humans, trees, structures, man-made objects, etc.

In order to exclude images with repetitive content and with a large CM, the S3D images were manually checked for repetition and analyzed automatically for CM with the method described in Section 3.6.1. In the end, we obtained 1035 undistorted S3D images.

In order to introduce CM in the undistorted S3D images, we applied the same color transformations used in [175, 148] but with more parameter values. Precisely, we modified the target images by applying six color modification operators found in Photoshop 2021 with different intensity levels that are reported in Table 4.1. In the end, we obtained 36225 distorted S3D images. The final dataset consists of a total of 37260 undistorted and distorted S3D images. Both types of S3D images were used for the training and evaluation. 80% of the dataset was randomly selected for the training set, 10% for the validation set, and another 10% for the test set.

Table 4.1: Photoshop 2021 operators and intensity levels used to generate the color mismatch dataset.

Photoshop Operator	Intensity Levels
Brightness	-90, -60, -30, 30, 60, 90
Color balance	-90, -60, -30, 30, 60, 90
Contrast	-60, -40, -20, 20, 40, 60
Exposure	-3, -2, -1, 1, 2, 3
Hue	-60, -40, -20, 20, 40, 60
Saturation	-40, -20, 20, 40, 60

Training Procedure

For the data augmentation, we applied random patch extraction and random vertical and horizontal flipping. Furthermore, we used the Adam optimizer [195] with a learning rate equal to 0.0001. The neural networks were implemented with Pytorch [196].

Color Correction Quality Metrics

For the evaluation of the color correction methods, four different full-reference quality metrics were applied between the color corrected and the ground truth target images. The first quality metric $\Delta\hat{E}^*_{ab}$ [197] is the mean of the color differences between corresponding pixels of a color corrected and the corresponding ground truth target image, where the color difference is defined as follows

$$\Delta E^*_{ab} = \sqrt{(L^*_1 - L^*_2)^2 + (a^*_1 - a^*_2)^2 + (b^*_1 - b^*_2)^2}, \tag{4.9}$$

with (L^*_1, a^*_1, b^*_1) and (L^*_2, a^*_2, b^*_2) representing colors defined in the CIELAB color space. $\Delta\hat{E}^*_{ab}$ was chosen because it is based on a perceptually uniform color space. The second quality metric is the structural similarity index measure (SSIM) [188], the third is the feature similarity index measure (FSIM) [198], and the last is the visual information fidelity (VIF) [199]. $\Delta\hat{E}^*_{ab}$ is an indicator for the correctness of the color information, while SSIM, FSIM, and VIF are more close to the human visual perception.

When evaluating the color corrected target images, left and right vertical borders likely not containing matching pixels with the corresponding reference images were not considered, since these borders are difficult to color correct and they are removed in the color correction method for S3D ODIs presented next in Section 4.5, where the methods for S3D standard images like the ones presented here are applied. Given the optical flow $(F_x(x,y), F_y(x,y))$ computed by SIFT-Flow [174] that maps (x,y) in a target image to $(x + F_x(x,y), y + F_y(x,y))$ in the corresponding reference image, and assuming that F_x^{min}

Table 4.2: Optimization of the depth of U-Net-PreProcNet.

	Depth		
	1	2	3
$\mathcal{L}_{CC}$	-0.9474	-0.9515	-0.9442

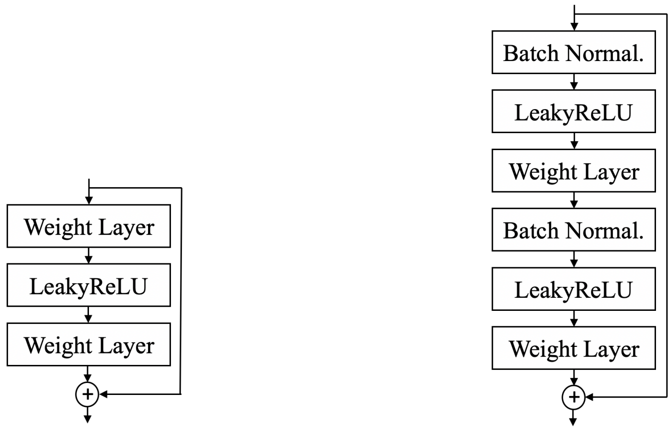
and F_x^{max} are the minimal and maximal x-axis displacements, in the color corrected target image a left vertical border with width equal to $\max(-F_x^{min}, 0)$ and a right vertical border with width equal to $\max(F_x^{max}, 0)$ were ignored.

Hyperparameter Optimization

In this section, the main hyperparameters are optimized. For this analysis, the best CM loss $\mathcal{L}_{CC}$ computed on the validation set in the first 50 epochs of the training is used.

For U-Net-PreProcNet, we set the number of channels in the first convolutional layer to 128, and we optimized the depth of U-Net, which is the number of times the resolution of the feature maps is halved in the encoder. According to Table 4.2, which shows the best loss $\mathcal{L}_{CC}$ of the validation set for different depths, the best performance is obtained with the depth equal to two.

For ResBSeq-E2ENet, we set the number of channels through the network to 64, and we optimized the type and the number of the residual blocks. Two different types of residual blocks were tested, ResBOne shown in Figure 4.10a, and ResBTwo with batch normalization shown in Figure 4.10b. Table 4.3 shows the best loss $\mathcal{L}_{CC}$ of the validation set for different network variants, and according to this table, the optimal model has the residual block type ResBOne, twelve residual blocks in the feature extraction component, and eight residual blocks in the color correction component.



(a) ResBOne

(b) ResBTwo

Figure 4.10: Residual blocks.

Table 4.3: Optimization of the type and number of the residual blocks in ResBSeq-E2ENet.

Feat. Extr.	Col. Corr.	$\mathcal{L}_{CC}$
4 ResBOne	4 ResBOne	-0.9624
4 ResBTwo	4 ResBTwo	-0.9498
8 ResBOne	4 ResBOne	-0.9645
4 ResBOne	8 ResBOne	-0.9659
8 ResBOne	8 ResBOne	-0.9666
12 ResBOne	8 ResBOne	-0.9690
8 ResBOne	12 ResBOne	-0.9689
12 ResBOne	12 ResBOne	-0.9680

Table 4.4: Optimization of EncDec-E2ENet.

Residual B.	Depth	I Layer Chan.	$\mathcal{L}_{CC}$
ResBOne	1	32	-0.9583
ResBTwo	1	32	-0.9226
ResBOne	1	64	-0.9707
ResBOne	1	128	-0.9402
ResBOne	2	32	-0.9650
ResBOne	2	64	-0.9592

For EncDec-E2ENet, we tested the residual blocks ResBOne and ResBTwo, different depths of the encoder (number of times when the feature map is downscaled to half resolution), and different numbers of channels in the first layer of the encoder. As can be seen from Table 4.4, which shows the best loss $\mathcal{L}_{CC}$ of the validation set for different network versions, ResBOne is better than ResBTwo in EncDec-E2ENet, and the best network has the depth equal to one and 64 channels in the first layer.

Ablation Study

For the ablation study, the different CNNs were trained with 100 epochs, and the color correction quality metrics presented before were applied to the corrected images of the test set.

For PreProcNet, we studied the influence of the globally corrected target image by replacing this input with the target image without any alteration. As can be seen in Table 4.5, which reports the metric values of PreProcNet with and without global correction, the globally corrected target image helps to improve $\Delta\hat{E}^{*}_{ab}$ a little, while the other metrics do not change too much. This can be explained by the fact that, in theory, the CNN extracts mainly structure information from the target image. Therefore, the global

Table 4.5: Performance of PreProcNet obtained using the identity function and Pitie *et al.* [3] for the global correction component .

Method	Glob. Corr.	$\Delta\hat{E}^*_{ab}$	SSIM	FSIM	VIF
SRCNN-PreProcNet	identity	4.9228	0.9543	0.9720	0.7304
SRCNN-PreProcNet	Pitie *et al.* [3]	4.7770	0.9527	0.9762	0.7339
U-Net-PreProcNet	identity	4.4693	0.9722	0.9870	0.8141
U-Net-PreProcNet	Pitie *et al.* [3]	3.9046	0.9725	0.9874	0.8085

Table 4.6: Study of the influence of the parallax attention mechanism.

Method	$\Delta\hat{E}^*_{ab}$	SSIM	FSIM	VIF
ResBSeq-E2ENet	4.4439	0.9710	0.9847	0.8283
ResBSeq-SIFT-Flow-E2ENet	5.3229	0.9657	0.9818	0.8114

color correction of the target image helps only marginally.

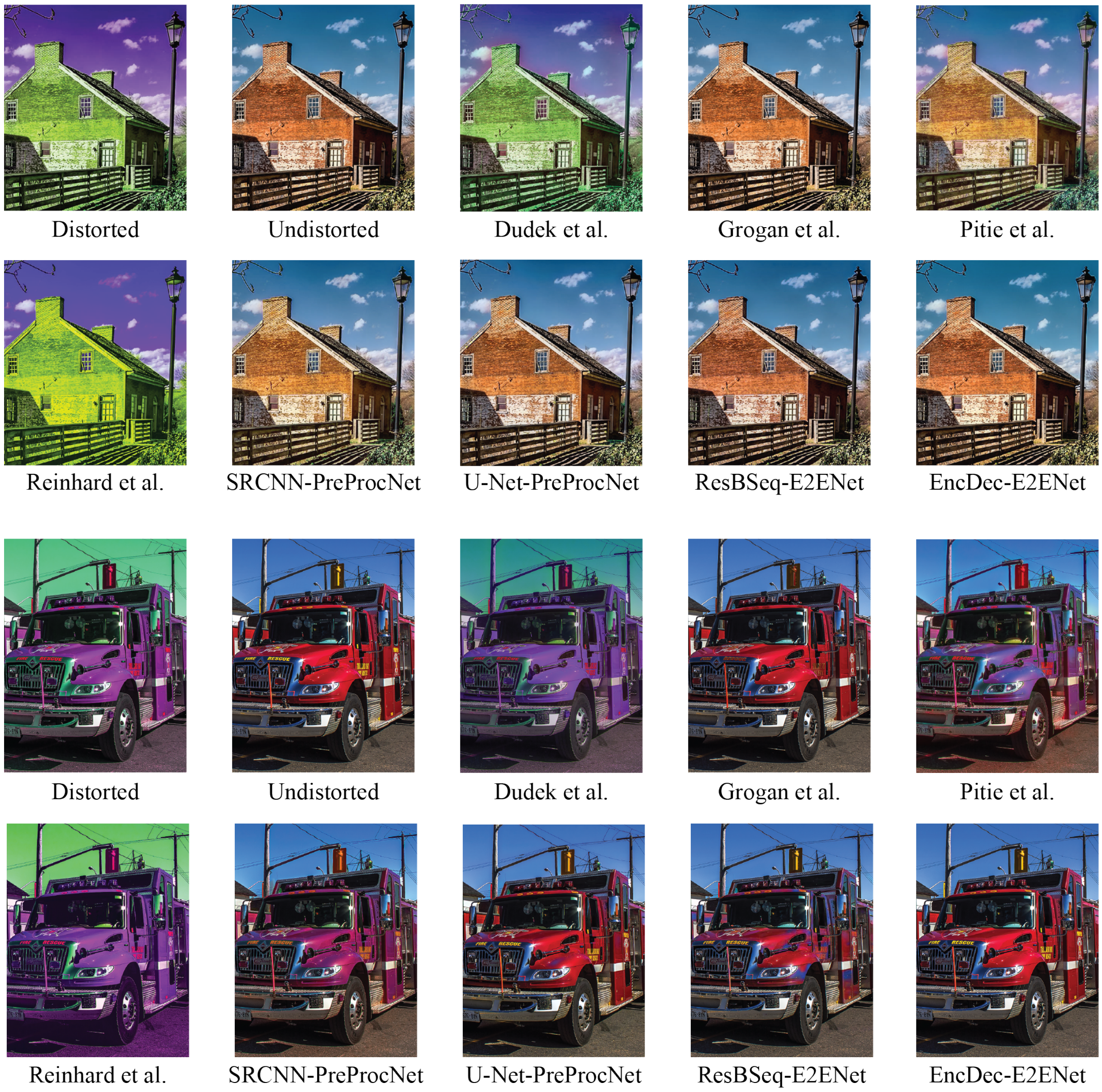
We also studied the contribution of PAM. For this study, we replaced PAM in ResBSeq-E2ENet with a component that warps the feature map **B** of the reference image based on correspondences obtained by SIFT-Flow [174]. We call this solution ResBSeq-SIFT-Flow-E2ENet. As can be seen from Table 4.6, which compares ResBSeq-E2ENet with ResBSeq-SIFT-Flow-E2ENet, PAM is important for improving the performance of ResBSeq-E2ENet.

Comparison

In this section, the optimized proposed solutions PreProcNet and E2ENet trained for 100 epochs are compared with four color correction methods based on the test set. These methods are Dudek *et al.*'s local approach [52], and the global approaches of Grogan *et al.* [147], Pitie *et al.* [3], and Reinhard *et al.* [142]. We did not compare against the most recent deep learning based method [148] as neither code nor data are available. The metric values are reported in Table 4.7. As can be seen in the table, U-Net-PreProcNet is better than SRCNN-PreProcNet. This is expected since SRCNN is less complex than U-Net. Between the two versions of E2ENet, EncDec-E2ENet reaches the best performance. And, also among all methods, EncDec-E2ENet is the best. Except for the proposed methods, the most competitive method is the one developed by Grogan *et al.*. As expected, Reinhard *et al.*'s method is the worst among all the evaluated methods due to the simplicity of this approach. Figure 4.11 shows the visual comparison of the methods. Also here, it is possible to observe that the best results are obtained by EncDec-E2ENet, U-Net-PreProcNet, and Grogan *et al.*'s method. Especially by looking at the figure with the truck, the quality of the color correction of the other methods is visibly worse.

Table 4.7: Comparison of the color correction methods.

Method	$\Delta\hat{E}^*_{ab}$	SSIM	FSIM	VIF
Dudek *et al.* [52]	6.4982	0.9366	0.9750	0.8258
Grogan *et al.* [147]	3.1523	0.9678	0.9822	0.7855
Pitie *et al.* [3]	5.8105	0.9350	0.9642	0.7038
Reinhard *et al.* [142]	13.4496	0.8247	0.9342	0.7384
SRCNN-PreProcNet	4.7770	0.9527	0.9762	0.7339
U-Net-PreProcNet	3.9046	0.9725	0.9874	0.8085
ResBSeq-E2ENet	4.4439	0.9710	0.9847	0.8283
EncDec-E2ENet	**2.7334**	**0.9841**	**0.9912**	**0.8799**



Distorted Undistorted Dudek et al. Grogan et al. Pitie et al.

Reinhard et al. SRCNN-PreProcNet U-Net-PreProcNet ResBSeq-E2ENet EncDec-E2ENet

Distorted Undistorted Dudek et al. Grogan et al. Pitie et al.

Reinhard et al. SRCNN-PreProcNet U-Net-PreProcNet ResBSeq-E2ENet EncDec-E2ENet

Figure 4.11: Visual comparison of the color correction methods.

Distortion-based Evaluation

In this section, the performance is studied separately for the different types of distortions that were applied in our dataset and described in Sec 4.4.4. Figure 4.12 and 4.13 show bar charts illustrating the distortion-based evaluation of EncDec-E2ENet and of Grogan *et al.* [147]. As can be noticed, for the images without CM (see undist in Figure 4.12 and 4.13) and for the images with a very low CM (cont+20 and cont-20), our approach is not able to reduce it. In reality, it slightly increases the CM. This behavior also characterizes Grogan *et al.* [147], and it can be explained by the inaccuracy of the correspondence estimation. Even in this evaluation, it is possible to notice the better performance of EncDec-E2ENet compared to Grogan *et al.* [147].

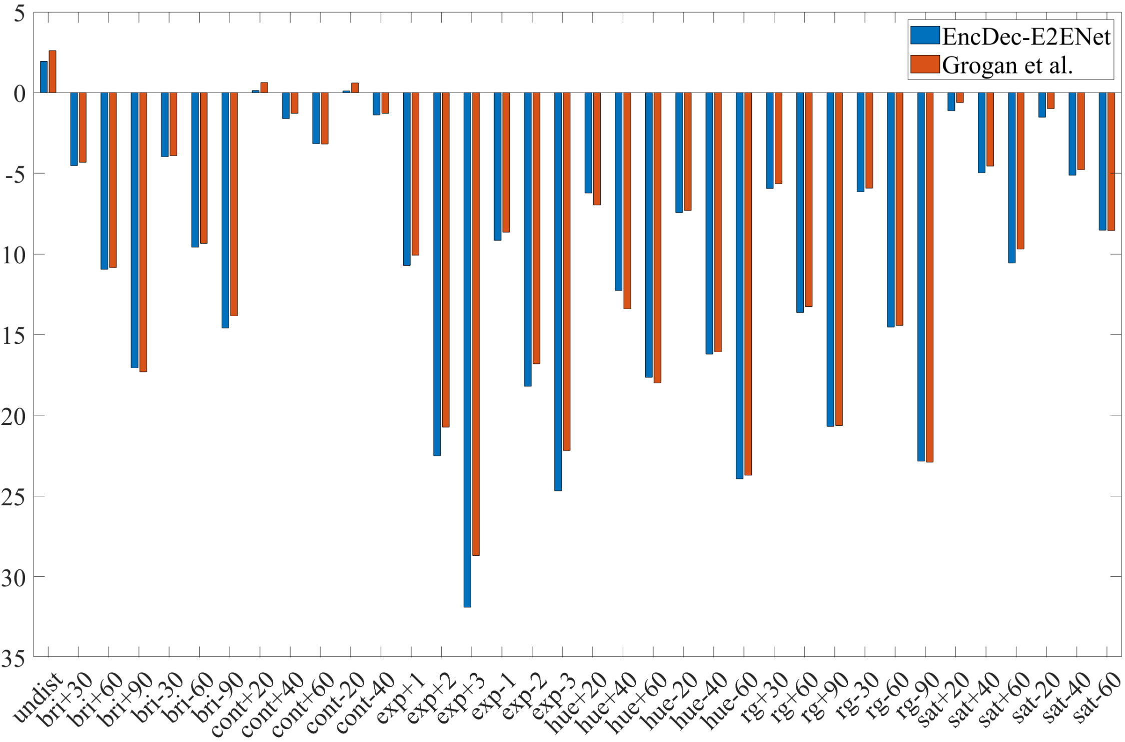


Figure 4.12: Difference between $\Delta\hat{E}^*_{ab}$ of the corrected and distorted target images for each distortion: $\Delta\hat{E}^*_{ab,\text{corrected}} - \Delta\hat{E}^*_{ab,\text{distorted}}$ (lower values are better). The labels of the horizontal axis refer to the different distortions with their intensity levels (undist corresponds to the undistorted images).

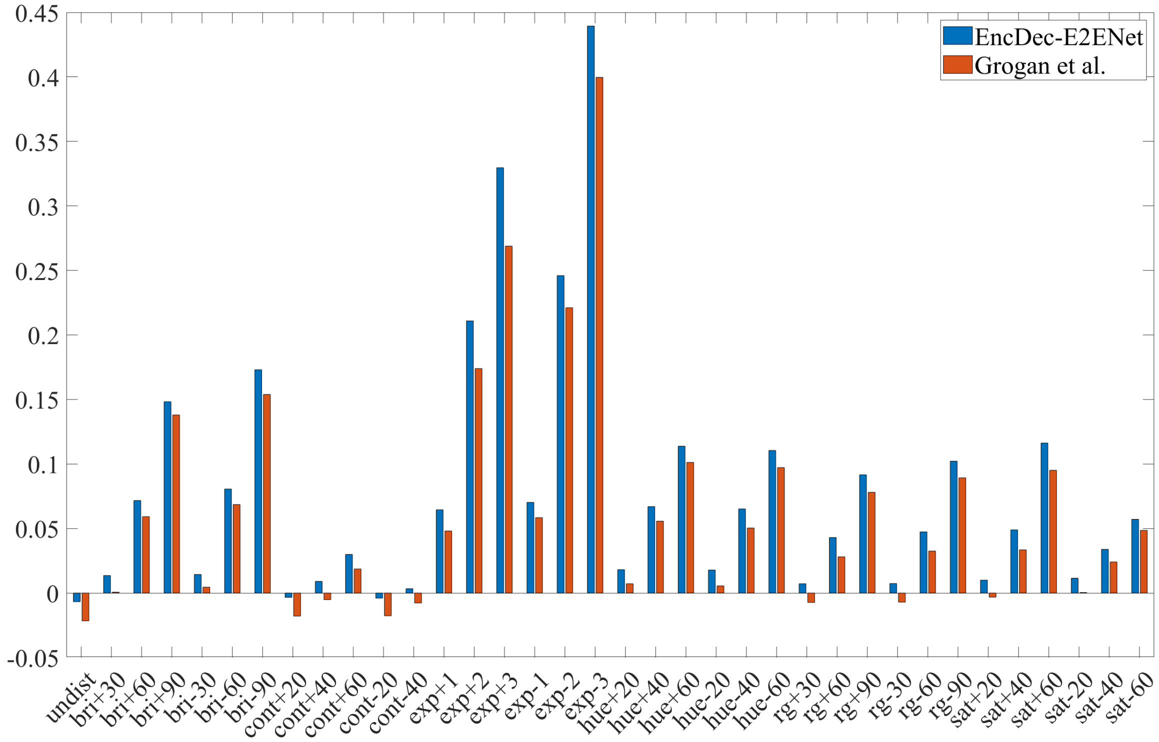


Figure 4.13: Difference between SSIM of the corrected and distorted target images for each distortion: $\text{SSIM}_{\text{corrected}} - \text{SSIM}_{\text{distorted}}$ (higher values are better). The labels of the horizontal axis refer to the different distortions with their intensity levels (undist corresponds to the undistorted images).

4.5 Extension to Stereoscopic 3D Omnidirectional Images

In general, the simplest approach to process an S3D ODI consists of taking as input the image in ERP format. For the proposed deep learning-based color correction solutions, this would require too much memory. Moreover, there could be problems at the left and right borders of the ERP format, since they would be processed as disconnected. For these reasons, we propose another approach based on patches similar to the one presented in Section 4.3. In this approach, planar patches are extracted from the ODI, and each patch belonging to the target image is color corrected. Afterward, a left and right vertical border are removed from each color corrected target patch, since they likely don't have matching pixels with the corresponding reference patch and they are difficult to color correct. In the end, the color corrected target patches without the removed borders are merged. We use planar patches evenly distributed on the spherical representation of the ODI according to the evenly distributed points used for the planar Voronoi patches of Voro-CMC (see

Section 2.2). Specifically, each planar patch is tangent on the sphere at one point of the evenly distributed points, and its size is specified by an user-defined horizontal and vertical field of view. The patches are extracted from the ODI based on the gnomonic projection [74], and their pixels are obtained by sampling the ODI in ERP format using bilinear interpolation. Once color corrected, the target patches without left and right vertical borders are projected back to the ERP format. In particular, the pixels of the projected patches in the ERP format are obtained by sampling the unprojected patches with bilinear interpolation. The projected patches without borders are then blended using linear combination. The patch pixel weights of the linear combination are obtained by applying the Gaussian function to the spherical distance between each patch pixel and the patch center on the sphere, like Voro-CMC.

We tested different patch parameters, and in the end, we decided to use 30 patches like Voro-CMC, with a vertical and horizontal field of view equal to 70 degrees, and a resolution 400×400. The left and right vertical borders of the color corrected target patches that are removed have a width equal to 80 pixels.

4.6 Comparison of the Approaches for Stereoscopic 3D Omnidirectional Images

This section compares the method Voro-CMC introduced in Section 4.3 and the best deep learning-based solution EncDec-E2ENet from Section 4.4 extended to S3D ODIs according to the approach described in Section 4.5. In order to evaluate the proposed methods, we selected 14 ODIs with the highest CM scores from the dataset introduced in Section 3.4, and one ODI that was captured with a 360° mirror-rig. For the method evaluation, we applied the CM detection method proposed in Section 3.6 assuming that the input visual attention map is uniform. It is worth mentioning, that while the color correction approaches work in the RGB color space, the CM detection method is applied in the Lab color space. This allows a more objective and independent evaluation of the still existing color discrepancies between the views.

Figure 4.14 shows a bar chart with the CM scores computed by our CM detection method before and after applying the proposed color correction methods, while Figure 4.15 shows the CM analysis before and after color correction of ODI 1 and ODI 2, together with some close-ups. Voro-CMC is able to reduce the CM scores in all ODIs by an average of 74%. The largest CM score reduction, equal to 89%, was observed for ODI 1. As can be noticed in Figure 4.15, Voro-CMC is able to drastically reduce the strong CM in ODI 1 and ODI 2 everywhere. Apart from the good results, we also observed some

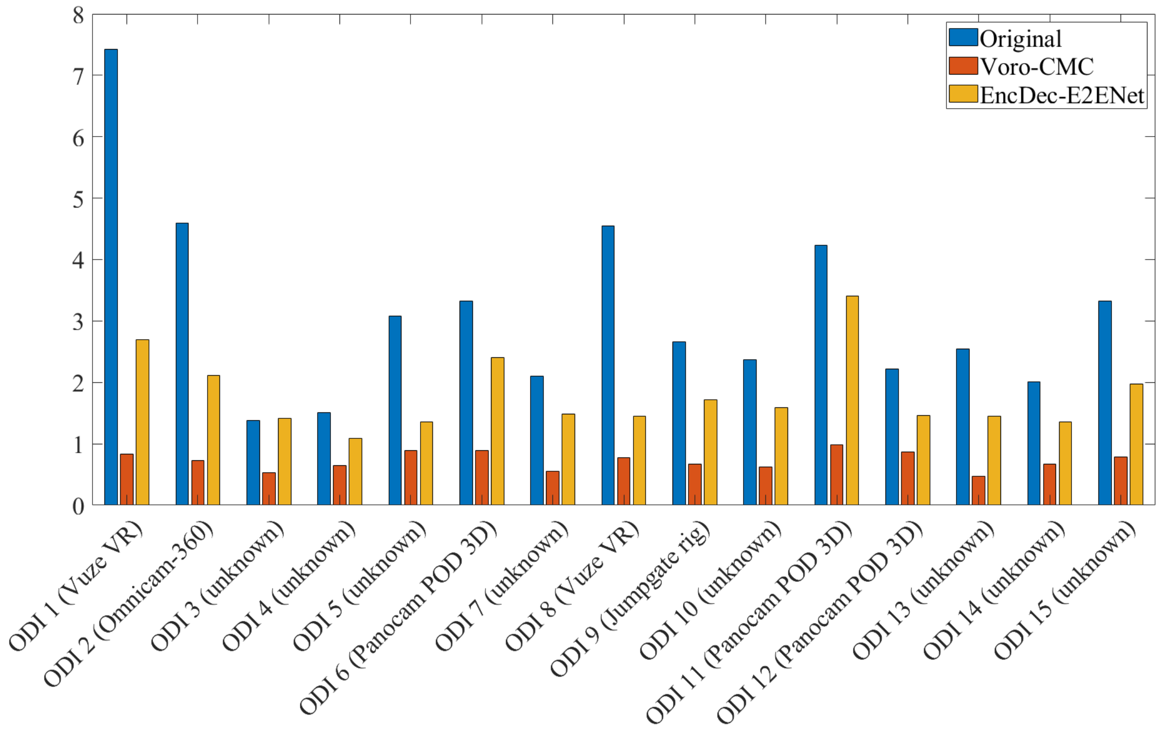


Figure 4.14: Bar chart with the performance of the CM correction methods for S3D ODIs based on a dataset with 15 ODIs (the used camera rigs are specified in brackets). The bar chart shows the CM scores computed by our CM detection method (Section 3.6) before and after color correction.

limitations of Voro-CMC. This method is based on a color transfer method that is robust to pixel correspondence inaccuracies, but these inaccuracies still influence the performance of Voro-CMC. Another limitation occurs when the patch is too large compared to the region with CM, or when the patch contains regions with different types of CM. In this case, the CM is reduced only partially.

In Figure 4.14, we can also notice that EncDec-E2ENet is able to reduce the CM scores by an average of 36%. While, in Figure 4.15, we can see that the corrected ODI 1 still has regions with large CM, and in ODI 2, moderate CM remains after the color correction. Compared to Voro-CMC, EncDec-E2ENet has worse performance as noticeable both in Figure 4.14 and in Figure 4.15. We have two main hypotheses about the reasons for this performance. The first is the presence of CM in the undistorted images of the training set mostly obtained with 3D cameras and not 2D-to-3D converted. Even if we selected images with low CM, they still have a little of it. The second hypothesis is the overfitting of the types of CM of the training dataset. On one hand, EncDec-E2ENet was trained with images characterized by global CM and not local CM, while the 15 ODIs contain local CM. On the other hand, the training dataset was obtained by applying a limited

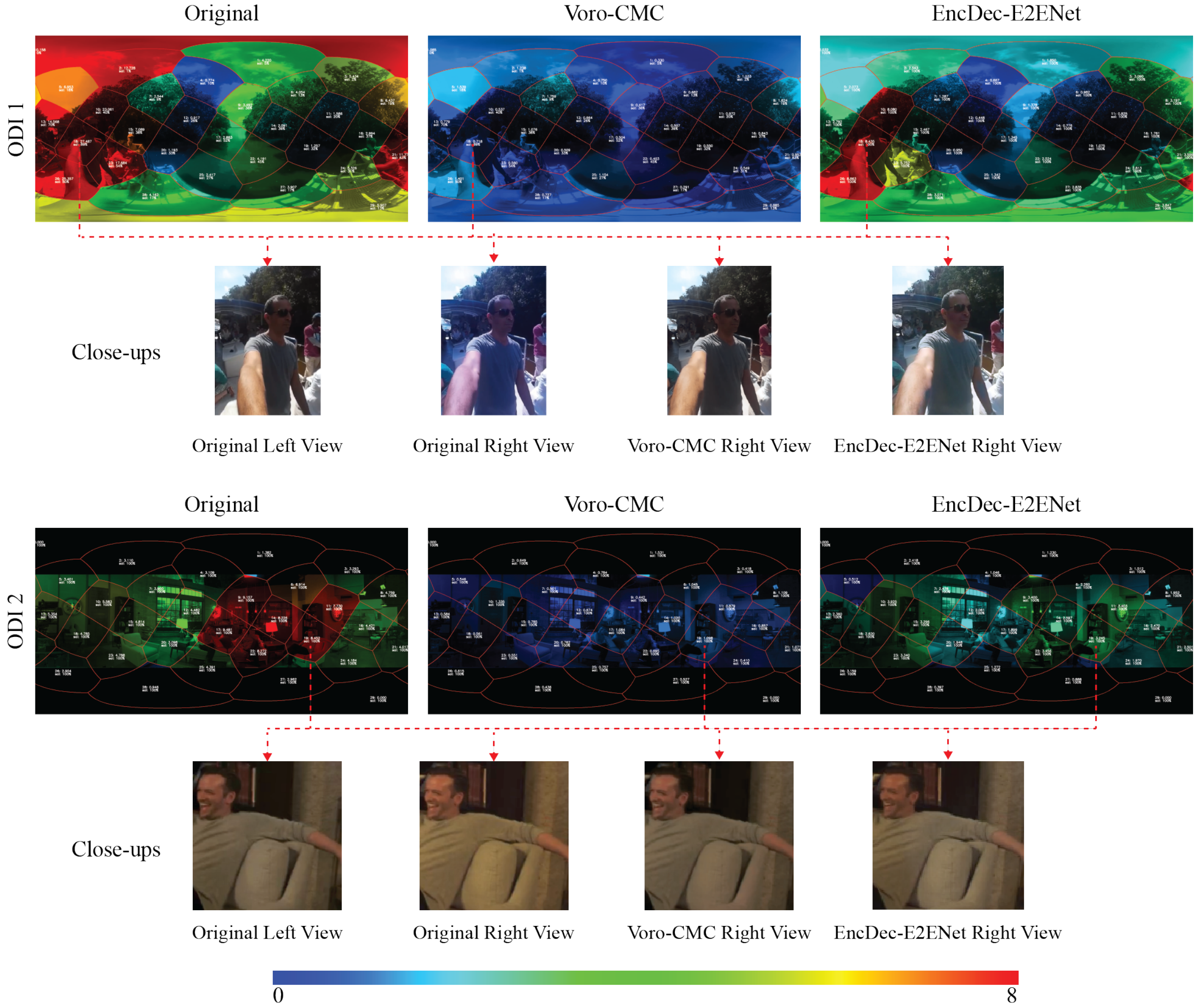


Figure 4.15: Sample ODIs with CM visualisation (red: strong mismatch, blue: no mismatch) based on the CM scores computed by our CM detection method (Section 3.6) and close-ups before and after color correction.

number (six) of Photoshop operators.

There are also some artifacts, like different glares in the two views or contamination (rain, dust, etc.), that can be identified as CM by our two methods but that cannot be completely removed. Problems can arise, for example, in the case of different glares in both views, or in the case of stitching artifacts within a patch. This can reduce the number of correct correspondences that are found in certain regions of the ODI, reducing the likelihood that they will be successfully corrected.

4.7 Conclusions

This chapter introduced Voro-CMC, which is a CM correction method for S3D ODIs based on traditional visual computing techniques. Moreover, it also presented two general deep learning-based methods for S3D standard images called PreProcNet and E2ENet, and their extension to S3D ODIs. According to the evaluation of the methods, Voro-CMC has a better performance than the deep learning-based methods.

Even if Voro-CMC reaches the best performance, it can still be improved by tackling some of its limitations, especially the misalignment of the patch with the region affected by CM. The possibility to have adaptable patches to the region with CM could be investigated.

Regarding the deep learning-based solutions, as previously discussed, there are two main hypotheses about the reasons for their lower performance: the presence of some low level of CM in the undistorted images of the training dataset, and the overfitting of the types of CM that characterize the training dataset. In the future, we plan to improve the performance of the deep learning-based approaches by creating a new training dataset. For this new dataset, we could take 2D-to-3D converted or rendered S3D images, since they do not have CM. In addition, more color transformations than the six Photoshop operators used in our old dataset could be applied. Finally, local CM could be introduced in the training dataset. This could be created by alpha blending globally transformed images with an alpha map containing zero and one regions with a small transition area between them, like in Figure 4.16.

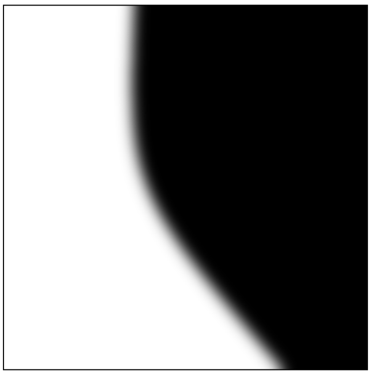


Figure 4.16: Alpha map.

Chapter 5

Voronoi-based Quality Metrics for Monoscopic Omnidirectional Video

5.1 Introduction

Compared to traditional video, ODV introduces new technical challenges especially for storage and transmission [20]. For the development and evaluation of new solutions to these technical challenges subjective and especially objective quality assessment methods are necessary. Currently, there are already quality metrics designed for monoscopic ODV like [34, 35, 36, 37, 38, 39], but these metrics have a limited correlation with the subjective quality scores. Thus, in order to improve the quality estimation performance, in this chapter, we propose a new objective quality assessment framework for monoscopic ODV. This framework was published in [54, 55, 56] and it was accepted as recommended method for monoscopic ODV quality assessment by MPEG [200]. Since the framework deals with monoscopic ODV, in the rest of the chapter, the term ODV refers to monoscopic ODV.

Quality assessment for ODV requires to consider its unique aspects, already mentioned in Chapter 1. First, ODV is inherently a spherical signal, but it is stored and transmitted in planar formats to be compatible with the existing video delivery pipelines. The spherical projection techniques presented in Section 2.1.1 could be used for the conversion into the planar formats, but they inevitably introduce distortions that must be taken into account to accurately estimate the ODV quality [201]. Second, HMDs allow the viewer to freely look around a scene [46], but they show only a part of the ODV, *i.e.*, the viewport. Therefore, for ODV it is important to consider visual attention. Various previous research works emphasize the importance of visual attention in quality assessment [201, 202], and existing studies show that visual attention improves the performance of quality assessment [38, 39, 37, 203, 204].

In this chapter, we propose an objective full-reference quality assessment framework for ODV that takes into account the spherical nature of ODV and its viewing characteristics. The framework first subdivides the ODV into planar Voronoi patches (Section 2.2) with low projection distortions. Afterward, the framework applies a quality metric for traditional video to each planar Voronoi patch, obtaining a quality score for each patch. To further consider the viewing characteristics of ODV, the proposed framework integrates visual attention by computing a weight for each patch that accounts for the probability of the patch being viewed. Finally, the framework computes a weighted average of the patch scores based on the visual attention weights obtaining the final ODV quality score. The results show that both the ODV subdivision into planar Voronoi patches and the integration of visual attention improve the performance of ODV quality assessment and are crucial for achieving state-of-the-art performance.

To evaluate the proposed framework, we created a dataset of ODVs with scaling and compression distortions, and we conducted subjective experiments in order to gather the subjective quality scores and the visual attention data for our ODV dataset. The evaluation consists in the analysis of the framework components, such as the number and angular resolution of the planar Voronoi patches, the visual attention estimation method, and the temporal pooling of the frame scores. We also performed a comparative analysis with existing quality metrics.

The rest of this chapter is organized as follows. Section 5.2 discusses the related work on both subjective and objective ODV quality assessment. Then, Section 5.3 describes the proposed quality assessment framework. The details of our ODV dataset and the related subjective experiments are explained in Section 5.4. Based on the proposed dataset, Section 5.5 presents the study of the framework components and the extensive comparative analysis with several existing quality metrics.

5.2 Background

Although there are many studies about subjective and objective ODV quality assessment, in the following, we outline only those that are most related to our work together with their limitations. For a comprehensive overview of recent research in the field, we recommend the overview paper of Li *et al.* [205].

5.2.1 Subjective Quality Assessment

Creating datasets and gathering subjective quality scores are fundamental requirements to understand the perceived quality of distorted omnidirectional images [202] and videos

[37, 206, 207, 208]. For this purpose, Li *et al.* [37] conducted a subjective experiment to establish an ODV quality dataset. Their dataset contains subjective scores for 600 compressed ODVs across 221 participants. Eye and head movement data were also gathered during the subjective experiment. Another recent work [206] established a dataset that contains subjective quality scores of 30 participants across 50 different ODVs compressed with the HEVC/H.265 video coding standard [69]. In this work, the optimal resolution of ODVs displayed by the HMD was used in order to reduce the sampling distortions when extracting the viewport from the ODV. Furthermore, Singla *et al.* [207] and Schatz *et al.* [208] conducted subjective experiments to assess the perceived quality of ODV streaming.

Various existing studies related to quality assessment, *e.g.*, [206, 209, 201, 210, 211], consider only compression distortions of ODVs with low spatial resolution due to the computational complexity of ODV rendering. However, hardware for the rendering of 8K ODV is now on the market, providing a higher quality of experience. In our research, we created an ODV dataset, which is based on the typical visual distortions in adaptive streaming systems, namely, compression and scaling distortions, applied to uncompressed ODVs with 8K resolution. We also organized subjective experiments to collect the subjective scores together with the viewport trajectories for the ODVs.

5.2.2 Objective Quality Assessment

Many quality metrics developed for ODV are the extended versions of the traditional PSNR metric. Sun *et al.* [34], for instance, developed the weighted spherical PSNR metric (WS-PSNR) with weights that consider the projection distortions of the pixels in the planar format. The Craster parabolic projection PSNR metric (CPP-PSNR) [35] computes the PSNR in the Craster parabolic projection characterized by low projection distortions. Furthermore, the spherical PSNR metric (S-PSNR) [36] estimates the PSNR for uniformly sampled points on the sphere. This quality metric has two different variants, namely, S-PSNR-NN and S-PSNR-I. When sampling pixels, they use the nearest neighbor or bicubic interpolation, respectively.

Subjective quality studies reported various findings on the PSNR-based quality metrics for ODV. On one hand, Zhang *et al.* [206] and Sun *et al.* [201] recently reported that the existing PSNR-based quality metrics for ODV have superior performance than the traditional PSNR. On the other hand, Tran *et al.* [210] claimed that the traditional PSNR is the most appropriate metric for quality evaluation in ODV communication. Furthermore, Upenik *et al.* [96] showed that the existing PSNR-based quality metrics for ODV do not have a high correlation with subjective scores. A similar conclusion was reached in another study [209].

In addition to the PSNR-based metrics, the structure similarity index metric (SSIM) was also extended to ODV by Chen *et al.* [212] based on weights that take into account the projection distortions. Moreover, a recent study [211] investigated the performance of the video multimethod assessment fusion metric (VMAF) [213] applied to ODV, which is a metric for traditional video developed to evaluate the distortions introduced by the adaptive streaming systems (*i.e.*, compression and scaling distortions) and characterized by high correlation with subjective scores [214, 215, 216]. Specifically, the study in [211] created a dataset of ODVs in ERP compressed using constant quantization parameters, and showed that VMAF can be used as a metric also for ODVs without modifications. In our research, we showed based on an ODV dataset with compression and scaling distortions, that the performance of VMAF can be improved using planar Voronoi patches. We did not only study VMAF, but we developed a new objective quality assessment framework for ODV based on planar Voronoi patches. With our framework, existing quality metrics for traditional video (*e.g.*, VMAF) can be applied to ODV based on planar Voronoi patches achieving a high correlation with subjective scores.

5.2.3 Visual Attention in Objective Quality Assessment

As already shown in [201, 202], visual attention is crucial when evaluating the quality of ODV. Similarly, Li *et al.* [37] showed that the incorporation of head and eye movement data in objective quality assessment, more specifically in PSNR, increases the prediction performance. Upenik *et al.* [38] also proposed to incorporate visual attention in PSNR for ODV quality assessment. Furthermore, Ozcinar *et al.* [39] developed a quality metric based on PSNR that considers visual attention and projection distortions, with the aim of ODV streaming optimization. However, these works [37, 38, 39] that use visual attention are based on PSNR, which does not correlate well with subjective scores. Differently, in our research, we developed a new quality assessment framework, which works with visual attention and robust quality metrics for traditional video.

5.3 Proposed Framework

This section introduces the proposed framework for objective full-reference quality assessment based on planar Voronoi patches (Section 2.2) and visual attention, which is illustrated in Figure 5.1. Initially, we introduce the Voronoi-based framework without visual attention, and then the Voronoi-based framework integrated with visual attention.

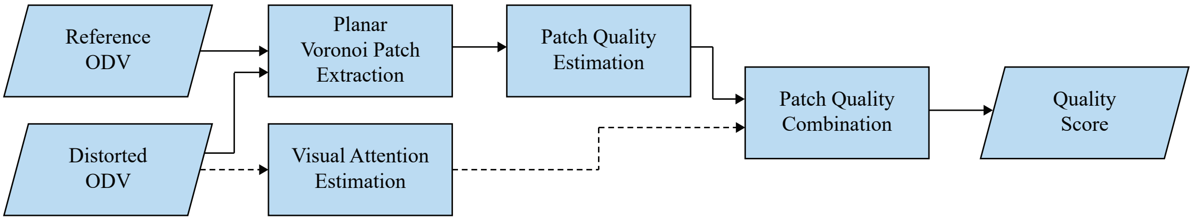

Reference ODV
Distorted ODV
Planar Voronoi Patch Extraction
Visual Attention Estimation
Patch Quality Estimation
Patch Quality Combination
Quality Score

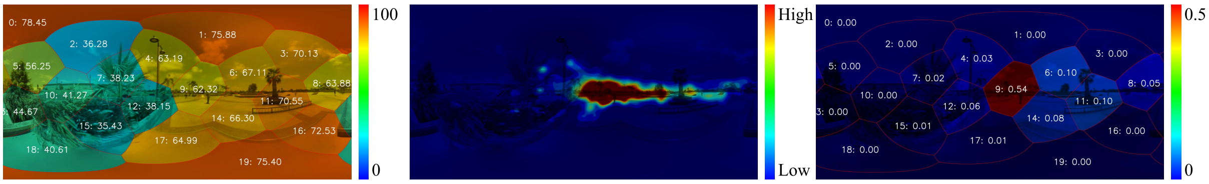
Figure 5.1: Voronoi-based framework with visual attention.

5.3.1 Voronoi-based Framework

The quality framework presented in this section extends full-reference metrics for traditional video to ODV. The extended metrics for ODV are called VI-METRIC, where VI stands for Voronoi, and METRIC $\in \{\text{PSNR}, \text{SSIM}, \text{MS-SSIM}, \text{VMAF}, \ldots\}$ is a full-reference metric for traditional video. Since we are dealing with full-reference quality assessment, the inputs of the framework are a distorted (*e.g.*, compressed) ODV and the corresponding undistorted reference ODV. Initially, the quality framework extracts M planar Voronoi patches Π'_k ($k = 1, \ldots, M$) from the distorted ODV and other M from the reference ODV according to the method presented in Section 2.2. Then, a full-reference metric for traditional video is applied to the planar Voronoi patches Π'_k of the distorted and reference ODV, obtaining M patch scores Γ_k. In our study, we apply the following full-reference metrics: PSNR, SSIM [188], MS-SSIM [217], and VMAF [213]. Since these metrics take rectangular video frames as input, we modified the first three of them, so that they can deal with any patch shape. While for VMAF, we take the bounding box of the patch as input, as it is not straightforward to modify VMAF for different patch shapes. Specifically, for PSNR we compute the mean squared error at the basis of this metric only inside the patches. SSIM computes different comparison measurements at each pixel, namely, luminance, contrast, and structure. We adapted SSIM to the planar Voronoi patches by considering only the comparison measurements inside the patches. Similar to SSIM, MS-SSIM computes comparison measurements for each pixel at different resolutions of the input images. MS-SSIM was adapted to the planar Voronoi patches by resizing these patches and by considering only the comparison measurements inside the resized patches. In the end, the final ODV quality score is obtained by computing the arithmetic mean of the patch scores Γ_k as follows:

$$\text{VI-METRIC} = \frac{\sum_{k=1}^{M} \Gamma_k}{M}. \tag{5.1}$$

5.3.2 Integration of Visual Attention



(a) VMAF scores of 20 Voronoi patches.

(b) Visual attention map generated with the Kent distribution method [45].

(c) Visual attention patch weights $\nu_{i,k}$ of 20 Voronoi patches from the visual attention map in (b).

Figure 5.2: Visualization of the VMAF patch scores, visual attention map, and the visual attention patch weights $\nu_{i,k}$. Please refer to the color bars beside the figures for the used color code.

As already mentioned in Section 5.1, since the viewers tend to look at the regions that attract their visual attention, and since different parts of the ODV can have different quality, it is important to consider visual attention during quality assessment and give more weight to the regions that are most likely viewed.

We now propose to integrate visual attention into the original Voronoi-based framework and refer to its metrics as VI-VA-METRIC, where VA stands for visual attention. Different methods can be used for the computation of visual attention maps. We investigate the effects of different visual attention estimation methods in Section 5.5.1. Figure 5.2 shows a sample visual attention map generated using the Kent distribution method [45].

For the computation of the VI-VA-METRICs, first a quality score for each video frame of the distorted ODV is computed based on visual attention, and then the frame scores are pooled into a final quality score. For the computation of the frame scores, initially M planar Voronoi patches Π'_k $(k = 1, \ldots, M)$ are extracted from each frame i of the distorted and reference ODV. Then, a full-reference metric for traditional video is applied to the planar Voronoi patches Π'_k of each frame i, obtaining M patch scores $\Gamma_{i,k}$ for each frame. At this point, the visual attention map Υ_i of each frame i of the distorted ODV is estimated. Then, M planar Voronoi patches Π'_k are extracted from each visual attention map Υ_i, and the sums $\nu_{i,k}$ of the visual attention pixel values inside each patch Π'_k of each map Υ_i are computed. The sum $\nu_{i,k}$ is related to the probability of patch Π'_k of frame i being viewed. Next, the frame scores T_i are obtained through a weighted average of the patch scores $\Gamma_{i,k}$ using the visual attention sums $\nu_{i,k}$ as weights according to the following equation:

$$T_i = \frac{\sum_{k=1}^{M} \nu_{i,k} \Gamma_{i,k}}{\sum_{k=1}^{M} \nu_{i,k}}. \tag{5.2}$$

In the last step, the frame scores T_i are combined using a temporal pooling approach P_{tempo} obtaining the final video score:

$$\text{VI-VA-METRIC} = P_{tempo}(T_1, T_2, \ldots, T_N), \tag{5.3}$$

where N is the number of frames. Different pooling approaches P_{tempo} can be applied, like the arithmetic and harmonic mean, the median, the minimum, etc. In this study, we analyze the following metrics obtained with the framework: VI-VA-PSNR, VI-VA-SSIM, VI-VA-MS-SSIM, and VI-VA-VMAF.

Figure 5.2 shows the patch scores obtained by applying VMAF to 20 Voronoi patches, the visual attention map computed by the Kent distribution method [45] from the viewport trajectories obtained in our subjective experiments, and the visual attention patch weights $\nu_{i,k}$ corresponding to 20 Voronoi patches. As can be seen in the figure, different regions of the ODV can have noticeably different qualities, and also clearly different visual attention values. For this reason, we integrate visual attention in our proposed quality assessment framework in a way to give more importance to patches that attract visual attention.

5.4 Dataset and Subjective Experiments

In this section, we introduce our dataset, and we describe the technical details of the two subjective experiments that we conducted in order to collect the subjective quality scores and the viewport trajectories for our dataset. This section terminates with the analysis of the collected subjective data.

5.4.1 Omnidirectional Video Quality Dataset

Considering a streaming application scenario, we built our dataset using ODVs with different spatial resolutions and different compression levels. For our dataset and subjective experiments, we first selected a total of nine *uncompressed* reference ODVs in YUV420p format of 10 *sec.* length, 8K×4K ERP resolution, and with different characteristics. These ODVs were selected from the videos of the Joint Video Exploration Team of ITU-T VCEG and ISO/IEC MPEG [218, 219, 220]. The selected videos are *Basketball*, *Dancing*, *Gaslamp*, *Harbor*, *JamSession*, *KiteFlite*, *SkateboardTrick*, *Train*, and *Trolley*. Sample frames of these videos are shown in Figure 5.3. *Basketball*, *Dancing*, *Harbor*, *JamSession*, *KiteFlite* were rated in the first subjective experiment, and *Gaslamp*, *SkateboardTrick*, *Trolley* were rated in the second experiment. The *Train* sequence was used

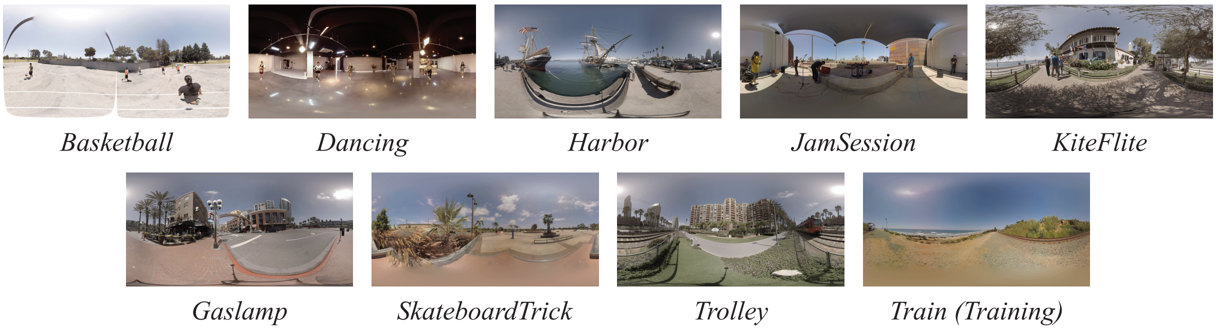


Figure 5.3: Sample frames of the nine reference ODVs used in the subjective experiments. The top five ODVs were rated in the first experiment, and the bottom left three ODVs were rated in the second experiment. *Train* was used for the experiment training.

only as training material in both experiments.

After the selection of the nine reference ODVs, they were downsampled to three different resolutions in ERP format: 8128×4064, 3600×1800, and 2032×1016. For the downsampling, we used the bicubic scaling algorithm of the FFmpeg software (*ver.* 4.0.3-1 18.04). Next, the ODVs were compressed with the HEVC/H.265 video coding standard [69]. For this, we used the *libx265* codec (*ver.* 2.9) [221] in FFmpeg [222] with the video buffering verifier method to set the target bitrates. As this database was created to consider possible cases that might be encountered in an adaptive streaming scenario, to ensure constant bitrate, each ODV was compressed using two-pass encoding with 150 percent constrained variable bitrate configuration, following the recommendations of streaming providers [223]. We also defined the buffer size during encoding to limit the output bitrate to twice the maximum bitrate for handling large bitrate spikes. To avoid any possible impact of the unknown resampling algorithm used by the video player, we upsampled the decoded ODVs to 8128×4064 resolution using the bicubic scaling algorithm of FFmpeg. For the downsampling and compression of the reference ODVs, we used the following FFmpeg commands:

```
ffmpeg -y -f rawvideo -pix_fmt iVideoFormat -s iVideoRes -r
    iVideoFramerate -i iVideoFn -c:v libx265 -preset medium -frames:v
     iVideoFrames -vf scale=oVideoRes -x265-params profile=main:
    keyint=48:min-keyint=48:scenecut=0:ref=5:bframes=3:b-adapt=2:
    bitrate=oVideoBitRate:vbv-maxrate=oVideoMaxRate:vbv-bufsize=
    oVideoBufSize:pass=1 -f mp4 /dev/null

ffmpeg -y -f rawvideo -pix_fmt iVideoFormat -s iVideoRes -r
    iVideoFramerate -i iVideoFn -c:v libx265 -preset medium -frames:v
     iVideoFrames -vf scale=oVideoRes -x265-params profile=main:
```

```
keyint=48:min-keyint=48:scenecut=0:ref=5:bframes=3:b-adapt=2:
bitrate=oVideoBitRate:vbv-maxrate=oVideoMaxRate:vbv-bufsize=
oVideoBufSize:pass=2  oVideoFn
```

where

- *iVideoFn*: filename of input video,
- *iVideoRes*: resolution of input video,
- *iVideoFormat*: format of input video (in our case yuv420p),
- *iVideoFramerate*: framerate of input video,
- *iVideoFrames*: number of frames of input video,
- *oVideoFn*: filename of output video,
- *oVideoRes*: resolution of output video,
- *oVideoBitRate*: target bitrate of output video in Kbps,
- *oVideoMaxRate*: maximum bitrate (in our case 1.5 × *oVideoBitRate*),
- *oVideoBufSize*: buffer size (in our case 2 × *oVideoMaxRate*).

To ensure that the distorted ODVs in our database are uniformly distributed across different quality levels, five different target bitrates were selected independently for each reference ODV in a pilot test with three experts using HTC Vive HMD. For this pilot test, before encoding, the reference ODVs were resized to the resolution 3600×1800, which was found to be the optimal ODV resolution for HTC Vive HMD by Zhang *et al.* [206], according to their calculation considering the HMD's display resolution and its field of view. The ODVs were then encoded with different bitrates $\in \{500, 1000, 2000, 5000, 7000, 10000, 13000, 15000\}$ Kbps, and among them five different bitrates were selected in the pilot test corresponding to five different quality levels, namely, *"bad"*, *"poor"*, *"fair"*, *"good"*, and *"excellent"*. The selected bitrates are reported in Table 5.1.

5.4.2 Subjective Experiments

This section describes the technical details of the two subjective experiments that we organized. Their main characteristics are shown in Table 5.2.

Table 5.1: Bitrates (in Kbps) for the selected ODVs.

ODV	BR1	BR2	BR3	BR4	BR5
Basketball *Dancing*	500	1000	2000	5000	13000
Harbor *JamSession* *Gaslamp* *SkateboardTrick* *Trolley*	500	1000	2000	7000	13000
KiteFlite	500	1000	5000	7000	13000

Table 5.2: Statistics of the stimuli and the participants in the subjective quality assessment experiments.

Subjective Experiment	# of Stimuli	# of Participants	Min – Mean – Max Age	Ratio of Women
First	75 + 5 Ref	24	22 – 29.7 – 38	16%
Second	45 + 3 Ref	23	25 – 31.6— 42	26%

Experiment Setup

The subjective experiments were conducted in a dedicated experiment room equipped with an HTC Vive HMD, which was used to present the stimuli to the viewers. Participants were seated in a swivel chair and allowed to turn freely. To ensure that the participants could vote without removing the HMD, we used the Virtual Desktop application. Virtual Desktop is an ODV player and an application that enables the users to watch and interact with the desktop using the HMD and VR controllers. Using this application and the open-source MATLAB GUI presented in [224, 225], participants were able to vote each stimulus. Additionally, with a special application, the viewport trajectories were also recorded during the presentation of each stimulus for the computation of the visual attention maps.

Methodology

The modified-absolute category rating (M-ACR) [226] methodology was chosen for our subjective experiments. We chose M-ACR, because it was demonstrated in the evaluations [226, 227] that it is more reliable than existing methods developed for traditional video. This methodology increases the duration of exposition time by showing each stimulus twice with a mid-gray screen displayed for three seconds in between the two presentations of each stimulus. The reference sequences were also included in the subjective experiments as hidden references. That is, the participants were not told of reference sequences, and they voted the hidden references like any other stimulus.

The subjective quality scores for all the videos were collected in two experiments with different ODVs and participants. The first experiment comprised of two sessions of 30 minutes, one hour in total. The second experiment had only one session of 30 minutes. At the beginning of both experiments, there was a training phase when the *Train* video sequence with five different quality levels was displayed. After the training phase, the experiment ODVs were randomly displayed avoiding consecutive presentation of the same content, and the quality scores were assigned by the participants based on a continuous grading scale in the range [0,100], with 100 corresponding to the best score, as recommended in ITU-R BT.500-13 [228].

Participants

24 participants, 20 males and four females, took part in the first experiment. These participants were aged between 22 and 38 years with an average of 29.7 years. 23 participants, 17 males and six females, took part in the second experiment. These participants were aged between 25 and 42 years with an average of 31.6 years. The gathered quality scores were screened for outliers using the outlier detection method recommended in ITU-R BT.500-13 [228]. Three outliers in the first experiment and two outliers in the second experiment were found and removed. All participants were screened for visual acuity and found to have normal or corrected-to-normal vision.

5.4.3 Subjective Quality Analysis

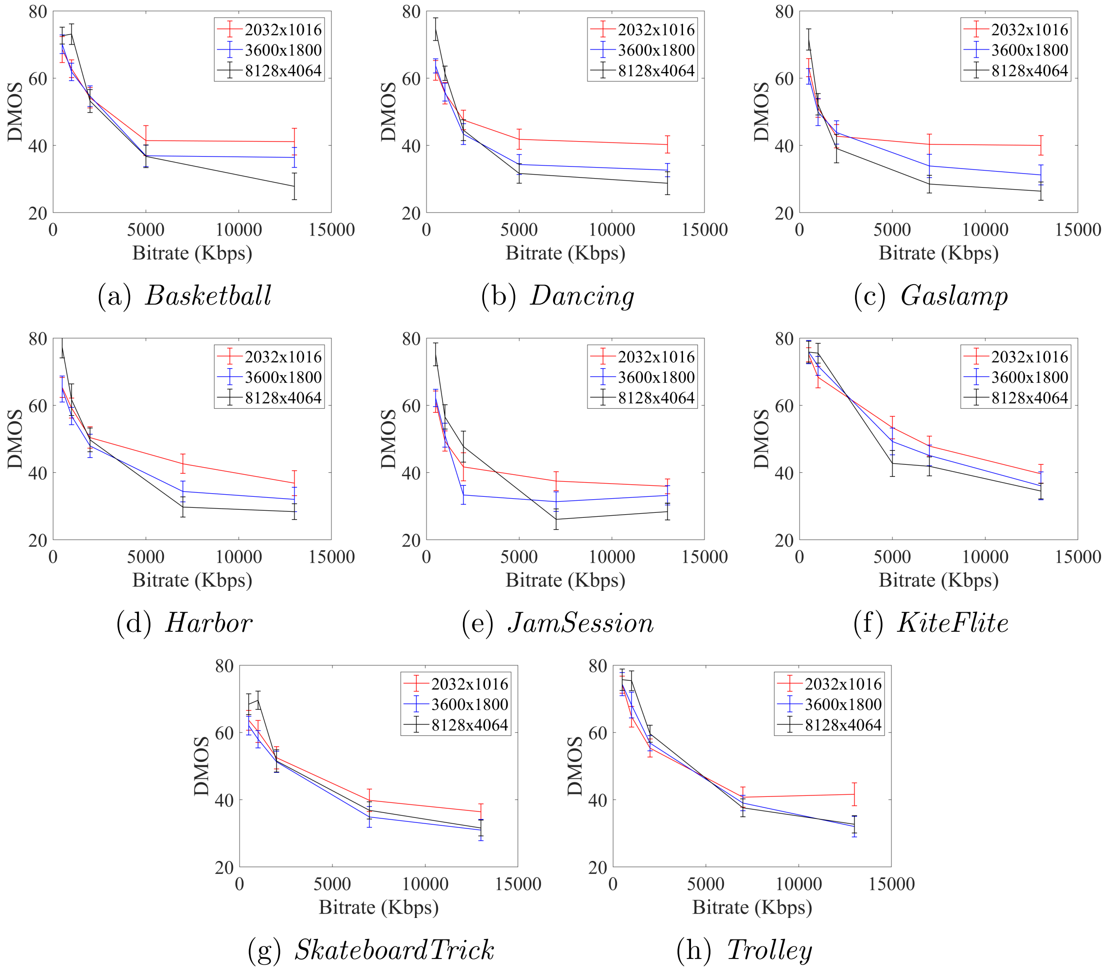
To represent the subjective quality of each stimulus, difference mean opinion scores (DMOS) [229] were calculated by applying the standard approach described in [230]. To calculate DMOS, first, the difference scores are computed as: $d_{ij} = s^r_{ij} - s_{ij}$, where s_{ij} and s^r_{ij} are the raw subjective score assigned by participant i to the distorted ODV j and the raw subjective score assigned to the corresponding hidden reference ODV, respectively. These difference scores d_{ij} are converted to z-scores as follows: $z_{ij} = (d_{ij} - \mu_i)/\sigma_i$, where μ_i and σ_i are the mean and standard deviation of the raw scores assigned by the participant i. Afterward, the outliers are detected based on the z-scores as recommended in ITU-R BT.500-13 [228]. Then, the z-scores are linearly rescaled in the interval [0,100] as follows: $z'_{ij} = 100(z_{ij}+3)/6$. The rescaling is based on the assumption that the z-scores z_{ij} are normally distributed with mean equal to zero and standard deviation equal to one, which means, that 99% of the z-scores z_{ij} are in the interval [-3,3], and consequently 99% of the rescaled z-scores z'_{ij} are in the interval [0,100]. The final DMOS of ODV j is then obtained by averaging the rescaled z-scores z'_{ij} of the K participants excluding the outliers as follows:

$$\mathrm{DMOS}_j = \frac{1}{K}\sum_{i=1}^{K} z'_{ij}. \tag{5.4}$$

Small DMOS indicate that the distorted stimulus is closer to the reference, and hence small DMOS is better. Figure 5.4 shows the DMOS of the ODVs included in the experiments. As expected, we can notice that there is an inverse relationship between DMOS and bitrate. From the plots, we can also see that the ODVs with the highest spatial resolution have the worst quality (highest DMOS) for low bitrate and the best quality for high bitrate. This shows that the 8128×4064 ODVs are coarsely compressed at the low bitrates due to the high number of pixels present. As the bitrate increases, the perceived quality of these videos gets better. Conversely, the perceived quality of the 2032×1016 ODVs becomes the worst at high bitrates, due to the scaling distortions [20]. These findings are especially important for ODV adaptive streaming systems [20], where the selection of the optimal encoding parameters is crucial.

5.4.4 Visual Attention Analysis

Table 5.3 shows the comparison between the visual attention maps of the reference ODVs and the corresponding ODVs with resolution 8128×4064 and encoded at the five bitrates reported in Table 5.1. For the comparison, first uniformly distributed points on the sphere are sampled from the visual attention maps, and then the Pearson's linear correlation coefficient (PLCC) and the Kullback–Leibler divergence (KLD) are applied to the sampled points [41]. Large PLCC values and small KLD values correspond to high similarity. As can be noticed from Table 5.3, the visual attention maps of the reference and corresponding distorted ODVs can be different, especially for the smallest bitrate BR1. This can also be noticed in Figure 5.5, where the visual attention maps of the *JamSession* reference ODV and the corresponding encoded ODVs at the smallest and largest bitrates with resolution 8128×4064 are shown. In Table 5.3, there is also the average of the PLCC and KLD values for each bitrate. It can be seen that by increasing the bitrate the average PLCC increases while the average KLD decreases. Based on these observations and to ensure the most accurate results, in our framework we use, for each undistorted and distorted ODV, the corresponding visual attention map and not only the visual attention maps of the undistorted ODVs.



(a) *Basketball* (b) *Dancing* (c) *Gaslamp*

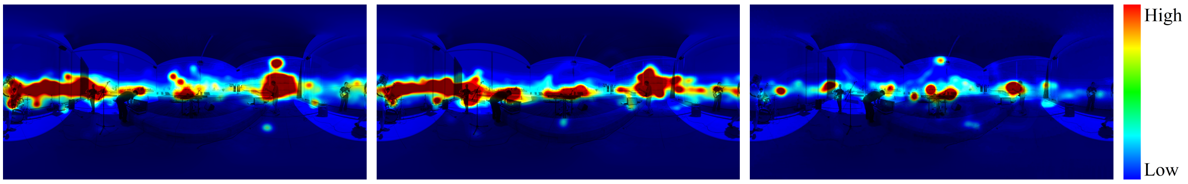
(d) *Harbor* (e) *JamSession* (f) *KiteFlite*

(g) *SkateboardTrick* (h) *Trolley*

Figure 5.4: Bitrate vs. DMOS plots of each ODV used in the subjective experiments. The vertical bars show 95% confidence intervals.

Table 5.3: PLCC and KLD computed between the visual attention maps of the reference ODVs and the corresponding ODVs with resolution 8128x4064 and encoded at the five bitrates reported in Table 5.1.

	BR1		BR2		BR3		BR4		BR5	
ODV	PLCC	KLD	PLCC	KLD	PLCC	KLD	PLCC	KLD	PLCC	KLD
Basketball	0.8914	0.5939	0.9134	0.6394	0.8838	0.7101	0.9019	0.8640	0.9195	0.6801
Dancing	0.6410	1.3625	0.6911	1.0891	0.7226	1.2005	0.7841	0.7137	0.7205	1.0115
Harbor	0.7316	0.7843	0.7134	0.6718	0.8341	0.4486	0.8348	0.5310	0.8536	0.4550
JamSession	0.5781	1.4356	0.8312	0.7140	0.7313	0.8753	0.8640	0.5990	0.8457	0.4435
KiteFlite	0.7273	0.8362	0.8136	1.0684	0.8486	0.5353	0.8352	0.6136	0.8557	0.5614
Gaslamp	0.7769	0.8339	0.7981	0.6213	0.8457	0.4773	0.8739	0.5137	0.8421	0.6517
SkateboardTrick	0.8705	0.7316	0.8713	1.0611	0.9413	0.4834	0.8901	0.5517	0.8976	0.3951
Trolley	0.8586	0.8713	0.7891	0.9610	0.8232	0.8207	0.8945	0.5879	0.9162	0.5906
Average	0.7594	0.9312	0.8026	0.8533	0.8288	0.6939	0.8598	0.6218	0.8564	0.5986



(a) *JamSession* 8128×4064 reference ODV.

(b) *JamSession* 8128×4064 encoded ODV at 13000 Kbps.

(c) *JamSession* 8128×4064 encoded ODV at 500 Kbps.

Figure 5.5: Comparison of the visual attention maps of the *JamSession* reference ODV and two corresponding encoded ODVs. See the color bar for the used color code.

5.5 Analysis and Evaluation

In this section, we first investigate the components of the proposed framework, and then we compare the metrics of the framework with existing quality metrics. With this aim, we use our ODV dataset with the gathered subjective quality scores presented in Section 5.4, and we analyze the correlation between the metric scores and the subjective scores. For the correlation analysis, we first convert the metric scores into the subjective scores by fitting the logistic function proposed in [141] and already presented in Section 3.5.3. Here, the subjective score predicted by the logistic function is the reversed DMOS (*i.e.*, subtracted from 100).

To evaluate how well the logistic function predicts the subjective scores, *i.e.*, how well the metric estimates the subjective quality, the following measures are applied to the real and predicted subjective scores: Pearson's linear correlation coefficient (PLCC), Spearman's rank ordered correlation coefficient (SROCC), root mean squared prediction error (RMSE), and mean absolute prediction error (MAE).

To visualize the relationship between the metric and subjective scores, sample plots are shown in Figure 5.6 for the metrics SSIM and VMAF applied to the ERP format (SSIM$_{ERP}$ and VMAF$_{ERP}$), and in the Voronoi-based quality assessment framework without and with visual attention. In these plots, the increase of the correlation between the metric scores and DMOS is noticeable for the VI-METRICs and the VI-VA-METRICs compared to the metrics calculated in ERP format.

5.5.1 Analysis of the Framework Components

In this section, we fine-tune the proposed framework by analyzing its components.

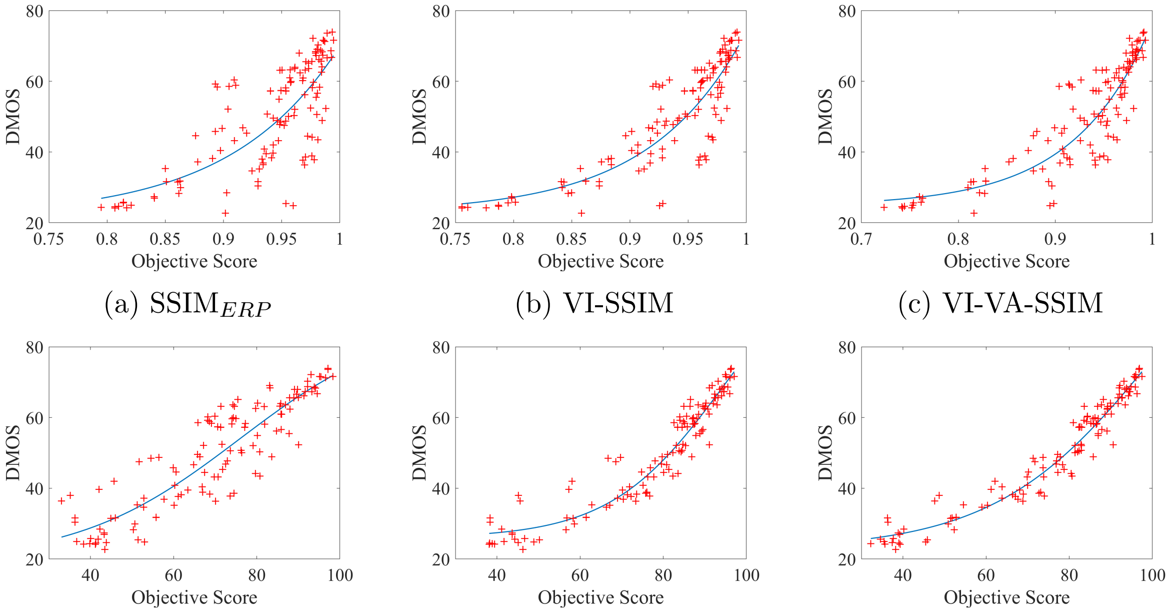
Estimation of the optimal angular resolution and number of planar Voronoi patches

We first analyze the two main parameters of the proposed framework that have an impact on the accuracy of the quality estimation, namely, the angular resolution and the number of the planar Voronoi patches.

For the Voronoi-based metrics obtained with the proposed framework without and with visual attention, *i.e.*, VI-METRICs and VI-VA-METRICs, Table 5.4 shows PLCC and SROCC for different parameter values. Three angular resolutions are investigated, namely $\{10, 15, 20\}$ pix/deg, which are close to the resolution of the HTC Vive HMD used in our subjective experiments. Moreover, we also consider three different numbers of planar Voronoi patches, that is, $M = \{10, 15, 20\}$.

Table 5.4: PLCC and SROCC of the Voronoi-based metrics with different angular resolutions and numbers of patches. The best performance values for each resolution (*i.e.*, each row) are in **bold**, while the best performance values among all the metrics are in **blue**.

		10 patches		15 patches		20 patches	
Metrics	Resolutions	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
VI-PSNR	10 pix/deg	0.8700	0.8584	**0.8775**	**0.8634**	0.8676	0.8551
	15 pix/deg	0.8700	0.8584	**0.8775**	**0.8636**	0.8675	0.8553
	20 pix/deg	0.8700	0.8584	**0.8775**	**0.8634**	0.8676	0.8553
VI-SSIM	10 pix/deg	0.8757	0.8667	0.8821	**0.8763**	**0.8823**	**0.8763**
	15 pix/deg	0.8423	0.8301	0.8509	0.8411	**0.8516**	**0.8414**
	20 pix/deg	0.8132	0.7995	0.8227	0.8072	**0.8237**	**0.8079**
VI-MS-SSIM	10 pix/deg	0.9468	0.9432	**0.9488**	0.9446	0.9486	**0.9450**
	15 pix/deg	0.9385	0.9361	**0.9411**	0.9381	0.9409	**0.9398**
	20 pix/deg	0.9314	0.9260	**0.9343**	**0.9303**	0.9339	0.9291
VI-VMAF	10 pix/deg	0.9634	0.9553	0.9615	0.9529	**0.9646**	**0.9581**
	15 pix/deg	0.9532	0.9444	0.9544	0.9470	**0.9581**	**0.9497**
	20 pix/deg	0.9387	0.9288	0.9435	0.9363	**0.9476**	**0.9401**
VI-VA-PSNR	10 pix/deg	**0.8977**	**0.8812**	0.8760	0.8563	0.8876	0.8712
	15 pix/deg	**0.8977**	**0.8817**	0.8760	0.8564	0.8876	0.8708
	20 pix/deg	**0.8977**	**0.8817**	0.8760	0.8564	0.8876	0.8707
VI-VA-SSIM	10 pix/deg	0.8947	0.8848	0.8921	0.8832	**0.9106**	**0.9007**
	15 pix/deg	0.8633	0.8510	0.8537	0.8426	**0.8777**	**0.8663**
	20 pix/deg	0.8353	0.8214	0.8188	0.8136	**0.8463**	**0.8323**
VI-VA-MS-SSIM	10 pix/deg	0.9563	0.9505	0.9628	0.9581	**0.9676**	**0.9635**
	15 pix/deg	0.9501	0.9438	0.9552	0.9506	**0.9627**	**0.9573**
	20 pix/deg	0.9445	0.9371	0.9482	0.9424	**0.9572**	**0.9517**
VI-VA-VMAF	10 pix/deg	0.9661	0.9589	0.9738	0.9667	**0.9773**	**0.9717**
	15 pix/deg	0.9580	0.9491	0.9678	0.9599	**0.9723**	**0.9658**
	20 pix/deg	0.9444	0.9349	0.9553	0.9482	**0.9623**	**0.9564**



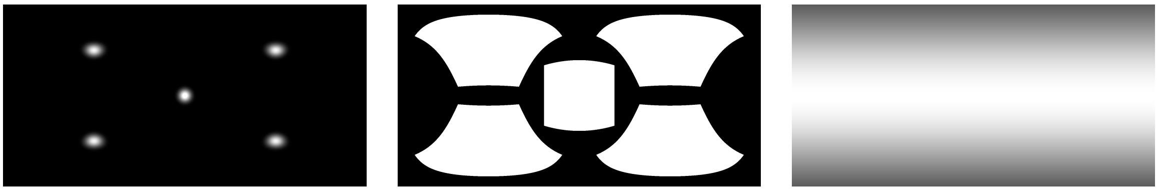
(a) SSIM$_{ERP}$ (b) VI-SSIM (c) VI-VA-SSIM

(d) VMAF$_{ERP}$ (e) VI-VMAF (f) VI-VA-VMAF

Figure 5.6: Metric vs subjective score plots with the fitted logistic functions. Red points indicate the data points, and blue lines indicate the logistic functions.

As can be seen in the table, the reduction of the patch resolution improves the performance of the Voronoi-based metrics in most of the cases. For the other cases, the performance remains almost constant. On the other hand, increasing the number of patches seems to positively influence the performance of the Voronoi-based metrics almost always, except for VI-PSNR and VI-VA-PSNR. This can be explained by the reduction of the projection distortions when the number of patches increases and consequently the patch size decreases. For the VI-VA-METRICs that use visual attention, the improvement of the performance can also be explained by the fact that with more patches the visual attention weights $\nu_{i,k}$ are localized to smaller regions and consequently more accurate.

As a result of this analysis, we select 10 pix/deg and 20 patches ($M = 20$) as the optimal parameter values for our proposed framework. We use these two parameters for the rest of this chapter. Please note that although we select these optimal parameter values, independently of the studied parameter values, the Voronoi-based metrics are characterized by a better performance than the performance of the corresponding original metrics for traditional video applied to the ERP and CMP formats, as shown later in Table 5.7.



(a) Kent distribution method.

(b) Uniform viewport method.

(c) Equator bias method.

Figure 5.7: Visual attention maps computed with different methods using as input five different viewport positions.

Investigation of applying different visual attention estimation methods

The proposed quality framework can make use of different visual attention estimation methods, as the visual attention weights $\nu_{i,k}$ can be computed from any visual attention map generated by different algorithms. Here, we investigate the effect of three different visual attention methods on VI-VA-METRIC performance, namely, the Kent distribution method [45], the uniform viewport method, and the equator-bias method. The first of the three estimation methods is based on the Kent distribution, which is a Gaussian distribution defined on the surface of a unit sphere, as explained in [45]. With this method, we compute the visual attention maps using the viewport trajectories and the default parameters proposed in [45]. For the uniform viewport method, we also use viewport trajectories from the viewers. In this method, each point of the viewport trajectories is replaced with a uniform viewport that is projected to ERP. The final visual attention map is obtained as the summation of the projected viewports. The equator-bias method does not require the viewport trajectories. Instead, it computes the visual attention map as a vertical bias from the equator defined by the Gaussian curve centered on the equator. Figure 5.7 shows the visual attention maps obtained with these three methods based on five discrete viewport positions.

Table 5.5 shows the performance of the Voronoi-based metrics integrated with visual attention. As can be noticed, both the Kent distribution method [45] and the uniform viewport method are able to improve the performance of the Voronoi-based metrics. On the other hand, the equator-bias method is capable to improve VI-PSNR and VI-SSIM, while the performance values of VI-MS-SSIM and VI-VMAF remain almost constant. In conclusion, these results show that adding a characterization of the actual parts of the ODV that are likely watched improves the performance of the Voronoi-based metrics. As can be seen from the table, the metrics of the proposed framework achieve the best performance when applying the Kent distribution method. Since this method is the most plausible and similar to the human eye-tracking results [44], it is expected to perform

Table 5.5: Performance evaluation of the Voronoi-based metrics integrated with visual attention estimated with three methods. The best performance values are in **bold**.

Metrics	Vis. Att.	PLCC	SROCC	RMSE	MAE
VI-PSNR	–	0.8676	0.8551	7.5743	5.8377
VI-VA-PSNR	Equator-bias	0.8781	0.8628	7.2995	5.5508
VI-VA-PSNR	Uniform	0.8774	0.8585	7.4141	5.7168
VI-VA-PSNR	Kent	0.8876	0.8712	7.1818	5.5072
VI-SSIM	–	0.8823	0.8763	7.1172	5.2867
VI-VA-SSIM	Equator-bias	0.8879	0.8850	6.9454	5.1687
VI-VA-SSIM	Uniform	0.8981	0.8929	6.8103	5.0647
VI-VA-SSIM	Kent	0.9106	0.9007	6.4345	4.8097
VI-MS-SSIM	–	0.9486	0.9450	4.8743	3.8475
VI-VA-MS-SSIM	Equator-bias	0.9486	0.9450	4.8790	3.8343
VI-VA-MS-SSIM	Uniform	0.9634	0.9583	4.1350	3.3506
VI-VA-MS-SSIM	Kent	0.9676	0.9635	3.8982	3.1526
VI-VMAF	–	0.9646	0.9581	4.2096	3.1548
VI-VA-VMAF	Equator-bias	0.9650	0.9576	4.1959	3.1393
VI-VA-VMAF	Uniform	0.9749	0.9671	3.5602	2.7569
VI-VA-VMAF	Kent	**0.9773**	**0.9717**	**3.3753**	**2.5948**

better than the other methods. Therefore, we use the visual attention maps estimated by the Kent distribution method in the rest of this thesis.

Investigation of different temporal pooling methods of the frame scores

Since the selection of the temporal pooling method P_{tempo} for the combination of the frame scores T_i (see Equation 5.3) might affect the overall performance, in this paper, we also investigate its effect. For this purpose, motivated by the pooling methods which are used in VMAF code [231], we evaluate the following ones: mean, harmonic mean, min, median, 5th percentile, 10th percentile, and 20th percentile. Table 5.6 shows the performance of VI-VA-VMAF with these pooling methods. As can be noticed, the performance is not influenced too much by the choice of the pooling method. Therefore, in the rest of the paper, we consider only the mean pooling method.

5.5.2 Comparison with Existing Metrics

This section evaluates the performance of our Voronoi-based metrics and existing well-known metrics used in ODV quality assessment studies. Four of the existing metrics that we evaluate were developed for traditional image/video quality assessment: PSNR, SSIM [188], MS-SSIM [217], and VMAF [213]. These metrics were applied to ODVs in two different formats, namely, ERP and CMP, and to distinguish them we use a subscript,

Table 5.6: Comparison of different temporal pooling methods for the combination of the frame scores applied in VI-VA-VMAF.

Pooling	PLCC	SROCC	RMSE	MAE
Mean	0.9773	0.9717	3.3753	2.5948
Harmonic Mean	0.9775	0.9718	3.3681	2.5911
Min	0.9753	0.9705	3.4920	2.6887
Median	0.9761	0.9715	3.4093	2.6275
5th Percentile	0.9759	0.9708	3.4489	2.6437
10th Percentile	0.9776	0.9711	3.3636	2.5776
20th Percentile	0.9764	0.9714	3.3866	2.6041

e.g. PSNR_{ERP} and PSNR_{CMP}. Moreover, we analyze extra four metrics which were specifically designed for ODV: S-PSNR-I [36], S-PSNR-NN [36], WS-PSNR [34], and CPP-PSNR [35]. The implementation used in our evaluation for PSNR, SSIM, and MS-SSIM is the one provided by the Video Quality Measurement Tool [232]; for VMAF we used the code provided by its developers [231]; while for S-PSNR-I, S-PSNR-NN, WS-PSNR, and CPP-PSNR, we used the 360Lib standard software [233].

Table 5.7 shows the performance evaluation of the selected existing metrics and our Voronoi-based metrics. By looking at the results, we can notice a slightly higher correlation between the subjective and metric scores when the metrics PSNR, SSIM, and VMAF are applied to the CMP format instead of the ERP format. The reason for this could be the lower projection distortions of CMP compared to ERP. We also observe that the performance of the PSNR-based metrics developed for ODV is better than the performance of the traditional PSNR. Furthermore, among all the evaluated metrics in Table 5.7, SSIM is characterized by the worst performance, even worse than PSNR. The reason might be that the inevitable projection distortions negatively affect the performance of SSIM, as some regions are stretched to much bigger areas (especially the top and bottom parts of ERP). Therefore, SSIM scores could be dominated by these regions, and this could cause SSIM to have lower correlation scores than PSNR, even though, for traditional video, SSIM is much closer to human perception than PSNR. On the other hand, among the selected existing metrics that are not Voronoi-based, MS-SSIM and VMAF have the best performance. This is not unexpected, since these metrics, which have state-of-the-art performance for traditional video [216], consider scaling and compression distortions that characterize our dataset. Between these two metrics, MS-SSIM is slightly better than VMAF for both projection formats. The reason can be explained by the fact that VMAF was neither modeled for 8K nor ODV.

The results also show that when the metrics are applied to planar Voronoi patches instead of the ERP and CMP formats, they achieve better performance. This is expected

Table 5.7: Performance evaluation of the selected existing metrics and our Voronoi-based metrics. The subscripts ERP and CMP indicate when the metric is applied to the corresponding projection formats. The best performance values are in **bold**.

Metrics	PLCC	SROCC	RMSE	MAE
PSNR$_{ERP}$	0.8408	0.8237	8.2326	6.3169
PSNR$_{CMP}$	0.8480	0.8323	8.0419	6.2085
S-PSNR-I	0.8580	0.8438	7.8207	5.9715
S-PSNR-NN	0.8584	0.8433	7.8066	5.9648
WS-PSNR	0.8582	0.8430	7.8107	5.9772
CPP-PSNR	0.8579	0.8439	7.8200	5.9779
SSIM$_{ERP}$	0.7659	0.7551	9.7734	7.7396
SSIM$_{CMP}$	0.7701	0.7546	9.6583	7.6036
MS-SSIM$_{ERP}$	0.9224	0.9160	5.8232	4.4205
MS-SSIM$_{CMP}$	0.9132	0.9081	6.1422	4.7378
VMAF$_{ERP}$	0.8978	0.8864	6.7433	5.3631
VMAF$_{CMP}$	0.9063	0.8945	6.5630	5.2229
VI-PSNR	0.8676	0.8551	7.5743	5.8377
VI-SSIM	0.8823	0.8763	7.1172	5.2867
VI-MS-SSIM	0.9486	0.9450	4.8743	3.8475
VI-VMAF	0.9646	0.9581	4.2096	3.1548
VI-VA-PSNR	0.8876	0.8712	7.1818	5.5072
VI-VA-SSIM	0.9106	0.9007	6.4345	4.8097
VI-VA-MS-SSIM	0.9676	0.9635	3.8982	3.1526
VI-VA-VMAF	**0.9773**	**0.9717**	**3.3753**	**2.5948**

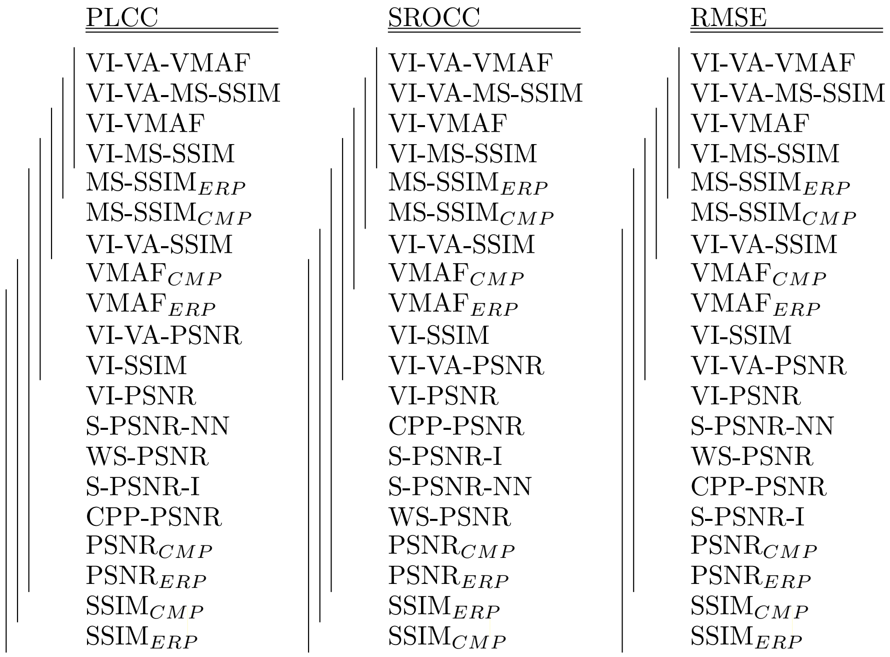


Figure 5.8: Statistical significance analysis of the difference between PLCC, SROCC, and RMSE of the quality metrics, obtained according to ITU-T Recommendation P.1401 [9]. There is statistically significant equivalence between two quality metrics, if they are aligned with the same vertical bar; e.g., there is a statistically significant difference between VI-VA-VMAF and MS-SSIM$_{ERP}$ in terms of PCC, SROCC, and RMSE.

because of the lower projection distortions of the planar Voronoi patches compared to ERP and CMP, and because of the similar angular resolutions of the patches and the HMD viewport. Moreover, as already noticed before, the Voronoi-based metrics integrated with visual attention (*i.e.*, VI-VA-METRICs) achieve better performance than the corresponding ones without visual attention (*i.e.*, VI-METRICs). The best performing metric among all compared is VI-VA-VMAF followed by VI-VA-MS-SSIM.

In addition to the numerical results, a statistical significance analysis of the difference between PLCC, SROCC, and RMSE of the quality metrics was conducted according to ITU-T Recommendation P.1401 [9]. Figure 5.8 illustrates the statistical significance analysis of the evaluated metrics in Table 5.7. The vertical bars show that there is no statistically significant difference between the metrics aligned with the same bar. As can be noticed in Figure 5.8, the first four best quality metrics are statistically equivalent. The significance analysis results also show that the addition of visual attention might not always yield a statistically significant difference. Nevertheless, the numerical results show that integrating visual attention improved the metric performance in all the cases, as we can also see in Table 5.4.

To further evaluate the Voronoi-based metrics in a different condition and analyze the effect of different spatial resolutions of the ODVs, we calculate the correlation coefficients separately for each spatial resolution of our dataset (*i.e.*, 2K, 4K, and 8K). The results of this analysis are shown in Table 5.8. It is interesting to notice that for most of the selected existing and Voronoi-based metrics the correlations PLCC and SROCC improve when the resolution is increased. This can be attributed to scaling distortions (blur) present at 2K and 4K resolutions. Assuming that most of the metrics were developed mainly for compression distortions and/or noise, the presence of scaling distortions could decrease the correlation between DMOS and metric scores in the cases of 2K and 4K. Nevertheless, we notice again that the integration of visual attention increases the performance of the Voronoi-based metrics.

Regardless of the case, the integration of visual attention (*i.e.*, VI-VA-METRICs) improves the Voronoi-based metrics (*i.e.*, VI-METRICs) in every situation. These improvements can be seen not only in Table 5.7 but also in Table 5.5 and 5.8. These consistent improvements show that the integration of visual attention is an important factor to consider in objective ODV quality assessment in the presence of compression and scaling distortions, *i.e.*, uniform artifacts, and it needs to be taken into account to increase the metric performance.

5.5.3 Limitations of the Proposed Framework and Future Improvements

As discussed in the previous section, the proposed framework integrated with visual attention achieves state-of-the-art performance. Nevertheless, it has also limitations that we plan to tackle in future work.

First, the current framework only considers visual attention maps generated using the viewport trajectories collected from the participants of subjective experiments. In practice, this type of data is not available, as it is not possible to find the viewport trajectories for new content without conducting a subjective experiment first. Instead, automatic visual attention estimation algorithms [234] might be used for most of the practical cases. Nevertheless, the integration of the said automatic visual attention estimation methods and the performance analysis in this case remain as future work.

Second, in our study and in particular in our dataset, we considered only the typical artifacts introduced by the encoding pipeline of the adaptive streaming systems, *i.e.*, compression and scaling distortions. However, the ODV processing pipeline can introduce other visual artifacts (see Section 3.2). The perceptual impact of the other visual artifacts can be investigated and integrated into our proposed framework.

Third, with the current unoptimized code, the computation of VI-VA-VMAF requires

Table 5.8: PLCC and SROCC of the evaluated metrics computed separately for the resolutions 2K, 4K, and 8K. The best performance values for each resolution are in **bold**.

	2K		*4K*		*8K*	
Metrics	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PSNR$_{ERP}$	0.7388	0.6139	0.8360	0.8343	0.9202	0.9183
PSNR$_{CMP}$	0.7517	0.6203	0.8431	0.8450	0.9221	0.9163
S-PSNR-I	0.7634	0.6469	0.8568	0.8615	0.9304	0.9228
S-PSNR-NN	0.7649	0.6433	0.8570	0.8574	0.9300	0.9227
WS-PSNR	0.7650	0.6366	0.8570	0.8574	0.9299	0.9230
CPP-PSNR	0.7638	0.6432	0.8567	0.8615	0.9302	0.9230
SSIM$_{ERP}$	0.6996	0.5570	0.7703	0.7951	0.8600	0.8482
SSIM$_{CMP}$	0.7011	0.5591	0.7714	0.7878	0.8565	0.8484
MS-SSIM$_{ERP}$	0.8841	0.7992	0.9150	0.9351	0.9652	0.9478
MS-SSIM$_{CMP}$	0.8673	0.7824	0.9071	0.9276	0.9583	0.9446
VMAF$_{ERP}$	0.9202	0.8735	0.9203	0.9071	0.9515	0.9240
VMAF$_{CMP}$	0.9226	0.8790	0.9309	0.9156	0.9567	0.9285
VI-PSNR	0.7640	0.6321	0.8660	0.8769	0.9358	0.9247
VI-SSIM	0.8346	0.7109	0.8794	0.9060	0.9367	0.9249
VI-MS-SSIM	0.8642	0.8807	0.8140	0.9437	0.9767	0.9557
VI-VMAF	0.9627	0.9287	0.9577	0.9458	0.9789	0.9500
VI-VA-PSNR	0.7960	0.6644	0.9050	0.9006	0.9451	0.9321
VI-VA-SSIM	0.8434	0.7326	0.9200	0.9321	0.9593	0.9392
VI-VA-MS-SSIM	0.9529	0.9105	0.8332	**0.9674**	0.9829	**0.9634**
VI-VA-VMAF	**0.9762**	**0.9493**	**0.9737**	0.9625	**0.9862**	0.9593

considerable computational resources. For an 8K ODV with 300 frames, the computation of VI-VA-VMAF with 20 patches and with 10 pix/deg patch resolution takes about three minutes using a PC with a 4GHz Intel Core i7-6700K processor. Moreover, VI-VA-VMAF requires as input also a visual attention map for each frame. On a machine with two Intel Xeon Gold 6134 processors, the parallel computation of 400x800 visual attention maps using the code of the Kent method provided in [45] takes about nine seconds per map.

5.6 Conclusions

This Chapter presented a framework for objective ODV quality assessment that takes into account the spherical nature of ODV and the ODV viewing characteristics. The proposed framework is based on the subdivision of ODV into planar Voronoi patches with low projection distortions obtained with the spherical Voronoi diagram. Furthermore, it also exploits visual attention to identify the regions that are consumed by the viewer with high probability, which have a big influence on the perception of the video quality. For the evaluation of the framework, we created an ODV dataset with a total of 120 distorted videos from 8 undistorted reference videos. Subjective scores and viewport trajectories for the new ODVs were also collected in subjective experiments.

In the evaluation of the framework, first the framework components were analyzed. This analysis showed how planar Voronoi patches and visual attention are important to achieve a high correlation between subjective and metric scores. Moreover, the framework was also compared with existing metrics, and this showed that our framework can achieve state-of-the-art performance.

As future work, we plan to further explore the visual attention estimation methods for ODV that do not require viewport trajectories, as visual attention maps obtained from subjective tests are not practical to obtain. We also intend to extend our framework to distortions different from the ones considered here, *i.e.*, compression and scaling distortions.

Chapter 6

Conclusions

Omnidirectional video has been gaining popularity especially in recent years. Nevertheless, there still remain some important challenges that can introduce distortions in ODV and compromise the quality of experience while watching it. This thesis proposes solutions that help overcome these challenges and improve the quality of ODV. This final chapter summarizes how the research objectives were realized and how the research question was answered. First, the thesis contributions are recapped, and then the future work is presented.

6.1 Summary

Thanks to the maturity of the ODV technology and the facility of accessing and consuming ODV, we are experiencing a marked growth in the popularity of this immersive imaging technique. The current technology of production, coding, and transmission of ODV has reached a level of maturity able to attract content producers and foster the success of ODV. Nevertheless, there are still various important challenges that need to be tackled. Examples are the artifacts introduced during capture and post-production, and the challenges related to coding and transmission. This thesis contributes to the solution of these challenges. Specifically, at the beginning of the thesis, three research areas were identified, and three corresponding research objectives were defined:

1. **Artifact Detection:** our objective was to develop methods usable by artists for the detection and localization of artifacts mostly introduced during capture and post-production. The realization of this objective, which is described in Chapter 3, is a general artifact detection framework that extends detection methods for S3D standard images to S3D ODIs based on a new subdivision of ODIs, namely planar Voronoi patches, and visual attention. Moreover, two methods for the detection of

sharpness mismatch and one method for the detection of color mismatch were developed. We also showed that one of our two sharpness mismatch detection methods reaches superior performance than the existing state-of-the-art approaches. For the evaluation of the framework, a dataset of S3D ODIs with visual attention data was created.

2. **Artifact Correction:** the aim was to design methods for the correction of common artifacts that can improve quality and reduce time and efforts in the production workflow. To achieve this goal, we decided to study one of the most common artifacts, namely color mismatch, and develop methods for its correction, which are described in Chapter 4. Precisely, one approach based on traditional visual computing techniques and two alternative deep learning-based solutions were developed.

3. **Quality Assessment:** the goal was to develop quality metrics for the optimization of coding and transmission solutions for ODV. To fulfill this objective, we developed a general framework for full-reference objective quality assessment of monoscopic ODV based on planar Voronoi patches and visual attention, which is described in Chapter 5. For its evaluation, a new ODV dataset with subjective quality scores and visual attention data was created. The comparison with commonly used quality metrics for ODV shows the state-of-the-art performance of our quality framework.

In conclusion, the original research question regarding ways to optimize the quality of ODV was answered by realizing the three research objectives with the contributions briefly summarized in this chapter.

6.2 Future Work

In the short term, some aspects of the contributions could be further studied and developed. For example, this thesis considers two common artifacts present in ODV in particular, namely color and sharpness mismatch. There are other common artifacts like stitching and blending artifacts that require better solutions to conceal them. In particular, stitching artifacts are almost inevitable in the case of monoscopic ODV captured with different cameras not sharing the same center of projection. Artists usually hide them in regions with low visual attention, but it would be better to find a way to remove them. This would probably require estimating the scene geometry or applying inpainting methods that would fill the regions occupied by the stitching lines.

Moreover, in our solutions, we used visual attention estimated from viewers. It would be useful to replace visual attention estimated from viewers with visual attention estimated by algorithms. In particular, it would be interesting to investigate the performance of the full-reference quality framework presented in Chapter 5 with visual attention estimated by algorithms instead of visual attention estimated from viewers.

Another point is the difficulty of correspondence estimation in S3D images in the presence of artifacts. In our solutions, we used correspondence estimation methods that seem to be accurate also in the presence of artifacts even if they were developed for artifact-free S3D images. We could study in-depth how artifacts influence these methods, and if necessary, develop new approaches robust to them.

Furthermore, deep learning started to become popular quite recently, but it has already been successfully applied in different fields like visual computing, speech processing, etc. We used deep learning only for the correction of color mismatch. Nevertheless, it could be applied also for the correction of other artifacts, and even for other tasks like artifact detection and quality assessment [235]. A requirement of deep learning-based solutions that must be taken into consideration is large training datasets. Creating them is not easy, especially for quality assessment, where ODIs or ODVs need to be annotated with subjective scores.

In the long term, the assessment of the QoE when watching ODV could be further studied by taking into account not only the visual quality considered in this thesis, but also other factors like the audio signal, the user discomfort, and the HMD ergonomics.

This research is a contribution to the improvement of the ODV technology. For the challenges that remain to be tackled, we hope that the solutions presented in this thesis could be further developed to solve at least some of them.

Appendix A

Abbreviations

Short Term	**Expanded Term**
CM	Color Mismatch
CNN	Convolutional Neural Network
CPBD	Cumulative Probability of Blur Detection
CPP-PSNR	Craster Parabolic Projection PSNR
DMOS	Difference Mean Opinion Score
EMD	Earth Mover's Distance
ERP	Equirectangular Projection
FSIM	Feature Similarity Index Measure
GMM	Gaussian Mixture Model
HMD	Head-Mounted Display
HSMD	Histogram-Based Sharpness Mismatch Detection Method
HVS	Human Visual System
JNB	Just Noticeable Blur
JNSM	Just Noticeable Sharpness Mismatch
KLD	Kullback-Leibler Divergence
M-ACR	Modified-Absolute Category Rating
MAE	Mean Absolute Prediction Error
MOS	Mean Opinion Score
MS-SSIM	Multi-Scale Structural Similarity Index Measure
MSE	Mean Squared Error
NN	Neural Network
ODI	Omnidirectional Image
ODV	Omnidirectional Video

OR	Outlier Ratio
PAM	Parallax Attention Mechanism
PLCC	Pearson's Linear Correlation Coefficient
PSM	Probability of Sharpness Mismatch
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
RMSE	Root Mean Squared Prediction Error
S-PSNR	Spherical PSNR
S-PSNR-I	Spherical PSNR with Bicubic Interpolation
S-PSNR-NN	Spherical PSNR with Nearest Neighbor Interpolation
S3D	Stereoscopic 3D
SM	Sharpness Mismatch
SROCC	Spearman's Rank Ordered Correlation Coefficient
SSIM	Structural Similarity Index Measure
TPS	Thin Plate Spline
VA	Visual Attention
VI-METRIC	Voronoi-Based Metric
VI-VA-METRIC	Voronoi-Based Metric with Visual Attention
VIF	Visual Information Fidelity
VMAF	Video Multimethod Assessment Fusion Metric
Voro-CMC	Voronoi-Based Color Mismatch Correction Method
VR	Virtual Reality
WS-PSNR	Weighted Spherical PSNR

Bibliography

[1] X. Wang, M. Yu, Y. Yang, and G. Jiang, "Research on subjective stereoscopic image quality assessment," in *Proceedings of the IS&T/SPIE Electronic Imaging: Multimedia Content Access: Algorithms and Systems III*, vol. 7255, 2009.

[2] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-reference quality assessment of natural stereopairs," *IEEE Transactions on Image Processing*, vol. 22, pp. 3379–3391, Sept. 2013.

[3] F. Pitié and R. Dahyot, "Automated colour grading using colour distribution transfer," *Computer Vision and Image Understanding*, vol. 107, 2007.

[4] K. Keodara, "History and future of 360°VR." https://visuon.com/s/history-and-future-of-360vr/. Accessed: 2021-05-01.

[5] Google Inc., *Rendering Omni-directional Stereo Content*.

[6] C. Richardt, "Omnidirectional stereo," in *Computer Vision* (K. Ikeuchi, ed.), pp. 1–4, Springer, Jan. 2020.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[9] ITU-T, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models." ITU-T Recommendation P.1401, Jul. 2012.

[10] Coursera, "VR and 360 video production." https://www.coursera.org/learn/360-vr-video-production. Accessed: 2021-05-01.

[11] K. Kane, "Robert Barker's Leicester Square Panorama: The Rotunda." https://regencyredingote.wordpress.com/2012/08/03/robert-barkers-leicester-square-panorama-the-rotunda/. Accessed: 2021-05-01.

[12] British Library, "The spectable of the panorama." https://www.bl.uk/picturing-places/articles/the-spectacle-of-the-panorama. Accessed: 2021-05-01.

[13] Library of Congress, "A brief history of panoramic photography." https://www.loc.gov/collections/panoramic-photographs/articles-and-essays/a-brief-history-of-panoramic-photography. Accessed: 2021-05-01.

[14] U. Shukla, "An introduction to 360° video." https://studio.knightlab.com/results/storytelling-layers-on-360-video/an-introduction-to-360-video/. Accessed: 2021-05-01.

[15] S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic, "Director's cut - a combined dataset for visual attention analysis in cinematic VR content," in *The 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018.

[16] A. Rana, C. Ozcinar, and A. Smolic, "Towards generating ambisonics using audio-visual cue for virtual reality," in *44th International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, 2019.

[17] "The invisible man - VR/360 Short Film." https://www.facebook.com/theinvisibleman360/, 2016.

[18] L. Sun, Y. Mao, T. Zong, Y. Liu, and Y. Wang, "Flocking-based live streaming of 360-degree video," in *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20, (New York, NY, USA), p. 26–37, Association for Computing Machinery, 2020.

[19] Arxel Tribe and Réunion des Musées Nationaux, "Pompei: The Legend of Vesuvius." https://en.wikipedia.org/wiki/Pompei:_The_Legend_of_Vesuvius, 2000.

[20] C. Ozcinar, A. De Abreu, S. Knorr, and A. Smolic, "Estimation of optimal encoding ladders for tiled 360° VR video in adaptive streaming systems," in *The 19th IEEE*

International Symposium on Multimedia (ISM 2017), (Taichung, Taiwan), Nov. 2017.

[21] D. E. Warburton, S. S. Bredin, L. T. Horita, D. Zbogar, J. M. Scott, B. T. Esch, and R. E. Rhodes, "The health benefits of interactive video game exercise," *Applied Physiology, Nutrition, and Metabolism*, vol. 32, no. 4, pp. 655–663, 2007.

[22] L. Freina and M. Ott, "A literature review on immersive virtual reality in education: state of the art and perspectives," in *The International Scientific Conference eLearning and Software for Education*, vol. 1, p. 133, " Carol I" National Defence University, 2015.

[23] F. Zhang and F. Liu, "Casual stereoscopic panorama stitching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2002–2010, June 2015.

[24] B. Mendiburu, *3D Movie Making.* Tayler & Francis, 2009.

[25] S. Knorr, K. Ide, M. Kunter, and T. Sikora, "The avoidance of visual discomfort and basic rules for producing "good 3D" pictures," *SMPTE Motion Imaging Journal*, vol. 121, no. 7, pp. 72–79, 2012.

[26] S. Möller and A. Raake, *Quality of experience: Advanced concepts, applications and methods.* Springer, 2014.

[27] M. Lambooij, W. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, p. 30201, 2009.

[28] D. Vatolin, A. Bokov, M. Erofeev, and V. Napadovsky, "Trends in S3D Movie Quality Evaluated on 105 Films Using 10 Metrics," *Electronic Imaging*, vol. 2016, pp. 1–10, 02 2016.

[29] K. Terzić and M. Hansard, "Methods for reducing visual discomfort in stereoscopic 3D: A review," *Signal Processing: Image Communication*, vol. 47, pp. 402–416, 2016.

[30] A. Voronov, D. Vatolin, D. Sumin, V. Napadovsky, and A. Borisov, "Methodology for stereoscopic motion-picture quality assessment," in *Proceedings of the SPIE, Stereoscopic Displays and Applications XXIV*, vol. 8648, 2013.

[31] M. Liu, K. Müller, and A. Raake, "Efficient no-reference metric for sharpness mismatch artifact between stereoscopic views," *Journal of Visual Communication and Image Representation*, vol. 39, pp. 132–141, 2016.

[32] X. Zheng, N. Yuzhen, J. Chen, and Y. Chen, "Color correction for stereoscopic image based on matching and optimization," in *IEEE International Conference on 3D Immersion (IC3D)*, (Brussels), 2017.

[33] C. Ozcinar, A. De Abreu, and A. Smolic, "Viewport-aware adaptive 360° video streaming using tiles for virtual reality," in *2017 International Conference on Image Processing (ICIP)*, (Beijing, China), Sept. 2017.

[34] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, pp. 1408–1412, Sept. 2017.

[35] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," *Proc.SPIE*, vol. 9970, 2016.

[36] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 31–36, Sept. 2015.

[37] 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," *CoRR*, vol. abs/1807.10990, 2018.

[38] E. Upenik and T. Ebrahimi, "Saliency driven perceptual quality metric for omnidirectional visual content," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4335–4339, Sep. 2019.

[39] C. Ozcinar, J. Cabrera, and A. Smolic, "Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, pp. 217–230, March 2019.

[40] A. Singla, S. Fremerey, A. Raake, P. List, and B. Feiten, "AhG8: Measurement of user exploration behavior for omnidirectinal (360°) videos with a head mounted display," tech. rep., International Telecommunication Union (ITU), Oct. 2017. Macau, China.

[41] J. Gutiérrez, E. David, Y. Rai, and P. Le Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images," *Signal Processing: Image Communication*, vol. 69, pp. 35–42, 2018.

[42] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 432–437, ACM, 2018.

[43] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, 2017.

[44] Y. Rai, P. Le Callet, and P. Guillotel, "Which saliency weighting for omni directional image quality assessment?," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, 2017.

[45] B. John, P. Raiturkar, O. Le Meur, and E. Jain, "A Benchmark of Four Methods for Generating 360° Saliency Maps from Eye Tracking Data," in *Proceedings of The First IEEE International Conference on Artificial Intelligence and Virtual Reality*, (Taichung, Taiwan), Dec. 2018.

[46] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *10th International Conference on Quality of Multimedia Experience (QoMEX 2018)*, (Sardinia, Italy), May 2018.

[47] S. Knorr, **Croci, Simone**, and A. Smolic, "A modular scheme for artifact detection in stereoscopic omni-directional images," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, (Maynooth, Ireland), 2017.

[48] **Simone Croci**, S. Knorr, and A. Smolic, "Saliency-based sharpness mismatch detection for stereoscopic omnidirectional images," in *14th European Conference on Visual Media Production (CVMP)*, (London, UK), 2017.

[49] **Simone Croci**, S. Knorr, L. Goldmann, and A. Smolic, "A framework for quality control in cinematic VR based on Voronoi patches and saliency," in *IEEE International Conference on 3D Immersion (IC3D)*, (Brussels, Belgium), 2017.

[50] **Simone Croci**, S. Knorr, and A. Smolic, "Sharpness mismatch detection in stereoscopic content with 360-degree capability," in *IEEE International Conference on Image Processing (ICIP)*, (Athens, Greece), 2018.

[51] **Simone Croci**, M. Grogan, S. Knorr, and A. Smolic, "Colour correction for stereoscopic omnidirectional images," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, (Belfast, UK), 2018.

[52] R. Dudek, **Simone Croci**, A. Smolic, and S. Knorr, "Robust global and local color matching in stereoscopic omnidirectional content," *Elsevier Signal Processing: Image Communication*, vol. 74, pp. 231–241, 2019.

[53] **Simone Croci**, S. Knorr, and A. Smolic, "Study on the perception of sharpness mismatch in stereoscopic video," in *11th International Conference on Quality of Multimedia Experience (QoMEX)*, (Berlin, Germany), 2019.

[54] **Simone Croci**, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Voronoi-based objective quality metrics for omnidirectional video," in *11th International Conference on Quality of Multimedia Experience (QoMEX)*, (Berlin, Germany), 2019.

[55] **Simone Croci**, C. Ozcinar, E. Zerman, S. Knorr, J. Cabrera, and A. Smolic, "Visual attention-aware quality estimation framework for omnidirectional video using spherical Voronoi diagram," *Springer Quality and User Experience (QUX)*, vol. 5, 2020.

[56] **Croci, Simone**, E. Zerman, and A. Smolic, "VIVA-Q: Omnidirectional video quality assessment based on Voronoi patches and visual attention." ISO/IEC JTC1/ SC29/AG2 M56165, 2021.

[57] **Simone Croci**, C. Ozcinar, E. Zerman, R. Dudek, S. Knorr, and A. Smolic, "Deep color mismatch correction in stereoscopic 3D images," in *IEEE International Conference on ImageProcessing (ICIP)*, (Anchorage, USA), 2021.

[58] M. Landy and J. A. Movshon, *The Plenoptic Function and the Elements of Early Vision*, pp. 3–20. MIT Press, 1991.

[59] H. Ishiguro, M. Yamamoto, and S. Tsuji, "Omni-directional stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 257–262, 1992.

[60] S. Peleg, M. Ben-Ezra, and Y. Pritch, "Omnistereo: Panoramic stereo imaging.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 279–290, 2001.

[61] R. Anderson, D. Gallup, J. T. Barron, J. Kontkanen, N. Snavely, C. Hernández, S. Agarwal, and S. M. Seitz, “Jump: Virtual reality video,” *ACM Trans. Graph.*, vol. 35, pp. 198:1–198:13, Nov. 2016.

[62] C. Schroers, J.-C. Bazin, and A. Sorkine-Hornung, “An omnistereoscopic video pipeline for capture and display of real-world VR,” *ACM Trans. Graph.*, vol. 37, pp. 37:1–37:13, Aug. 2018.

[63] Y. Ye, E. Alshina, and J. Boyce, “Algorithm descriptions of projection format conversion and video quality metrics in 360lib,” Tech. Rep. JVET-F1003, ISO/IEC JTC1/SC29/WG11/N16888, Hobart, AU, March 2017.

[64] D. Fenna, *Cartographic Science, A Compendium of Map Projections, with Derivations.* CRC Press, 2007.

[65] E. W. Grafarend and F. W. Krumm, *Map Projections, Cartographic Information Systems.* Springer, 2006.

[66] H. Ukida, N. Yamato, Y. Tanimoto, T. Sano, and H. Yamamoto, “Omni-directional 3D measurement by hyperbolic mirror cameras and pattern projection,” in *2008 IEEE Instrumentation and Measurement Technology Conference*, pp. 365–370, 2008.

[67] S. Chan, X. Zhou, C. Huang, S. Chen, and Y. Li, “An improved method for fisheye camera calibration and distortion correction,” in *2016 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 579–584, 2016.

[68] J.-W. Chen, C.-Y. Kao, and Y.-L. Lin, “Introduction to H.264 advanced video coding,” in *Asia and South Pacific Conference on Design Automation, 2006.*, pp. 6 pp.–, 2006.

[69] J.-R. Ohm and G. Sullivan, “Vision, applications and requirements for high efficiency video coding (HEVC),” Tech. Rep. MPEG2011/N11891, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, March 2011.

[70] I. 23009-1, ““Information technology — dynamic adaptive streaming over http (DASH) — part 1: Media presentation description and segment formats,” tech. rep., ISO/IEC JTC1/SC29/WG11, 2014.

[71] C. Zhou, Z. Li, J. Osgood, and Y. Liu, “On the effectiveness of offset projections for 360-degree video streaming,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, June 2018.

[72] C. Ozcinar, A. De Abreu, and A. Smolic, "Viewport-aware adaptive 360° video streaming using tiles for virtual reality," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2174–2178, 2017.

[73] J. M. Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V. K. M. Vadakital, and L. Yu, "MPEG immersive video coding standard," *Proceedings of the IEEE*, pp. 1–16, 2021.

[74] F. Pearson, *Map Projections: Theory and Applications*. CRC Press, 1990.

[75] F. Aurenhammer, "Voronoi Diagrams - A Survey of a Fundamental Data Structure," *ACM Comput. Surv.*, vol. 23, no. 3, pp. 345–405, 1991.

[76] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry, Algorithms and Applications*. Springer, 2008.

[77] S. Ghali, *Introduction to Geometric Computing*. Springer, 2008.

[78] H.-S. Na, C.-N. Lee, and O. Cheong, "Voronoi diagrams on the sphere," *Computational Geometry*, vol. 23, no. 2, pp. 183–194, 2002.

[79] P. Su and R. L. Scot Drysdale, "A comparison of sequential Delaunay triangulation algorithms," *Computational Geometry*, vol. 7, no. 5, pp. 361–385, 1997. 11th ACM Symposium on Computational Geometry.

[80] "CGAL 5.2.1 - 2D Triangulation." https://doc.cgal.org/latest/Triangulation_2/. Accessed: 2021-05-10.

[81] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 101, pp. 2058–2067, 09 2013.

[82] M. S. Gide and L. J. Karam, "Computational visual attention models," *Foundations and Trends® in Signal Processing*, vol. 10, no. 4, pp. 347–427, 2017.

[83] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience* (L. M. Vaina, ed.), pp. 115–141, Dordrecht: Springer Netherlands, 1987.

[84] M. Carrasco, "Visual attention: The past 25 years," *Vision Research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[85] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[86] L. Itti and A. Borji, "Computational models: Bottom-up and top-down aspects," *The Oxford Handbook of Attention*, pp. 1–20, 2015.

[87] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 679–700, 2021.

[88] 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 (PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[89] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–12, 2011.

[90] J. Tilke, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2106–2113, 2009.

[91] K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 914–921, 2011.

[92] J. Wang, A. Borji, C. C. Kuo, and L. Itti, "Learning a combined model of visual saliency for fixation prediction," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1566–1579, 2016.

[93] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," *ArXiv e-prints*, pp. 1–16, 2016.

[94] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 262–270, 2015.

[95] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," in *arXiv*, January 2017.

[96] E. Upenik, M. Rerábek, and T. Ebrahimi, “A testbed for subjective evaluation of omnidirectional visual content,” in *Proceedings of the Picture Coding Symposium (PCS)*, 2016.

[97] X. Corbillon, F. De Simone, and G. Simon, “360-degree video head movement dataset,” in *Proceedings of the 8th ACM Multimedia Systems Conference (MMSys)*, pp. 199–204, 2017.

[98] E. Upenik and T. Ebrahimi, “A simple method to obtain visual attention data in head mounted virtual reality,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, (Hong Kong), 2017.

[99] Y. Rai, J. Gutiérrez, and P. Le Callet, “A dataset of head and eye movements for 360 degree images,” in *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys’17, (New York, NY, USA), pp. 205–210, ACM, 2017.

[100] A. De Abreu, C. Ozcinar, and A. Smolic, “Look around you: Saliency maps for omnidirectional images in VR applications,” in *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2017.

[101] A. Bur, A. Tapus, N. Ouerhani, R. Siegwart, and H. Hügli, “Robot navigation by panoramic vision and attention guided features,” in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 695–698, 2006.

[102] I. Bogdanova, A. Bur, and H. Hugli, “Visual attention on the sphere,” *IEEE Transactions on Image Processing*, vol. 17, pp. 2000–2014, Nov. 2008.

[103] I. Bogdanova, A. Bur, H. Hügli, and P.-A. Farine, “Dynamic visual attention on the sphere,” *Computer Vision and Image Understanding*, vol. 114, no. 1, pp. 100 – 110, 2010.

[104] “Salient360!: Visual attention modeling for 360○ images grand challenge.” http://www.icme2017.org/grand-challenges/, 2017.

[105] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, “SalNet360: Saliency Maps for omni-directional images with CNN,” *Signal Processing: Image Communication*, 2018.

[106] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, “Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks,” in *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 01–04, 2018.

[107] 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," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1420–1429, 2018.

[108] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, "Gaze prediction in dynamic 360° immersive videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5333–5342, 2018.

[109] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks, "The zone of comfort: Predicting visual discomfort with stereo displays," *Journal of Vision*, vol. 11, no. 8, pp. 1–29, 2011.

[110] G. Kramida and A. Varshney, "Resolving the vergence-accommodation conflict in head mounted displays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 7, pp. 1–16, 2015.

[111] D. Khaustova, J. Fournier, E. Wyckens, and O. Le Meur, "An objective method for 3D quality prediction using visual annoyance and acceptability level," in *Proceedings of the IS&T/SPIE Electronic Imaging: Stereoscopic Displays and Applications XXVI*, vol. 9391, 2015.

[112] Q. Dong, T. Zhou, Z. Guo, and J. Xiao, "A stereo camera distortion detecting method for 3DTV video quality assessment," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4, 2013.

[113] S. A. Fezza, M.-C. Larabi, and K. M. Faraoun, "Stereoscopic image quality metric based on local entropy and binocular just noticeable difference," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2002–2006, 2014.

[114] F. Battisti, M. Carli, A. Stramacci, A. Boev, and A. Gotchev, "A perceptual quality metric for high-definition stereoscopic 3D video," in *Proceedings of the SPIE Image Processing: Algorithms and Systems XIII, 939916*, 2015.

[115] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 591–602, 2014.

[116] D. Vatolin and A. Bokov, "Sharpness mismatch and 6 other stereoscopic artifacts measured on 10 Chinese S3D movies," in *Proceedings of the IS&T/SPIE Electronic Imaging: Stereoscopic Displays and Applications XXVIII*, pp. 137–144, 2017.

[117] A. Bokov, D. Vatolin, A. Zachesov, A. Belous, and M. Erofeev, "Automatic detection of artifacts in converted S3D video," in *Proceedings of the IS&T/SPIE Electronic Imaging: Stereoscopic Displays and Applications XXV*, vol. 9011, pp. 901112–901114, 2014.

[118] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of Just Noticeable Blur (JNB)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.

[119] N. Narvekar and L. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience, QoMEx 2009*, 2009 International Workshop on Quality of Multimedia Experience, QoMEx 2009, pp. 87–91, Nov. 2009. 2009 International Workshop on Quality of Multimedia Experience, QoMEx 2009 ; Conference date: 29-07-2009 Through 31-07-2009.

[120] J. Chen, J. Zhou, J. Sun, and A. C. Bovik, "Binocular mismatch induced by luminance discrepancies on stereoscopic images," in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2014.

[121] W. Foerstner and B. P. Wrobel, *Photogrammetric Computer Vision, Statistics, Geometry, Orientation and Reconstruction.* Springer, 2016.

[122] P. N. Binh Do and Q. Chi Nguyen, "A review of stereo-photogrammetry method for 3-d reconstruction in computer vision," in *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 138–143, 2019.

[123] K. Konolige, "Small vision systems: Hardware and implementation," in *Robotics Research* (Y. Shirai and S. Hirose, eds.), (London), pp. 203–212, Springer London, 1998.

[124] Y. Song, R. Wilson, R. Edmondson, and N. Parsons, "Surface modelling of plants from stereo images," in *Sixth International Conference on 3-D Digital Imaging and Modeling (3DIM 2007)*, pp. 312–319, 2007.

[125] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 508–515 vol.2, 2001.

[126] Q. Yang, "A non-local cost aggregation method for stereo matching," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1402–1409, 2012.

[127] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 2, pp. 328–341, 2008.

[128] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 31–36, Sept. 2015.

[129] J. F. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 8, no. 6, pp. 679–698, 1986.

[130] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to JPEG2000," *Signal Processing: Image Communication*, vol. 19, pp. 163–172, 2004.

[131] O. Pele and M. Werman, "Fast and robust Earth Mover's Distances," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 460–467, Sept. 2009.

[132] D. De Silva, H. Kodikara Arachchi, E. Ekmekcioglu, A. Fernando, S. Dogan, A. Kondoz, and S. Savas, "Psycho-physical limits of interocular blur suppression and its application to asymmetric stereoscopic video delivery," in *2012 19th International Packet Video Workshop, PV 2012*, 05 2012.

[133] B. Julesz, "Cyclopean perception and neurophysiology," *Investigative Ophthahalmology & Visual Science*, vol. 11, no. 6, p. 540, 1972.

[134] C. Schor, L. Landsman, and P. Erickson, "Ocular dominance and the interocular suppression of blur in monovision," *American journal of optometry and physiological optics*, vol. 64, p. 723, Nov. 1987.

[135] M. J Collins and A. Goode, "Interocular blur suppression and monovision," *Acta ophthalmologica*, vol. 72, p. 376, Jul 1994.

[136] S. A. Fezza and M. Larabi, "Perceptually driven nonuniform asymmetric coding of stereoscopic 3D video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, pp. 2231–2245, Oct. 2017.

[137] T. Sakamoto, "Model for spherical aberration in a single radial gradient-rod lens.," *Applied Optics*, vol. 23, no. 11, p. 1707, 1984.

[138] M. M. Subedar and L. J. Karam, “3D blur discrimination,” *ACM Trans. Appl. Percept.*, vol. 13, pp. 12:1–12:13, Apr. 2016.

[139] A. T. Duchowski, B. Pelfrey, D. H. House, and R. Wang, “Measuring gaze depth with an eye tracker during stereoscopic display,” in *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, APGV ’11, (New York, NY, USA), p. 15–22, Association for Computing Machinery, 2011.

[140] J. Kollin and A. Hollander, “Re-engineering the wheatstone stereoscope,” *SPIE Newsroom*, 2007.

[141] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” tech. rep., International Telecommunication Union, COM 9-80-E, Geneva, Switzerland, 2000.

[142] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Computer Graphics and Applications*, vol. 21, pp. 34–41, Sept. 2001.

[143] W. Xu and J. Mulligan, “Performance evaluation of color correction approaches for automatic multi-view image and video stitching,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 263–270, 2010.

[144] C. Mouffranc and V. Nozick, “Colorimetric Correction for Stereoscopic Camera Arrays,” in *Computer Vision - ACCV 2012 Workshops*, pp. 206–217, 2013.

[145] Q. Wang, P. Yan, Y. Yuan, and X. Li, “Robust color correction in stereo vision,” in *18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium, September 11-14, 2011* (B. Macq and P. Schelkens, eds.), pp. 965–968, IEEE, 2011.

[146] M. Grogan and R. Dahyot, “Robust registration of Gaussian mixtures for colour transfer,” *CoRR*, vol. abs/1705.06091, 2017.

[147] M. Grogan and R. Dahyot, “L2 divergence for robust colour transfer,” *Computer Vision and Image Understanding*, vol. 181, pp. 39 – 49, 2019.

[148] Y. Fan, P. Liu, and Y. Niu, “Deep residual optimization for stereoscopic image color correction,” in *Parallel Architectures, Algorithms and Programming* (H. Shen and Y. Sang, eds.), (Singapore), pp. 147–158, Springer Singapore, 2020.

[149] R. Dudek, C. Cuenca-Hernández, and F. Quintana-Domínguez, “Stereoscopic rectification brought to practical: A method for real film production environments,” in *Computer Aided Systems Theory – EUROCAST 2017* (R. Moreno-Díaz, F. Pichler, and A. Quesada-Arencibia, eds.), (Cham), pp. 157–165, Springer International Publishing, 2018.

[150] S. Ye, S. P. Lu, and A. Munteanu, “Color correction for large-baseline multiview video,” *Elsevier Signal Process.: Image Commun.*, vol. 53, no. January, pp. 40–50, 2017.

[151] M. Xia, J. Yao, R. Xie, and M. Zhang, “Color Consistency Correction Based on Remapping Optimization for Image Stitching,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[152] F. Pitie, A. Kokaram, and R. Dahyot, “N-dimensional probability density function transfer and its application to color transfer,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1434–1439 Vol. 2, Oct. 2005.

[153] Y.-W. Tai, J. Jia, and C.-K. Tang, “Local color transfer via probabilistic segmentation by expectation-maximization,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 747–754 vol. 1, June 2005.

[154] F. Pitié and A. Kokaram, “The linear monge-kantorovitch linear colour mapping for example-based colour transfer,” in *Visual Media Production, 2007. IETCVMP. 4th European Conference on*, pp. 1–9, Nov. 2007.

[155] M. Grogan, R. Dahyot, and A. Smolic, “User interaction for image recolouring using L2,” in *14th European Conference on Visual Media Production (CVMP)*, CVMP 2017, (New York, NY, USA), pp. 6:1–6:10, ACM, 2017.

[156] N. Bonneel, G. Peyré, and M. Cuturi, “Wasserstein barycentric coordinates: Histogram regression using optimal transport,” *ACM Transactions on Graphics*, vol. 35, pp. 71:1–71:10, July 2016.

[157] B. Wang, Y. Yu, T.-T. Wong, C. Chen, and Y.-Q. Xu, “Data-driven image color theme enhancement,” *ACM Transactions on Graphics*, vol. 29, pp. 146:1–146:10, Dec. 2010.

[158] Y. Shih, S. Paris, F. Durand, and W. T. Freeman, “Data-driven hallucination of different times of day from a single outdoor photo,” *ACM Transactions on Graphics*, vol. 32, pp. 200:1–200:11, Nov. 2013.

[159] M. Grogan, M. Prasad, and R. Dahyot, “L2 registration for colour transfer,” in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1–5, Aug. 2015.

[160] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[161] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, pp. 533–536, 1986.

[162] F.-F. Li, “Convolutional neural networks for visual recognition.” http://cs231n.stanford.edu/. Accessed: 2021-05-01.

[163] D. H. Hubel, *Eye, brain, and vision*. Scientific American Library: Distributed by W.H. Freeman New York, 1988.

[164] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, Nov. 1998.

[165] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’12, (USA), pp. 1097–1105, Curran Associates Inc., 2012.

[166] “IMAGENET large scale visual recognition challenge.” http://www.image-net.org/challenges/LSVRC/.

[167] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[168] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *CoRR*, vol. abs/1409.4842, 2014.

[169] 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)*, pp. 770–778, June 2016.

[170] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 6154–6162, IEEE Computer Society, Jun. 2018.

[171] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[172] K. Zhang, L. Van Gool, and R. Timofte, “Deep unfolding network for image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3217–3226, 2020.

[173] Y. Hu, R. Song, and Y. Li, “Efficient coarse-to-fine patch match for large displacement optical flow,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5704–5712, June 2016.

[174] C. Liu, J. Yuen, and A. Torralba, “SIFT Flow: Dense correspondence across scenes and its applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[175] Y. Niu, H. Zhang, W. Guo, and R. Ji, “Image quality assessment for color correction based on color contrast similarity and color value difference,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 849–862, 2018.

[176] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *CVPR*, 2017.

[177] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, “Learning parallax attention for stereo image super-resolution,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[178] L. Wang, Y. Guo, Y. Wang, Z. Liang, Z. Lin, J. Yang, and W. An, “Parallax attention for unsupervised stereo correspondence learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[179] A. M. Andrew, “Artificial vision for mobile robots: Stereo vision and multisensory perception by nicholas ayache mit press, cambridge, mass., 1991, hard cover, 342 pp. (£40.50).,” *Robotica*, vol. 10, no. 5, p. 472–473, 1992.

[180] M. Pollefeys, R. Koch, and L. Van Gool, “A simple and efficient rectification method for general motion.,” in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 496–501, 01 1999.

[181] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, “A stereo attention module for stereo image super-resolution,” *IEEE Signal Processing Letters*, vol. 27, pp. 496–500, 2020.

[182] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12031–12038, Apr. 2020.

[183] W. Xie, J. Zhang, Z. Lu, M. Cao, and Y. Zhao, "Non-local nested residual attention network for stereo image super-resolution," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2643–2647, 2020.

[184] Y. Pang, J. Nie, J. Xie, J. Han, and X. Li, "BidNet: Binocular image dehazing without explicit disparity estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5930–5939, 2020.

[185] G. Wu, Y. Liu, L. Fang, and T. Chai, "Spatial-angular attention network for light field reconstruction," *CoRR*, 2020.

[186] Y. Nakano, "Stereo vision based single-shot 6D object pose estimation for bin-picking by a robot manipulator," *CoRR*, 2020.

[187] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 525–536, Curran Associates, Inc., 2018.

[188] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

[189] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *International Conference on Computer Vision Workshops*, pp. 3852–3857, Oct 2019.

[190] W. Bao, W. Wang, Y. Xu, Y. Guo, S. Hong, and X. Zhang, "InStereo2K: A large real dataset for stereo matching in indoor scenes," *Science China Information Sciences*, vol. 63, 11 2020.

[191] Y. J. Jung, H. Sohn, S. Lee, , and Y. M. Ro, "IVY Lab stereoscopic image database." `http://ivylabprev.kaist.ac.kr/demo/3DVCA/3DVCA.htm`, 2012.

[192] `https://www.flickr.com/`. Accessed: 2021-05-01.

[193] L. Guo, Z. Zha, S. Ravishankar, and B. Wen, “Self-convolution: A highly-efficient operator for non-local image restoration,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1860–1864, 2021.

[194] J. Watson, O. M. Aodha, D. Turmukhambetov, G. J. Brostow, and M. Firman, “Learning stereo from single images,” in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 722–740, Springer International Publishing, 2020.

[195] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[196] https://pytorch.org/. Accessed: 2021-05-01.

[197] G. Sharma and R. Bala, *Digital Color Imaging Handbook.* CRC Press, 2002.

[198] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[199] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, pp. 430–444, Feb 2006.

[200] “Draft overview of quality metrics and methodologies for immersive visual media (v2).” ISO/IEC JTC 1/SC 29/AG 5 N00013, January 2021.

[201] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai, “A Large-Scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison,” in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, Aug. 2018.

[202] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, “Perceptual quality assessment of omnidirectional images,” in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, May 2018.

[203] G. Luz, J. Ascenso, C. Brites, and F. Pereira, “Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment,” in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, Oct 2017.

[204] H. Lim, H. G. Kim, and Y. M. Ro, “VR IQA NET: deep virtual reality image quality assessment using adversarial learning,” *CoRR*, vol. abs/1804.03943, 2018.

[205] C. Li, M. Xu, S. Zhang, and P. L. Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *CoRR*, vol. abs/1905.00161, 2019.

[206] Y. Zhang, Y. Wang, F. Liu, Z. Liu, Y. Li, D. Yang, and Z. Chen, "Subjective panoramic video quality assessment database for coding applications," *IEEE Transactions on Broadcasting*, vol. 64, pp. 461–473, June 2018.

[207] A. Singla, S. Goring, A. Raake, B. Meixner, R. Koenen, and T. Buchholz, "Subjective quality evaluation of tile-based streaming for omnidirectional videos," in *10th ACM Multimedia Systems Conference (MMSys 2019)*, 2019.

[208] R. Schatz, A. Sackl, C. Timmerer, and B. Gardlo, "Towards subjective quality of experience assessment for omnidirectional video streaming," in *Proc. 9th Int. Conf. Qual. Multimedia Exp.(QoMEX)*, pp. 1–6, 2017.

[209] E. Upenik, M. Rerabek, and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[210] H. T. T. Tran, N. P. Ngoc, C. M. Bui, M. H. Pham, and T. C. Thang, "An evaluation of quality metrics for 360 videos," in *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 7–11, July 2017.

[211] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video multimethod assessment fusion (VMAF) on 360VR contents," *CoRR*, vol. abs/1901.06279, 2019.

[212] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2018.

[213] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric." Netflix Technology Blog https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, June 2016.

[214] N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller, "An evaluation of video quality assessment metrics for passive gaming video streaming," in *Proceedings of the 23rd Packet Video Workshop*, pp. 7–12, ACM, 12 June 2018.

[215] R. Rassool, “VMAF reproducibility: Validating a perceptual practical video quality metric,” in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–2, June 2017.

[216] C. G. Bampis, Z. Li, and A. C. Bovik, “Spatiotemporal feature integration and model fusion for full reference video quality assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.

[217] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, pp. 1398–1402 Vol.2, Nov. 2003.

[218] A. Abbas and B. Adsumilli, “AhG8: New GoPro test sequences for virtual reality video coding,” Tech. Rep. JVET-D0026, JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Oct. 2016.

[219] E. Asbun, H. He, H. Y., and Y. Ye, “AhG8: InterDigital test sequences for virtual reality video coding,” Tech. Rep. JVET-D0039, JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Oct. 2016.

[220] G. Bang, G. Lafruit, and M. Tanimoto, “Description of 360 3D video application exploration experiments on divergent multiview video,” Tech. Rep. MPEG2015/M16129, JTC1/SC29/WG11, ISO/IEC, Chengdu, China, Feb. 2016.

[221] “x265 HEVC Encoder / H.265 Video Codec.” http://x265.org/, Jan. 2018.

[222] “FFmpeg.” https://ffmpeg.org. Accessed: 2019-01-15.

[223] “HLS Authoring Specification for Apple Devices.” https://developer.apple.com, Jan. 2018.

[224] 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*, vol. 29, 12 2018.

[225] https://github.com/Archer-Tatsu/Evaluation_VR-onebar-vive. Accessed: 2019-01-15.

[226] A. Singla, S. Fremerey, W. Robitza, P. Lebreton, and A. Raake, “Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution,” in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Thematic Workshops ’17, (New York, NY, USA), pp. 511–519, ACM, 2017.

[227] A. Singla, W. Robitza, and A. Raake, "Comparison of subjective quality evaluation methods for omnidirectional videos with DSIS and modified ACR," in *IS&T Electronic Imaging, Human Vision and Electronic Imaging XXII*, International Society for Optics and Photonics, 2018.

[228] ITU-R, "Methodology for the subjective assessment of the quality of television pictures." ITU-R Recommendation BT.500-13, Jan. 2012.

[229] ITU-T, "Subjective video quality assessment methods for multimedia applications." ITU-T Recommendation P.910, Apr 2008.

[230] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, pp. 1427–1441, June 2010.

[231] "Video multi-method assessment fusion (VMAF)." https://github.com/Netflix/vmaf. Accessed: 2019-01-15.

[232] "Video quality measurement tool (VQMT)." https://mmspg.epfl.ch/vqmt. Accessed: 2019-01-15.

[233] "360Lib." https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/trunk. Accessed: 2019-01-15.

[234] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360° videos," in *Computer Vision – ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 504–520, Springer International Publishing, 2018.

[235] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal CNN for 360° video quality assessment," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10169–10178, 2019.