

Quality Assessment of Stereoscopic (3D) Multimedia

M.Tech Thesis

by

Saish Dilip Kajrolkar



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

May 2025

Quality Assessment of Stereoscopic (3D) Multimedia

A THESIS

*Submitted in partial fulfillment of the
requirements for the award of the degree
of*

Master of Technology

by

Saish Dilip Kajrolkar

2302102007



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

May 2025




INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Quality Assessment of Stereoscopic (3D) Multimedia** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF TECHNOLOGY** and submitted in the **DEPARTMENT OF ELECTRICAL ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2023 to July 2025 under the supervision of Dr. Balasubramanyam Appina, Indian Institute of Technology Indore, India and Dr. Nagendra Kumar, Indian Institute of Technology, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.


22/05/2025

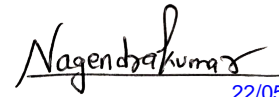
Signature of the student with date
Saish Dilip Kajrolkar

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.



22/05/2025

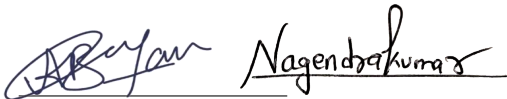
Signature of the Supervisor of
M.Tech. thesis #1 (with date)
Dr. Balasubramanyam Appina



22/05/2025

Signature of the Supervisor of
M.Tech. thesis #2 (with date)
Dr. Nagendra Kumar

Saish Dilip Kajrolkar has successfully given his/her M.Tech. Oral Examination held on **07 May 2025**.



Signature(s) of Supervisor(s) of M.Tech. thesis
Date: 22/05/2025



Convener, DPGC
Date: 22-05-2025

ACKNOWLEDGEMENT

I want to take this moment to sincerely thank everyone who has supported me along this journey, making it both joyful and rewarding. Their cooperation has been invaluable to my scientific progress and personal development, and I am deeply grateful for their assistance.

Above all, I extend my heartfelt gratitude to my supervisor, Dr. Balasubramanyam Appina, whose persistent support, wise counsel, and steadfast guidance have been crucial throughout this endeavor. His mentorship has provided me with the direction and confidence needed to navigate the challenges of my research. This work would not have been possible without his expertise and dedication, for which I am immensely appreciative. I would also like to express my sincere thanks to my co-supervisor, Dr. Nagendra Kumar, for his valuable suggestions, support, and encouragement throughout the course of this work. I am equally thankful to my lab mates, whose collaboration, feedback, and camaraderie have greatly enriched the quality of my research. Their helpful suggestions and considerate assistance have made this experience both memorable and intellectually rewarding. I am also deeply grateful to Dr. Vivek Kanhangad, Head of the Department of Electrical Engineering, for his invaluable guidance and supervision, which have significantly contributed to the success of this project. Special recognition goes to the Director of the Indian Institute of Technology Indore, Prof. Suhas Joshi, for fostering an excellent learning environment and providing me with the opportunity to develop my research skills. My experience has been greatly enhanced by the resources and support offered by this institution.

Last but not least, I want to express my sincere gratitude to my parents for their constant love, support, and encouragement. Throughout my life, their confidence in my skills has continuously given me strength and inspired me. Their steadfast presence has enabled me to pursue my aspirations, and I will always remain grateful for their sacrifices and dedication. To everyone who has contributed to this work in any capacity, from their guidance and advice to their encouragement and patience, thank you. Each individual has played a crucial role in making this journey a success. I am truly grateful for your support during this significant chapter of my academic life.

Saish Dilip Kajrolkar

Dedicated to My Family

List of Publications

Publications from Thesis

International Conferences

- C1. Saish Kajrolkar**, Venkatakiran Madana, and Balasubramanyam Appina. “An ‘Unsupervised’ 3D Image Quality Assessment using Spectral Decompositions of Scene Components”. In: *Proceedings of the 2025 National Conference on Communications (NCC)*. IEEE, New Delhi, India. Mar. 2025, pp. 1–6. DOI: 10.1109/NCC63735.2025.10983225.

ABSTRACT

Reliable quality assessment of stereoscopic 3D (S3D) images is essential for immersive media applications such as virtual reality, 3D video, and depth-based visualization systems. This thesis presents a no-reference image quality assessment (NR-IQA) framework that operates without access to ground truth reference images and captures human-like perception of stereoscopic content. The work begins with the generation of cyclopean images by simulating binocular fusion of stereo pairs. A novel tensor-based approach is used to compute pixel-level disparity maps, where gradients and chrominance-depth differences across the HSV channels are processed to extract dominant structural cues using eigen decomposition. These disparity maps are used to align stereo views and synthesize perceptually unified cyclopean images for quality analysis.

To capture perceptual features from these synthesized images, a multi-orientation steerable pyramid decomposition is applied to isolate spectral subbands of chrominance and luminance. These are evaluated using entropy and PIQE-based statistics to derive a final quality score that correlates well with subjective human ratings. Building upon this, a deep learning model based on a U-Net-inspired autoencoder is introduced to predict SSIM maps directly from cyclopean images. Unlike traditional SSIM computation which requires a reference image, the model is trained to infer structural distortion patterns in a completely blind setting. A large-scale dataset of 29,400 cyclopean images was constructed using IVY stereo image pairs with synthetically applied distortions, enabling effective training of the model on diverse degradation scenarios.

The proposed hybrid framework combines interpretable, unsupervised signal processing with data-driven learning to advance the state of blind S3D image quality assessment. Experimental results demonstrate that both the statistical model and deep learning architecture generalize well across distortion types and correlate strongly with human visual judgments. This thesis lays the foundation for efficient, real-time, and reference-free quality evaluation in stereoscopic imaging workflows.

Contents

List of Figures	v
Acronyms	vii
1 Introduction	1
2 Literature Survey	5
2.1 Stereoscopic Image Quality Assessment (S3D-IQA)	5
2.1.1 Early Approaches and Limitations	6
2.1.2 Incorporation of Disparity and Cyclopean Models	6
2.1.3 No-Reference and Reduced-Reference Methods	6
2.1.4 Comparative Analysis of IQA Metrics	7
2.1.5 Advancements in Deep Learning-Based S3D-IQA	7
2.2 Cyclopean Image Generation Techniques	8
2.3 No-Reference IQA Models for 2D and 3D	10
2.4 SSIM and Its Limitations in NR-IQA	12
2.5 Deep Learning Approaches in Blind IQA	14
3 Methodology	17
3.1 Cyclopean Image Generation Pipeline	17
3.1.1 HSV-Based Gradient Extraction	17
3.1.2 Tensor Formation and Eigenvalue Analysis	18
3.1.3 Disparity Map Computation	19
3.1.4 Disparity-Guided View Alignment	20

3.1.5	Cyclopean Image Synthesis	20
3.2	Unsupervised Quality Estimation using Spectral Features	22
3.2.1	Steerable Pyramid Decomposition	22
3.2.2	Chrominance and Luminance Quality Metrics	23
3.2.3	Final SDSC3D Score Computation	24
3.3	Dataset Creation for Deep Learning Training	25
3.3.1	Synthetic Distortion Application	26
3.3.2	Stereo Image Pair Formation	26
3.3.3	Disparity Map Generation	26
3.3.4	Cyclopean Image Synthesis	27
3.3.5	Computational Challenges and Validation	27
3.3.6	Dataset Utility and Future Steps	27
3.4	Deep Learning-Based SSIM Map Prediction	28
3.4.1	Dataset Expansion Using the IVY Dataset	28
3.4.2	Cross-Distortion Generalization	29
3.4.3	SSIM Ground Truth Generation	31
3.4.4	U-Net Based Autoencoder Architecture	32
3.4.5	Training Setup and Loss Function	33
4	Results	35
4.1	Dataset Details	35
4.1.1	LIVE Phase I Dataset	35
4.1.2	LIVE Phase II Dataset	37
4.1.3	Dataset Summary and Comparative Analysis	37
4.2	Quantitative Evaluation Metrics	39
4.2.1	Correlation-Based Metrics	40
4.2.2	Error-Based Metric	40
4.2.3	Use of Logistic Mapping	41
4.2.4	Evaluation Protocol	41
4.3	Comparative Evaluation with Benchmark Models	42

4.4	Symmetric vs Asymmetric Distortion Performance	46
4.4.1	Understanding Symmetric and Asymmetric Distortions	46
4.4.2	SDSC _{3D} Performance Breakdown	47
4.4.3	Orientation-Wise Analysis of Spectral Components	47
4.5	SSIM Map Prediction Evaluation	48
4.5.1	Motivation and Evaluation Goals	48
4.5.2	Cross-Distortion Generalization	49
4.5.3	Boundary Precision and Structural Integrity	49
4.5.4	Visual Perception and Interpretability	50
4.5.5	Limitations and Edge Cases	50
4.6	Training Curves and Optimization Performance	51
4.6.1	Training and Validation Loss Trends	51
4.6.2	Optimization Strategy	52
4.6.3	Early Stopping and Model Convergence	53
4.6.4	Comparison of SSIM Maps and Predicted Maps	53
5	Conclusion	57

List of Figures

2.1	General pipeline for cyclopean image generation from stereo pairs	8
2.2	CNN-based architecture for learning cyclopean fusion from stereo image pairs [1]	10
2.3	Generalized CNN-based pipeline for blind stereoscopic IQA [2, 3]	15
3.1	Algorithmic pipeline for cyclopean image generation	21
3.2	Flowchart illustrating unsupervised quality estimation using spectral features	23
3.3	Comparison of MSE and SSIM across different types of distortions. Despite having similar MSE values, SSIM varies significantly, reflecting perceptual differences more accurately. Source: https://www.researchgate.net	29
3.4	Example of SSIM ground truth map generation (Left: Reference Image, Center: Test Image, Right: SSIM Error Map)	31
3.5	General structure of the U-Net architecture (Conv: Convolution, BN: Batch Normalization, ReLU: Rectified Linear Unit)	32
4.1	Illustration of pristine and distorted stereoscopic images. Top: Reference and distorted left views. Middle: Corresponding disparity maps computed via structure tensor analysis. Bottom: Synthesized cyclopean images showing quality degradation due to distortions.	39
4.2	Scatter plots showing correlation between SDSC _{3D} scores and human-rated DMOS for two S3D datasets.	45
4.3	Training vs Validation Loss curve across 34 epochs.	52

4.4	Side-by-side visual comparison of SSIM maps (left column) and predicted SSIM maps (right column) across increasing JPEG2000 distortion severity levels (1–4).	54
-----	---	----

Acronyms

3D	Three-Dimensional
AE	Autoencoder
CNN	Convolutional Neural Network
DMOS	Difference Mean Opinion Score
DL	Deep Learning
FR	Full-Reference
GPU	Graphics Processing Unit
HSV	Hue Saturation Value
IQA	Image Quality Assessment
IVY	IIT Indore Vision and Imaging Dataset
LIVE	Laboratory for Image and Video Engineering
LR	Learning Rate
LCC	Linear Correlation Coefficient
MSE	Mean Squared Error
NR	No-Reference
PIQE	Perception-based Image Quality Evaluator
PSNR	Peak Signal-to-Noise Ratio
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
S3D	Stereoscopic 3D
SSIM	Structural Similarity Index Measure
SROCC	Spearman Rank Order Correlation Coefficient
U-Net	U-shaped Convolutional Network
VR	Virtual Reality

Chapter 1

Introduction

The evaluation of visual quality in stereoscopic 3D (S3D) images plays a critical role in modern multimedia systems, where user perception dictates overall experience. With the rapid growth of immersive content in domains such as virtual reality (VR), telemedicine, entertainment, automotive safety, and surveillance, ensuring high perceptual image quality has become increasingly important [4, 5]. Unlike traditional 2D media, S3D content introduces a binocular dimension that enhances realism through depth perception, but this comes at the cost of increased complexity in quality evaluation.

Classical image quality assessment (IQA) algorithms, such as the Structural Similarity Index (SSIM) [6], Peak Signal-to-Noise Ratio (PSNR), and others, have long relied on full-reference settings where a pristine copy of the image is accessible for comparison. However, practical scenarios such as real-time video streaming, bandwidth-limited environments, and wireless sensor networks often lack a reference image, prompting a shift toward no-reference (NR) IQA models [7, 8]. These NR approaches are even more challenging in the stereoscopic setting due to the added disparity cues and binocular interactions involved in human visual system (HVS) processing.

The HVS combines slightly offset images from the left and right eyes into a single

perceived view known as the cyclopean image [9, 10]. This fusion encapsulates chromatic, structural, and depth-related information, offering a perceptually holistic view of the scene. Computationally modeling this process allows us to approximate perceptual quality from a stereoscopic pair. However, effectively simulating the HVS demands accurate disparity estimation, chrominance handling, and luminance preservation tasks that are difficult to generalize using classical vision techniques.

Several earlier attempts constructed cyclopean images through simple averaging, energy-based fusion, or geometric warping [11, 12], but these overlooked perceptual subtleties such as misalignment artifacts and chromatic imbalance. Additionally, many hand-crafted feature based NR-IQA methods, such as BRISQUE [13] and DIIVINE [14], are insufficient when applied to complex stereoscopic scenes and tend to generalize poorly across distortion types [15].

In light of these challenges, there exists a clear and pressing need for a novel paradigm of NR-IQA that bridges computational efficiency with perceptual accuracy. The primary motivation of this thesis stems from this gap the lack of scalable, interpretable, and accurate no-reference methods that mimic HVS behavior in cyclopean image perception. Existing NR methods are not only computationally constrained, but also typically fall short in interpreting complex, mixed distortions that frequently appear in real world multimedia systems. Moreover, the absence of publicly available large scale datasets for stereoscopic image analysis makes it difficult to train and benchmark machine learning models effectively.

To address these limitations, this thesis introduces a data driven approach using deep learning. A UNet inspired convolutional autoencoder model is proposed to learn the mapping from distorted cyclopean images to SSIM like distortion maps in an unsupervised manner [16]. The central hypothesis is that a network can infer SSIM maps typically computed using reference images directly from distorted inputs, thereby functioning as a blind quality

estimator [17].

One major challenge was the limited availability of large stereoscopic datasets. The LIVE Phase II dataset, though well curated, contains only 400 stereo image pairs [10]. To address this, a larger dataset was synthesized using the IVY dataset and extended with a suite of distortions including Gaussian blur, white noise, JPEG compression, and fast fading. Disparity maps were estimated through a gradient-based eigenvector decomposition in the HSV domain, and cyclopean images were generated through disparity guided fusion. This process yielded a new dataset of over 29,000 image samples with varied content and distortion levels.

Disparity computation was carried out using GPU accelerated structure tensor analysis, where the eigenvectors of each pixel's local gradient tensor determined the dominant disparity direction. These disparity values informed precise alignment during cyclopean image formation, preserving key features critical to human perception [16, 18].

The proposed deep learning model employs an encoder-decoder architecture with skip connections to maintain spatial resolution. It was trained using mean squared error (MSE) loss between the predicted and reference SSIM maps. Training convergence and generalization were verified through cross-validation on held-out samples. The predicted SSIM maps closely resembled the ground truth, both visually and statistically, validating the feasibility of no-reference quality prediction [6].

Results demonstrated that the model could effectively localize perceptual distortions without any reference image, making it viable for real-time applications in streaming and display technologies. This represents a significant advancement in the domain of blind S3D IQA.

This thesis presents an end to end pipeline for perceptual quality prediction in stereoscopic 3D images. Contributions include a novel disparity aware cyclopean fusion method,

a large curated dataset for model training, and a deep learning model capable of generating SSIM equivalent quality maps without references. These innovations offer a new direction for perceptual optimization in stereoscopic content delivery and lay the foundation for future work in 3D video assessment, adaptive bit allocation, and real-time quality monitoring.

Chapter 2

Literature Survey

The exponential growth of 3D multimedia technologies has necessitated the development of robust methods for assessing the perceptual quality of stereoscopic content. Accurate evaluation of 3D image and video quality is vital for various applications including virtual reality, immersive entertainment, telemedicine, and remote surveillance. The literature review in this chapter provides a structured understanding of key developments in the domain, focusing on stereoscopic image quality assessment (S3D-IQA), cyclopean image generation, no-reference IQA models, and the role of SSIM and deep learning in advancing blind perceptual assessment. This review helps in identifying the existing gaps and motivates the formulation of a novel deep learning-based SSIM map prediction model using cyclopean inputs.

2.1 Stereoscopic Image Quality Assessment (S3D-IQA)

Stereoscopic Image Quality Assessment (S3D-IQA) focuses on evaluating the perceptual quality of stereoscopic images, which are designed to provide a three-dimensional viewing experience by presenting slightly different images to each eye. Unlike traditional 2D images, stereoscopic images introduce complexities such as binocular disparity, depth per-

ception, and potential visual discomfort due to inconsistencies between the left and right views [4, 12, 9].

2.1.1 Early Approaches and Limitations

Initial S3D-IQA methods extended 2D image quality metrics like PSNR and SSIM by applying them separately to each view and averaging the results. However, these approaches failed to account for the binocular fusion process of the human visual system (HVS), leading to discrepancies between objective assessments and subjective experiences [19, 11].

2.1.2 Incorporation of Disparity and Cyclopean Models

To address these limitations, more advanced models integrated disparity maps and simulated the cyclopean image the single mental image perceived by the HVS when combining the left and right views. Full-reference S3D-IQA models, such as those proposed by Chen et al. [20] and Akhter et al. [21], evaluated structural consistency, luminance fidelity, and disparity coherence. These models demonstrated improved correlation with human perception but required access to pristine reference images, limiting their applicability in scenarios like real-time streaming.

2.1.3 No-Reference and Reduced-Reference Methods

Recognizing the need for quality assessment without reference images, researchers developed reduced-reference and no-reference (NR) methods. These approaches leveraged statistical priors from natural scenes to estimate quality. For instance, Mittal et al. introduced the BRISQUE model, which utilized natural scene statistics for NR image quality assessment [17]. However, designing robust NR S3D-IQA algorithms remains challenging due to the complexity of modeling binocular perception without ground truth data.

2.1.4 Comparative Analysis of IQA Metrics

You et al. conducted a comprehensive evaluation of 11 popular 2D IQA metrics across various distortion types, including Gaussian blur, JPEG, JPEG2000 compression, and white noise. Their study revealed that the accuracy of these metrics deteriorated significantly when applied to stereoscopic content, primarily due to the absence of disparity information [22]. To mitigate this, they proposed integrating disparity maps with 2D IQMs using global and local fusion strategies. The results, summarized in Table 2.1, demonstrated that incorporating disparity improved the correlation of objective predictions with subjective scores.

Table 2.1: RMSE of IQMs on Stereoscopic Images (adapted from [22])

IQM	Blurring	JPEG	JPEG2000	Noise
PSNR	1.97	5.97	5.09	2.74
SSIM	3.00	7.70	8.91	2.43
MSSIM	1.91	4.94	4.97	2.51
VIF	1.84	4.39	6.45	3.41

2.1.5 Advancements in Deep Learning-Based S3D-IQA

Recent advancements have seen the application of deep learning techniques to S3D-IQA. For example, Shen et al. proposed a novel no-reference quality assessment metric that considers comprehensive 3D quality information, including the quality of the cyclopean image and 3D visual perceptual information like binocular fusion and rivalry [23]. Their method achieved high correlation with subjective assessments, demonstrating the potential of deep learning in this domain.

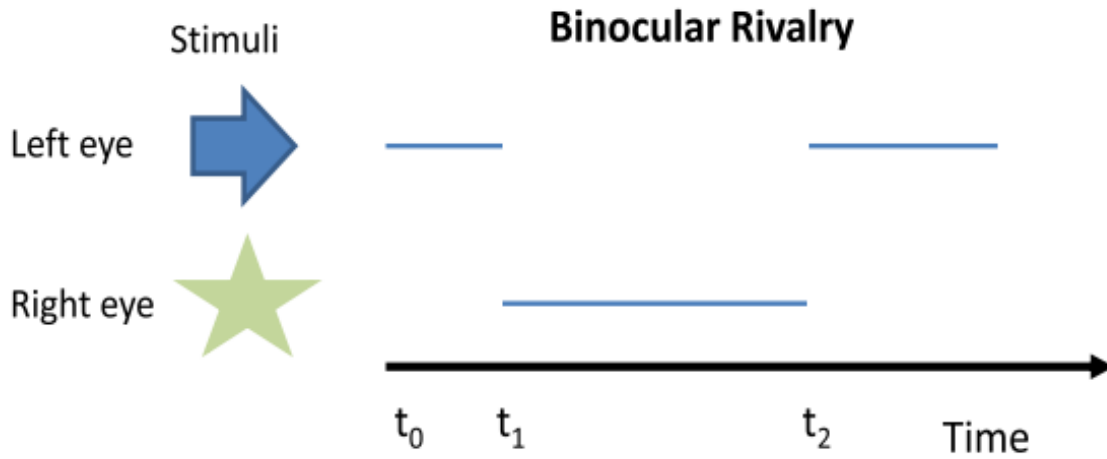


Figure 2.1: General pipeline for cyclopean image generation from stereo pairs [20]

2.2 Cyclopean Image Generation Techniques

Cyclopean image generation serves as a fundamental process in stereoscopic image quality assessment (S3D-IQA), as it simulates how the human visual system (HVS) perceives a unified three-dimensional scene from two slightly different left and right views. The fusion of these views results in what is known as the cyclopean image, capturing the binocular disparity and chrominance cues that are central to depth perception and perceptual realism.

Traditional cyclopean image generation techniques typically include basic methods such as pixel averaging, weighted fusion, or geometric warping of one view to the coordinate system of the other. These simplistic approaches, while computationally inexpensive, often fail to capture essential perceptual cues like edge misalignments, disparity inconsistencies, and depth-based fusion artifacts. Consequently, these methods do not accurately reflect the visual quality perceived by human observers.

As illustrated in Figure 2.1, the process begins with disparity map estimation, followed by warping of one view and fusion into a single perceptual representation. Such pipelines

more closely approximate human visual fusion compared to basic averaging.

To evaluate the effectiveness of these methods, Su et al. [11] compared several traditional techniques, highlighting their limitations in preserving structure and color fidelity under different distortion types. A comparative analysis of these methods is shown in Table 2.2.

Table 2.2: Comparison of Cyclopean Image Generation Methods [11]

Method	Handles Disparity	Color Preservation	Computational Cost
Pixel Averaging	No	Poor	Low
Warped Fusion	Yes	Moderate	Medium
DIBR-Based Warping	Yes	High	High
Bivariate Statistical Fusion	Yes	High	Medium

Modern developments have shifted toward deep learning-based approaches, which learn cyclopean representations directly from data. These methods leverage convolutional neural networks (CNNs) to fuse stereo images by capturing disparity, texture, and structural information in a data-driven manner. Figure 2.2 illustrates a typical CNN-based pipeline proposed in [1].

These approaches significantly outperform traditional methods in capturing complex perceptual distortions, occlusions, and misalignments. However, they require large annotated datasets and high computational resources for training. Despite these challenges, deep learning has opened new avenues in perceptually accurate, no-reference S3D image quality assessment.

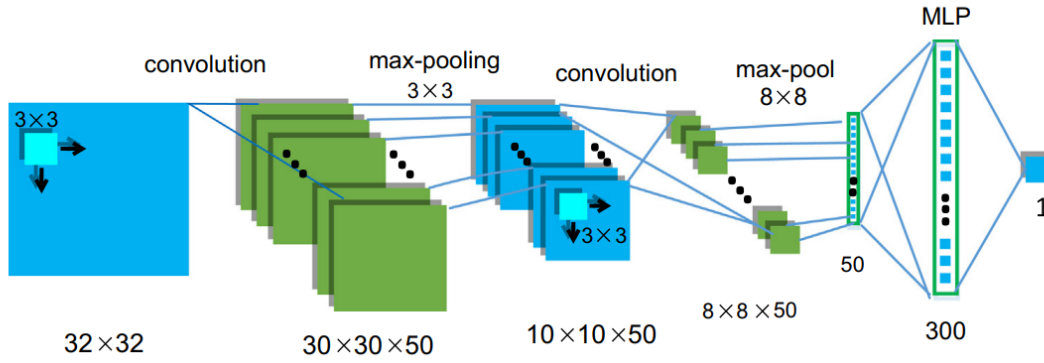


Figure 2.2: CNN-based architecture for learning cyclopean fusion from stereo image pairs [1]

Overall, cyclopean image generation remains a cornerstone in the accurate modeling of 3D perceptual experience. The transition from handcrafted rules to learning-based methods marks a pivotal evolution in this domain, with substantial implications for S3D display technologies, quality control, and immersive media applications.

2.3 No-Reference IQA Models for 2D and 3D

No-Reference (NR) Image Quality Assessment (IQA) methods are designed to evaluate visual quality without access to a pristine reference image. These models are especially valuable in practical applications such as streaming, broadcasting, or real-time rendering where ground truth is unavailable. The goal is to predict perceptual quality in a manner consistent with human visual judgments using only the distorted input.

2D No-Reference IQA Models

In 2D IQA, early NR approaches were built on Natural Scene Statistics (NSS), where deviations from statistical regularities in undistorted images are interpreted as perceptual degradation. Notable NSS-based methods include BRISQUE, NIQE, and DIIVINE,

each modeling different aspects of image statistics such as local luminance normalization, sharpness, and entropy. These techniques are computationally efficient and remain widely adopted due to their interpretability and simplicity.

Machine learning-based extensions, such as BLIINDS-II and FRIQUEE, leverage Support Vector Regression (SVR) or neural networks to improve prediction accuracy. These models use handcrafted features derived from wavelet transforms, DCT statistics, or spatial derivatives to map distortions to subjective quality scores.

3D No-Reference IQA Models

Extending NR-IQA to stereoscopic 3D introduces added complexity due to binocular fusion, disparity inconsistencies, and viewer discomfort. Initial efforts in NR S3D-IQA adapted 2D features to both left and right views independently and aggregated the scores. However, this approach fails to account for cyclopean perception and depth-based artifacts.

Table 2.3: Representative No-Reference IQA Models in 2D and 3D Domains

Model	Domain	Year	Key Features
BRISQUE [13]	2D	2012	Spatial NSS statistics
NIQE [17]	2D	2013	Unsupervised statistical baseline
FRIQUEE [24]	2D	2017	Feature-rich model using SVR
Appina et al. [10]	3D	2020	Disparity and chroma statistics
CoDIQE3D [25]	3D	2023	Joint depth-color statistical fusion

Recent models aim to capture binocular interactions by constructing statistical representations from fused or disparity-aware views. For instance, Appina et al. proposed a completely blind S3D-IQA model that integrates disparity entropy, chroma variance, and color-depth correlation into a no-reference framework [10]. These handcrafted features are evaluated using regression models trained on subjective quality scores.

Another line of work, such as CoDIQE3D [25], uses joint statistics of depth and chrominance to emulate perceptual fusion. These models extract features from the fused cyclopean domain and estimate perceptual quality using deep learning.

Current Challenges

Despite advancements, NR-IQA in 3D remains less mature than its 2D counterpart. The lack of large-scale annotated stereoscopic datasets, varying degrees of viewer discomfort, and the subjective nature of depth perception pose serious challenges. Additionally, the diversity of distortions including asymmetric artifacts across views requires more robust fusion and modeling strategies.

As deep learning continues to evolve, future models are likely to integrate spatiotemporal, binocular, and semantic cues into unified architectures, enabling real-time and perceptually aligned quality prediction.

2.4 SSIM and Its Limitations in NR-IQA

The Structural Similarity Index (SSIM), introduced by Wang et al. [26], marked a major shift in image quality assessment (IQA) by moving away from traditional pixel-wise error metrics such as PSNR, toward perceptually motivated criteria. SSIM operates by comparing local patterns of pixel intensities that have been normalized for luminance and contrast. It considers three perceptual components - luminance, contrast, and structure to quantify the similarity between a distorted image and its pristine reference.

Although SSIM has become one of the most widely adopted full-reference IQA metrics, it was never designed to operate in the absence of a reference image. Its strong dependence on a pristine counterpart renders it unsuitable for no-reference (NR) scenarios, where access to undistorted ground truth is either impractical or impossible such as in real-time

transmission, surveillance feeds, or legacy multimedia archives.

In the context of stereoscopic 3D (S3D) images, the shortcomings of SSIM become even more apparent. The metric is inherently monocular and fails to account for binocular visual processing, including disparity fusion and depth cues that significantly affect perceived image quality in stereo content. Attempts to extend SSIM to 3D content often involve computing SSIM scores for the left and right views independently and averaging them. However, such approaches ignore the cyclopean fusion process that takes place in the human visual system (HVS), leading to a poor correlation with human perception [12, 20].

Several studies have attempted to modify SSIM for better compatibility with NR-IQA tasks. For instance, metrics such as NIQE and BRISQUE have incorporated statistical properties of natural images inspired by the perceptual framework of SSIM [17, 13]. Nonetheless, these derivatives still do not fully replicate SSIM’s ability to localize structural distortions without access to reference data.

A fundamental challenge in NR adaptation of SSIM lies in its formulation it compares pixel-wise local statistics across two inputs. In the absence of a reference, the algorithm lacks a baseline to quantify what constitutes a deviation from visual “normalcy.” This makes direct extension of SSIM to NR paradigms ill-posed.

Recent research, including our own, explores the hypothesis that it may be possible to learn SSIM like perceptual maps from distorted images alone. By training deep neural networks (such as autoencoders) on large datasets of cyclopean images and their corresponding SSIM maps, the goal is to enable NR-IQA systems to implicitly infer perceptual distortions. This approach mimics the human visual system’s internal baseline and serves as a bridge between full-reference clarity and reference-free practicality.

Despite its historical significance and widespread utility, SSIM’s role in the evolution of NR-IQA is largely foundational. The metric provides a perceptually relevant structure but

requires augmentation through learning-based strategies or statistical modeling to function effectively in blind image quality assessment scenarios, especially in S3D contexts.

2.5 Deep Learning Approaches in Blind IQA

In recent years, deep learning has revolutionized the field of image quality assessment (IQA), particularly in no-reference (NR) or blind IQA, where the absence of a ground truth reference image makes perceptual modeling extremely challenging. Traditional handcrafted feature-based NR-IQA models rely on natural scene statistics (NSS), which, although effective to a degree, are limited in their ability to model complex distortions, semantic content, or higher-order perceptual interactions [13, 4].

Deep neural networks (DNNs), especially convolutional neural networks (CNNs), have emerged as powerful tools for learning hierarchical feature representations directly from data. These models are capable of capturing both low-level distortion cues and high-level contextual dependencies. Early CNN-based IQA models, such as those by Zhang et al. [1], trained networks to predict subjective quality scores using patch-based input. More recent architectures exploit attention mechanisms, multi-scale features, and perceptual embeddings to improve robustness across diverse distortion types.

For stereoscopic 3D (S3D) images, deep learning offers unique advantages by enabling joint learning of disparity, depth, and binocular fusion patterns from stereo pairs. Zhou et al. [2] proposed a self-similarity based CNN that extracted binocular features from stereo images, achieving strong correlation with human opinion scores. Similarly, Shi et al. [3] presented a multi-task CNN that learned both quality and distortion type simultaneously, using registered disparity-aware feature maps.

Figure 2.3 shows a typical pipeline of CNN-based blind IQA, where stereo image pairs are passed through a feature extractor, followed by regression layers to predict quality

scores.

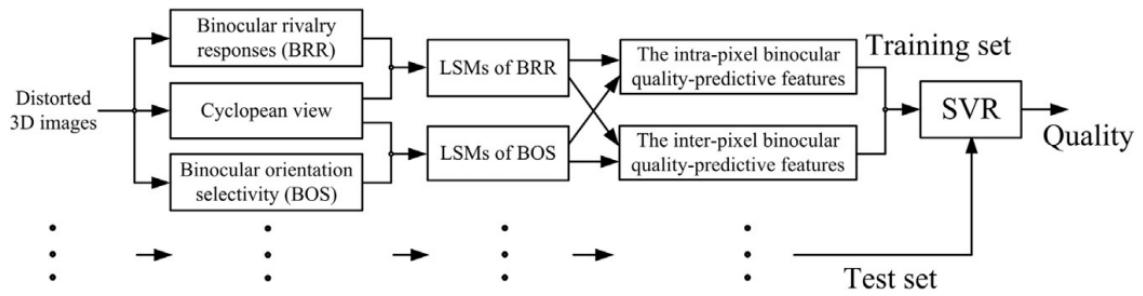


Figure 2.3: Generalized CNN-based pipeline for blind stereoscopic IQA [2, 3]

Despite their effectiveness, deep learning models for NR-IQA face several practical challenges:

To overcome data limitations, recent research has explored synthetic data generation, transfer learning from natural image domains, and weakly supervised learning approaches. For example, autoencoders and GANs (Generative Adversarial Networks) have been used to learn distortion manifolds in an unsupervised manner, enabling models to generalize without explicit labels [7, 25].

Overall, deep learning has not only improved prediction accuracy but also enabled the design of more perceptually aligned NR-IQA systems. As computational resources become more accessible and large-scale datasets become available, the fusion of deep learning with HVS-inspired priors will continue to redefine the benchmarks in blind stereoscopic quality assessment.

Chapter 3

Methodology

3.1 Cyclopean Image Generation Pipeline

This section provides an in-depth exploration of the cyclopean image generation pipeline, designed to replicate human binocular vision by integrating stereoscopic images. This process meticulously incorporates structural and depth information from left and right image pairs, creating a synthesized, perceptually unified image.

3.1.1 HSV-Based Gradient Extraction

Initially, RGB images from left (I_L) and right (I_R) views are transformed into the HSV (Hue, Saturation, and Value) color space. The HSV color space is selected for its superior alignment with human perceptual color interpretations, effectively segregating chromatic and luminance information. This segregation allows for precise extraction of visual characteristics relevant to human visual perception.

Following HSV conversion, horizontal (∇_x) and vertical (∇_y) gradients are computed for each HSV channel. These gradients provide essential local variation information, pivotal for accurate depth and structural analysis. Moreover, a structural gradient component (Z), representing the pixel-wise difference between corresponding pixels of the left and right

images, is calculated as:

$$Z_{i \in H,S,V} = I_{L,i \in H,S,V} - I_{R,i \in H,S,V} \quad (3.1)$$

This structural gradient is crucial in highlighting disparities between stereoscopic views.

3.1.2 Tensor Formation and Eigenvalue Analysis

The computed gradients form the basis for constructing an extended 3×3 tensor matrix (T) at each pixel location. This tensor encapsulates local variations in intensity and structure across HSV channels, defined mathematically as:

$$T = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{12} & T_{22} & T_{23} \\ T_{13} & T_{23} & T_{33} \end{bmatrix}, \quad (3.2)$$

Individual tensor components are explicitly detailed as:

- $T_{11} = \sum_{i \in \{H,S,V\}} (\nabla_{x_i})^2$ represents the sum of squared horizontal gradients for the H, S, and V channels.
- $T_{12} = \sum_{i \in \{H,S,V\}} (\nabla_{x_i} \cdot \nabla_{y_i})$ represents the product of horizontal and vertical gradients for each H,S and V channels that captures the correlation between horizontal and vertical variations.
- $T_{13} = \sum_{i \in \{H,S,V\}} (\nabla_{x_i} \cdot Z_i)$ represents the product of horizontal gradients of each H, S and V channels and structural-based variations to capture the directional relationship between horizontal variations and structural attributes.

- $T_{22} = \sum_{i \in \{H, S, V\}} (\nabla_{y_i})^2$ represents the sum of squared vertical gradients for the H, S, and V channels.
- $T_{23} = \sum_{i \in \{H, S, V\}} (\nabla_{y_i} \cdot Z_i)$ represents the product of vertical gradients of each H, S and V channels and structural-based variations to capture the directional relationship between vertical changes and structural attributes.
- $T_{33} = \sum_{i \in \{H, S, V\}} (Z_i)^2$ represents the sum of squared structural-based components of each H, S and V attributes.

Eigenvalue decomposition is then performed on each tensor T , yielding eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ and eigenvectors (v_1, v_2, v_3) . The largest eigenvalue and its corresponding eigenvector indicate the dominant orientation and magnitude of disparity.

3.1.3 Disparity Map Computation

Utilizing the eigen decomposition, a disparity map $d(a, b)$ is computed at each pixel as:

$$d(a, b) = |v_{\max}(a, b)| \cdot \lambda_{\max}(a, b) \quad (3.3)$$

where $\|v_{\max}(a, b)\|$ is the magnitude of the dominant eigenvector and $\lambda_{\max}(a, b)$ is the associated eigenvalue.

Using the disparity map, the HSV components of right view $I_{R_{i \in H, S, V}}$ are shifted horizontally by an amount proportional to the computed disparity $d(a, b)$ at each pixel location.

This map captures depth variations effectively, highlighting spatial differences essential for accurate stereoscopic alignment.

3.1.4 Disparity-Guided View Alignment

The computed disparity map guides the horizontal realignment of the right-view image I_R to match the spatial arrangement of the left-view image I_L . This alignment is implemented by horizontally shifting pixels proportionally to their calculated disparity:

$$I_{R_{\text{aligned}_{i \in H, S, V}}}(a, b) = I_{R_{i \in H, S, V}}((a - d(a, b)), b), \quad (3.4)$$

where $I_{R_{\text{aligned}_{i \in H, S, V}}}$ is aligned right image with respect to the left view $I_{L_{i \in H, S, V}}$.

This alignment step ensures accurate spatial correlation between stereoscopic pairs, crucial for synthesizing a coherent cyclopean image.

3.1.5 Cyclopean Image Synthesis

Finally, the cyclopean image (I_C) is synthesized by averaging the aligned right-view image and the original left-view image in the HSV color space:

$$I_C = \frac{I_{L_{i \in H, S, V}} + I_{R_{\text{aligned}_{i \in H, S, V}}}}{2} \quad (3.5)$$

where I_C is the estimated cyclopean image in the HSV format.

This averaging integrates depth cues and structural details from both views into a unified representation, effectively mimicking the perceptual fusion experienced by human binocular vision. Visual validation through examples clearly demonstrates the effectiveness and perceptual relevance of this pipeline.

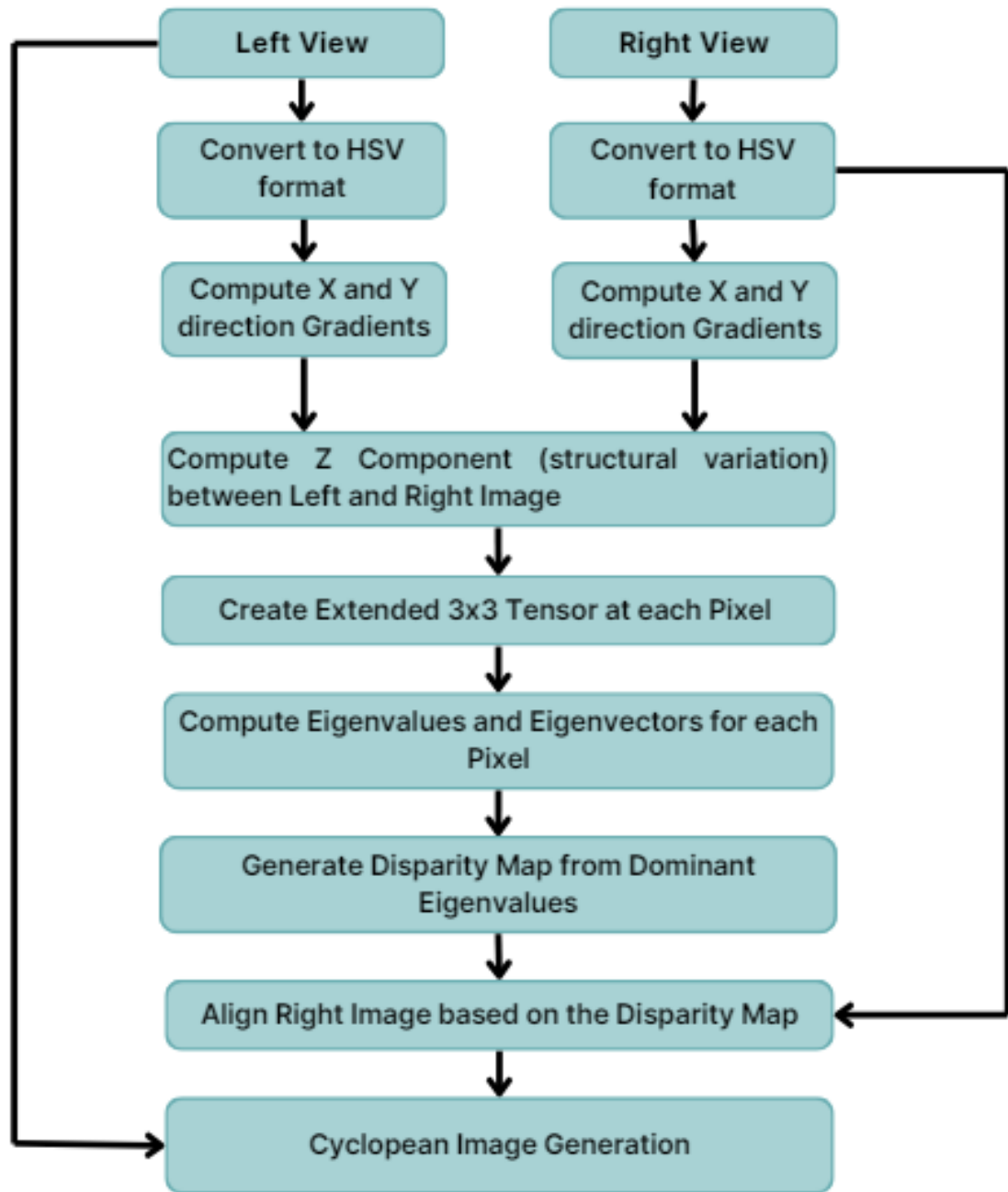


Figure 3.1: Algorithmic pipeline for cyclopean image generation

3.2 Unsupervised Quality Estimation using Spectral Features

Building upon the cyclopean image generation process described earlier, this section details an unsupervised approach for assessing stereoscopic image quality. This methodology exploits spectral decomposition techniques, particularly steerable pyramid decomposition, to capture perceptual attributes relevant to human visual processing, focusing on chrominance and luminance information.

3.2.1 Steerable Pyramid Decomposition

Steerable pyramid decomposition serves as the foundational spectral analysis method within the proposed pipeline. This multi-resolution analysis effectively captures detailed textural and structural information by decomposing the cyclopean image into multiple subbands at distinct orientations. Specifically, decomposition is performed across six orientations: 0° , 30° , 60° , 90° , 120° , and 150° .

The decomposition is mathematically represented as:

$$\{I_{CH_\theta}, I_{CS_\theta}, I_{CV_\theta}\} = F_\theta(I_C), \quad (3.6)$$

where F_θ denotes the steerable pyramid decomposition function at orientation θ , and I_{CH_θ} , I_{CS_θ} , I_{CV_θ} represent the subbands of the Hue, Saturation, and Value channels respectively.

This decomposition emulates the functionality of human cortical neurons, capturing orientation-specific perceptual features crucial for detailed image quality assessment.

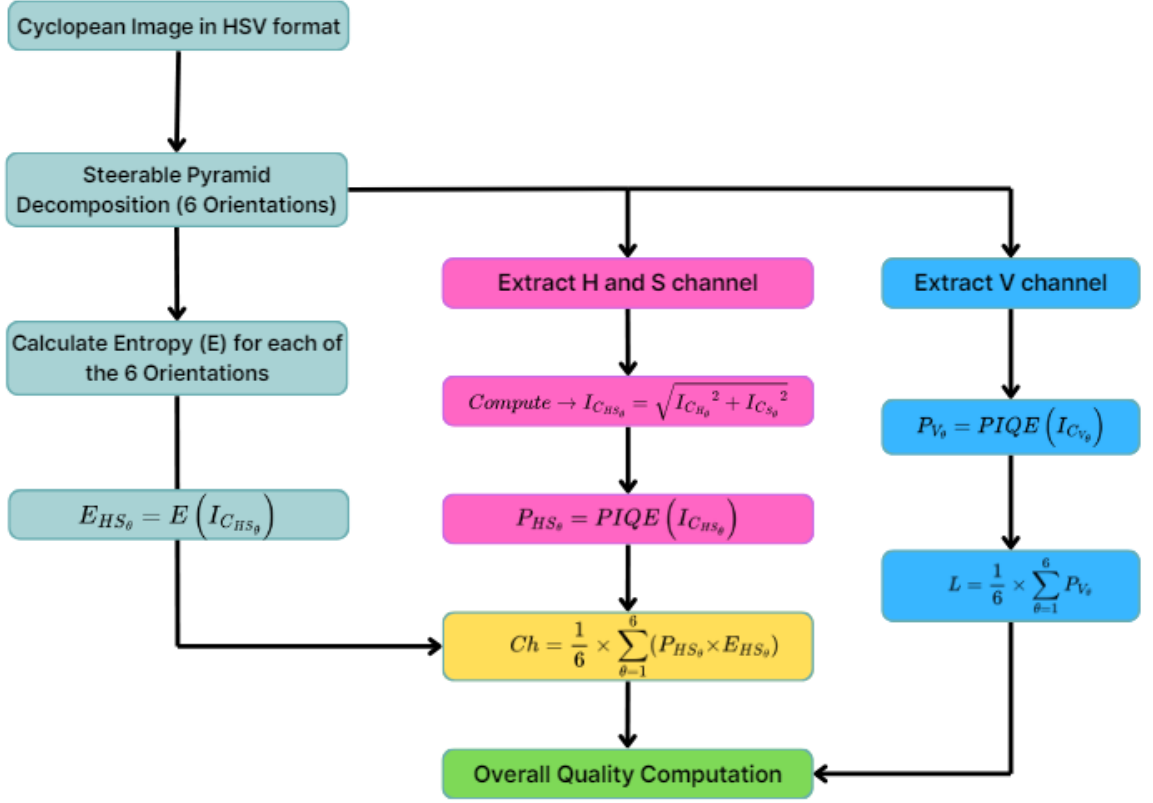


Figure 3.2: Flowchart illustrating unsupervised quality estimation using spectral features

3.2.2 Chrominance and Luminance Quality Metrics

To quantify the quality attributes of chrominance and luminance, spectral subbands from the decomposition are processed separately.

For chrominance quality estimation, the magnitude of Hue and Saturation subbands is combined as follows:

$$I_{CHS_\theta} = \sqrt{I_{CH_\theta}^2 + I_{CS_\theta}^2}, \quad (3.7)$$

where I_{CHS_θ} captures comprehensive chrominance strength at each orientation θ .

The perceptual quality features of the chrominance components are computed using two metrics: Perception-based Image Quality Evaluator (PIQE) scores and entropy measure-

ments:

$$P_{HS_\theta} = PIQE(I_{CHS_\theta}), \quad (3.8)$$

$$E_{HS_\theta} = E(I_{CHS_\theta}), \quad (3.9)$$

where P_{HS_θ} and E_{HS_θ} represent PIQE and entropy scores, respectively. These metrics quantify perceptual degradation and information content effectively.

The final aggregated chrominance quality metric (Ch) is calculated as the average product of PIQE and entropy scores across all orientations:

$$Ch = \frac{1}{6} \sum_{\theta=1}^6 (P_{HS_\theta} \times E_{HS_\theta}). \quad (3.10)$$

where Ch is the overall chrominance quality score of a given I_C . For luminance quality estimation, PIQE scores are independently calculated for each orientation-specific subband of the Value channel:

$$P_{V_\theta} = PIQE(I_{CV_\theta}). \quad (3.11)$$

where P_{V_θ} represents perceptual luminance quality at given θ . The overall luminance quality metric (L) is computed as the mean of PIQE scores across all orientations:

$$L = \frac{1}{6} \sum_{\theta=1}^6 P_{V_\theta}. \quad (3.12)$$

where L represents overall perceptual luminance quality of a given I_C image.

3.2.3 Final SDSC3D Score Computation

The combined chrominance (Ch) and luminance (L) quality scores offer complementary perceptual insights. Empirical analyses have shown that multiplying these two scores yields

a robust unified quality metric, referred to as $SDSC_{3D}$ (Spectral Decomposition Scene Components 3D Quality Score):

$$SDSC_{3D} = Ch \times L. \quad (3.13)$$

This final score effectively integrates both color and brightness-based perceptual qualities, aligning closely with subjective human evaluations.

The efficacy of the $SDSC_{3D}$ score has been validated across various distortion types, demonstrating its consistent performance and robustness in capturing perceptual image quality nuances in stereoscopic content.

3.3 Dataset Creation for Deep Learning Training

Creating a robust and comprehensive dataset is a crucial step in training deep learning models, particularly for stereoscopic image quality assessment tasks. Initially, the LIVE Phase 2 dataset was employed, which consisted of merely 400 cyclopean images. Recognizing the inadequacy of this dataset size for effective deep learning training, we aimed to substantially expand it. The main objective behind expanding the dataset was to enhance the diversity and quantity of training samples, thereby enabling more effective learning of the perceptual quality characteristics specific to stereoscopic images.

To achieve this, we utilized the IVY dataset, which originally comprises 240 images, containing both left and right stereoscopic views. This dataset was systematically extended through a rigorous process of synthetic distortion application and cyclopean image generation. The objective was to replicate and expand the types and levels of distortions similar to those observed in the LIVE Phase 2 dataset.

3.3.1 Synthetic Distortion Application

Each image in the IVY dataset was subjected to five different distortion types, consistent with those found in the LIVE Phase 2 dataset: White Noise, JPEG2000 (JP2K), JPEG compression, Gaussian Blur, and Fast Fading. To ensure comprehensive coverage, each distortion was applied across seven distinct, equidistant levels, ranging from low to high intensity. Consequently, each original IVY image was expanded to 35 distorted variants, yielding a total of 8,400 distorted images.

3.3.2 Stereo Image Pair Formation

The distorted images were systematically organized to maintain stereo pairing integrity. Each distorted left-view image was paired precisely with its corresponding distorted right-view counterpart under identical distortion conditions. This pairing was crucial to maintain accurate stereoscopic relationships, ensuring depth coherence and facilitating precise disparity map computation. Consequently, we obtained 4,200 right-view images and 4,200 left-view images, forming coherent stereoscopic pairs.

3.3.3 Disparity Map Generation

Accurate disparity maps are indispensable in synthesizing cyclopean images and crucial for reliable quality estimation. The disparity maps represent pixel-wise depth variations between stereo pairs and facilitate proper alignment during cyclopean synthesis. Disparity maps were computed using eigenvalue decomposition of extended structure tensors, capturing intricate variations in depth and structure between stereo images. A total of 29,400 disparity maps were generated, providing a comprehensive basis for the subsequent cyclopean image synthesis step.

3.3.4 Cyclopean Image Synthesis

Utilizing the computed disparity maps, cyclopean images were synthesized through disparity-guided alignment and averaging of corresponding stereo image pairs. This resulted in a dataset consisting of 29,400 cyclopean images, amounting to approximately 94.2 GB. The synthesized dataset exhibits considerable diversity in terms of distortion types, levels, and image content, thereby significantly enhancing the depth and breadth of data available for deep learning training.

3.3.5 Computational Challenges and Validation

The dataset creation involved substantial computational demands, with approximately 11 days required to process all images on a CPU-based setup. This substantial computational overhead posed significant logistical challenges. Additionally, ensuring a uniform quality distribution across various distortions and intensity levels was a critical consideration, requiring thorough validation checks to guarantee dataset integrity and reliability.

3.3.6 Dataset Utility and Future Steps

The resulting extensive dataset serves as an invaluable resource for training robust deep learning models. Subsequent steps involve employing this dataset to train classification and regression models capable of predicting distortion types and precise quality scores, respectively. Various deep learning architectures will be explored to optimize performance.

This systematic dataset creation approach lays a strong foundation for the subsequent phases of deep learning model training, ensuring comprehensive coverage of the quality-related nuances pertinent to stereoscopic image assessment tasks.

IVY Dataset	Apply Distortions	Apply different distortions at different levels	Make combinations of left and right view with each other
contains 240 images including left and right views	5 distortions namely White Noise, JP2K, JPEG, Gaussian Blur, Fast Fading are applied onto each image of the IVY Dataset	7 different equidistant levels applied under each distortion going from low to highest compression. Total: $240 \text{ images} \times 5 \text{ distortions} \times 7 \text{ levels} = 8,400$ images.	The dataset currently contains 4,200 right-view images and 4,200 left-view images. Each left-view image is paired with its corresponding right-view image to form stereoscopic image pairs under same distortion type.

Table 3.1: Sequential processing of the IVY Dataset for stereoscopic image quality evaluation

3.4 Deep Learning-Based SSIM Map Prediction

The previous sections outlined the processes involved in generating the cyclopean images and dataset creation for training deep learning architectures. This section presents a deep learning-based framework for predicting Structural Similarity Index Measure (SSIM) maps directly from distorted cyclopean images. The objective is to achieve high-fidelity perceptual quality assessment without relying on pristine reference images at inference time.

3.4.1 Dataset Expansion Using the IVY Dataset

To effectively train a deep learning model, a diverse and extensive dataset is essential. Initially, the LIVE Phase II dataset provided limited data, containing only around 400 cyclopean images. Therefore, this dataset was significantly expanded using the IVY dataset. The

expansion involved generating additional stereo pairs by applying similar distortion patterns as found in the LIVE dataset, such as Gaussian noise, JPEG compression, JPEG2000 compression, Gaussian blur, and Fast Fading. Consequently, the extended dataset comprises approximately 29,400 images, offering a substantial improvement in diversity, content variety, and distortion complexity for robust training.

3.4.2 Cross-Distortion Generalization

To evaluate the generalization capability of the proposed model across diverse types of distortions, we analyze how SSIM and MSE behave when applied to perceptually different test images, as shown in Figure 3.3.

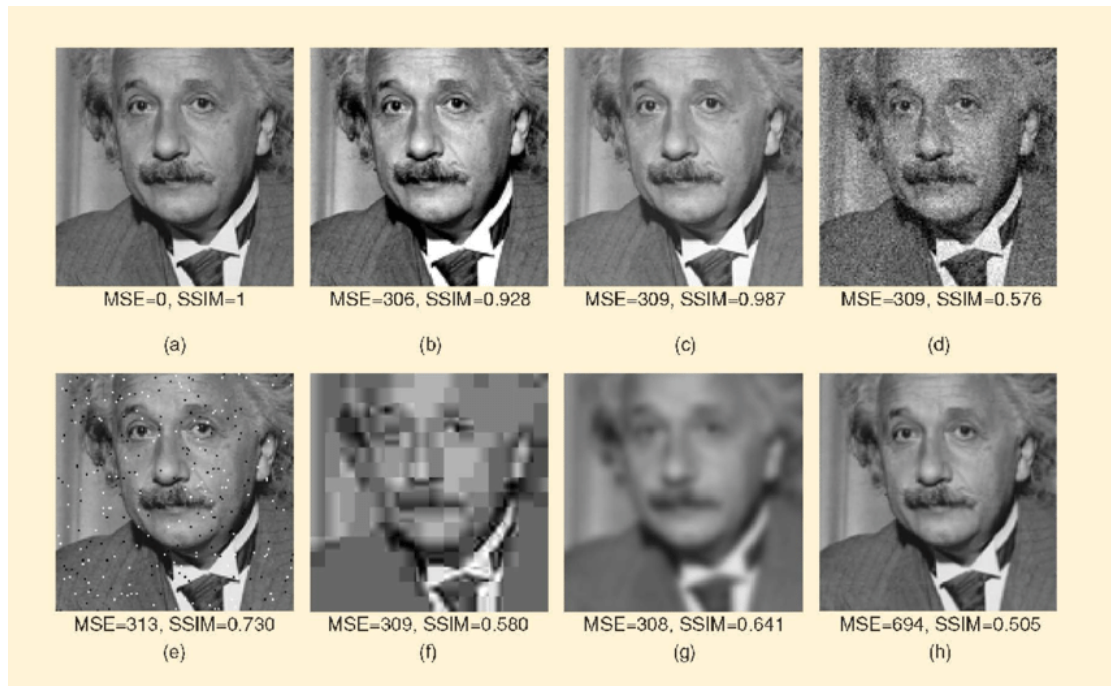


Figure 3.3: Comparison of MSE and SSIM across different types of distortions. Despite having similar MSE values, SSIM varies significantly, reflecting perceptual differences more accurately.

Source: <https://www.researchgate.net>

This figure juxtaposes eight versions of the same image, each affected by a unique distortion type ranging from Gaussian noise and blur to compression artifacts and structural

degradation. The Mean Squared Error (MSE) and SSIM values are annotated below each variant.

From the results, several critical observations emerge:

- **Perceptual Sensitivity of SSIM:** Images (c) and (f) have almost identical MSE values (309), yet vastly different SSIM scores (0.987 vs. 0.580). This reveals that MSE, as a pixel-wise metric, does not align well with human perception. In contrast, SSIM accurately reflects the significant quality drop in image (f), caused by strong block artifacts.
- **Distortion-Type Discrimination:** Blurred images like (g) exhibit relatively low SSIM despite having a comparable MSE to images (b) and (c). This suggests that SSIM captures the structural loss of fine details more effectively. Our model, trained on cyclopean views, mimics this behavior by assigning low predicted SSIM values to perceptually degraded but numerically similar distortions.
- **Structural Awareness:** In images with localized artifacts, such as salt-and-pepper noise in (e), SSIM reduction is moderate (0.730), which matches subjective perception. Such insights reinforce the importance of structural modeling, which is at the core of our network’s feature extraction.
- **MSE’s Shortcomings:** Images (d), (e), and (f) all have MSE values around 309–313, yet their perceived quality differs drastically. SSIM addresses this discrepancy by assigning scores that correlate more closely with subjective evaluation.

This comparative analysis validates the motivation behind adopting SSIM-based distortion maps for training. By modeling SSIM-like outputs, our network internalizes perceptual sensitivity, especially to local structures, edges, and texture loss across multiple distortion types.

Hence, the use of SSIM in our unsupervised training framework enables the model to not just estimate distortion severity but also discern its perceptual characteristics. This makes it particularly suitable for blind quality assessment in scenarios involving heterogeneous content and unseen distortions.

3.4.3 SSIM Ground Truth Generation

To train our model effectively, ground truth SSIM maps were generated using pristine and distorted cyclopean images. SSIM evaluates structural degradation between reference and distorted images at the pixel level. Figure 3.4 demonstrates an example of SSIM map generation. The reference and test images are compared pixel-by-pixel to compute a spatially localized distortion map highlighting regions of significant structural discrepancies.



Figure 3.4: Example of SSIM ground truth map generation (Left: Reference Image, Center: Test Image, Right: SSIM Error Map)

Source: <https://research.nvidia.com>

3.4.4 U-Net Based Autoencoder Architecture

To predict SSIM maps from distorted images, a deep convolutional neural network inspired by U-Net architecture was employed. U-Net effectively captures both global context and fine-grained spatial details, which are essential for accurately modeling local structural distortions. Figure 3.5 depicts the general architecture of the U-Net used.

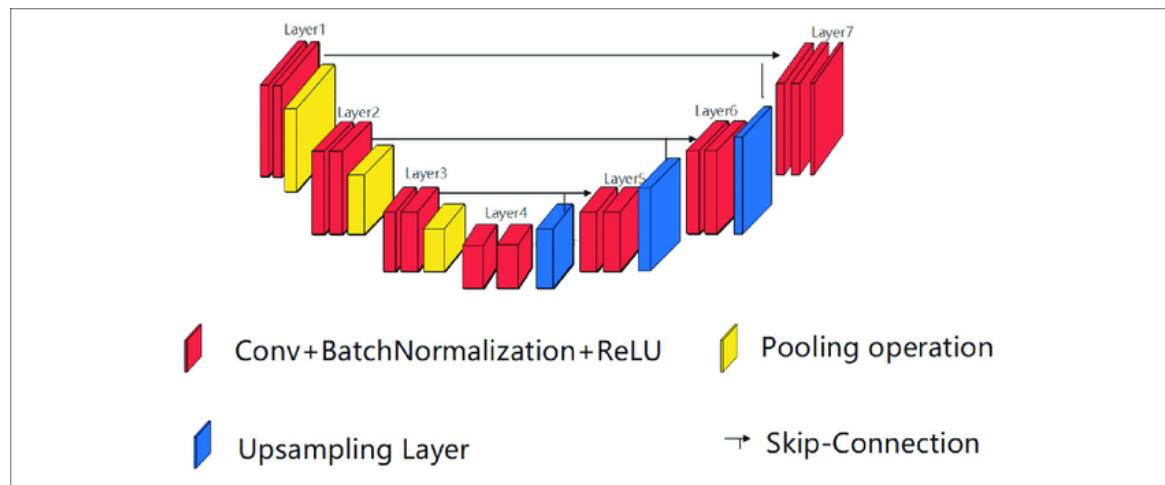


Figure 3.5: General structure of the U-Net architecture (Conv: Convolution, BN: Batch Normalization, ReLU: Rectified Linear Unit)

Source: https://www.researchgate.net/figure/Structure-diagram-of-UNet_fig1_359148875

The detailed architectures of the encoder and decoder components are presented below in tabular format.

Encoder architecture details:

Layer Name	Operation	Input Channels	Output Channels	Output Size
enc1	DoubleConv	3	32	512×512
pool1	MaxPool2d	-	-	256×256
enc2	DoubleConv	32	64	256×256
pool2	MaxPool2d	-	-	128×128
enc3	DoubleConv	64	128	128×128
pool3	MaxPool2d	-	-	64×64
enc4	DoubleConv	128	256	64×64
pool4	MaxPool2d	-	-	32×32
enc5	DoubleConv	256	512	32×32

Table 3.2: Encoder part of the U-Net Architecture

Decoder architecture details:

Table 3.3: Decoder Architecture

Layer Name	Operation	Input Channels	Output Channels	Output Size	Skip Connection
up4	ConvTranspose2d	512	256	64×64	Yes (from enc4)
dec4	DoubleConv	$256 + 256 = 512$	256	64×64	-
up3	ConvTranspose2d	256	128	128×128	Yes (from enc3)
dec3	DoubleConv	$128 + 128 = 256$	128	128×128	-
up2	ConvTranspose2d	128	64	256×256	Yes (from enc2)
dec2	DoubleConv	$64 + 64 = 128$	64	256×256	-
up1	ConvTranspose2d	64	32	512×512	Yes (from enc1)
dec1	DoubleConv	$32 + 32 = 64$	32	512×512	-
outc	Conv2d (1×1) + Sigmoid	32	1	512×512	Final output

3.4.5 Training Setup and Loss Function

The training of the U-Net model was conducted on GPU infrastructure to leverage computational efficiency. The training setup included batch normalization layers to stabilize and accelerate training. Rectified Linear Units (ReLU) activation functions were applied to introduce non-linear transformations.

A Mean Squared Error (MSE) loss function was utilized, defined mathematically as:

$$L_{MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3.14)$$

where y_i is the ground truth SSIM map and \hat{y}_i is the predicted SSIM map. This loss measures pixel-wise differences, ensuring the model learns fine-grained, spatially accurate representations.

The training setup included:

- **Batch Size:** 8
- **Epochs:** 150
- **Early Stopping:** Implemented after 10 epochs without validation loss improvement
- **Optimizer:** Adam optimizer with an initial learning rate of 1e-3
- **Learning Rate Scheduler:** ReduceLROnPlateau strategy, halving the learning rate after three epochs without improvement
- **Hardware:** Training conducted on a Linux-based local GPU environment configured with PyTorch

The training was closely monitored for improvements in validation loss, adopting the best model configurations iteratively. Early stopping was utilized effectively to prevent overfitting, ensuring optimal model performance.

The training process included careful monitoring of validation loss to avoid overfitting and to achieve optimal generalization performance.

Chapter 4

Results

4.1 Dataset Details

The evaluation of stereoscopic image quality assessment (S3D-IQA) methodologies necessitates robust and comprehensive datasets that accurately reflect real-world perceptual quality. To ensure the validity and reliability of our approach, we have utilized two benchmark datasets: LIVE Phase I and LIVE Phase II. These datasets are widely acknowledged within the image quality assessment community for their rigorously annotated distortion profiles and comprehensive coverage of common image degradations encountered in practical scenarios.

4.1.1 LIVE Phase I Dataset

The LIVE Phase I dataset constitutes a foundational repository developed explicitly for assessing the quality of stereoscopic content. This dataset comprises 20 original (pristine) stereo image pairs alongside 365 distorted stereo images. Each pristine image has been subjected to a range of artificially induced distortions, meticulously designed to simulate real-world impairments that commonly arise in stereoscopic image transmission and storage.

Specifically, the dataset includes the following distortion categories:

- **White Gaussian Noise (WN):** This distortion represents random pixel-level variations, often introduced due to sensor noise or transmission errors.
- **JPEG2000 Compression (JP2K):** JP2K artifacts arise from lossy compression schemes, characterized by wavelet-based compression blocks leading to noticeable visual impairments at higher compression ratios.
- **JPEG Compression (JPEG):** Block-based compression artifacts typical of JPEG compression, commonly observed in web-based image storage and sharing platforms.
- **Gaussian Blur (BLUR):** Simulates out-of-focus imaging conditions or poor camera optics resulting in blurry visual features.
- **Fast Fading (FF):** Reflects distortions similar to those observed during wireless signal transmission, manifesting as patchy visual degradation.

Each distortion type is applied at multiple severity levels, creating a graded scale of perceptual impairment, from mild to severe distortions. This gradation facilitates a nuanced understanding of how image quality algorithms perform across a spectrum of quality impairments.

Additionally, subjective evaluations have been conducted on this dataset to generate Difference Mean Opinion Scores (DMOS). DMOS scores represent human perceptual judgments of image quality, providing essential ground-truth data for supervised or semi-supervised training and rigorous algorithmic evaluation. The availability of these subjective quality scores significantly enhances the relevance and applicability of this dataset for evaluating quality assessment models.

4.1.2 LIVE Phase II Dataset

Building upon the robustness of the Phase I dataset, the LIVE Phase II dataset introduces further complexity, featuring an extensive variety of distortion scenarios including both symmetric and asymmetric distortions. This dataset includes 8 original stereo pairs and, similar to Phase I, comprises 365 distorted images. The notable difference in Phase II is the deliberate inclusion of asymmetric distortions, where the left and right views exhibit varying levels or types of impairments. This scenario closely mirrors real-world applications such as stereoscopic streaming services, where bandwidth constraints or transmission issues can lead to discrepancies between the stereo pair.

The distortion categories included in LIVE Phase II are identical to Phase I but with an emphasis on diverse perceptual complexity introduced by asymmetric distortion scenarios. The presence of asymmetric distortions introduces additional challenges for image quality assessment algorithms, which must accurately model perceptual fusion processes inherent in binocular vision.

DMOS scores are also provided for LIVE Phase II, enabling robust supervised evaluations and comparative analysis between different image quality assessment approaches. These subjective scores facilitate the precise benchmarking of algorithmic performance against human visual perception, making this dataset particularly valuable for validating the perceptual alignment of quality prediction models.

4.1.3 Dataset Summary and Comparative Analysis

To facilitate a clear and concise comparison, Table 4.1 summarizes key aspects of both LIVE Phase I and Phase II datasets, highlighting their similarities, distinctions, and overall utility for developing and validating stereoscopic image quality assessment algorithms.

Table 4.1: Summary of LIVE Phase I and Phase II Datasets

Attribute	LIVE Phase I	LIVE Phase II
Pristine Image Pairs	20	8
Distorted Images	365	365
Types of Distortions	5 (WN, JP2K, JPEG, BLUR, FF)	5 (WN, JP2K, JPEG, BLUR, FF)
Distortion Symmetry	Symmetric only	Symmetric and Asymmetric
Distortion Levels	Multiple levels	Multiple levels
DMOS Availability	Yes	Yes

These datasets collectively offer a comprehensive resource, enabling thorough evaluation and refinement of stereoscopic IQA algorithms, particularly emphasizing the challenges posed by asymmetric distortions in LIVE Phase II. As such, they provide an essential foundation for the supervised evaluation and validation of the proposed methodologies in this thesis.

In conclusion, the extensive and well-documented LIVE datasets, featuring robust subjective DMOS annotations and diverse distortion scenarios, provide a critical infrastructure for rigorous validation and comparative assessment of the proposed stereoscopic IQA model.

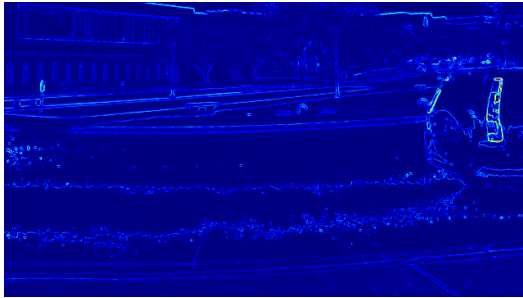
To illustrate the impact of distortions on the fused perceptual output, Figure 4.1 presents a comparative visualization of pristine and distorted scenes. The reference and distorted left images (top row) are taken from the LIVE Phase II dataset. Their corresponding disparity maps (middle row) exhibit noticeable structural differences due to distortions such as noise or compression artifacts. These disparities ultimately affect the final cyclopean image (bottom row), where perceptual quality visibly degrades between the reference and distorted outputs.



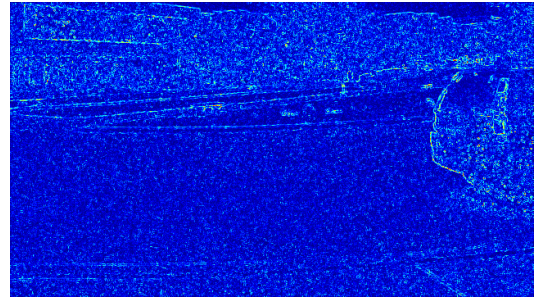
(a) Reference left scene.



(b) Distorted left scene.



(c) Reference disparity map.



(d) Distorted disparity map.

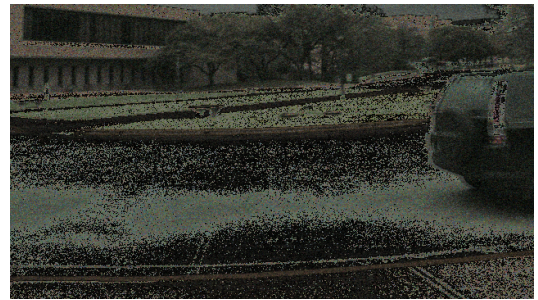
(e) Reference I_C view.(f) Distorted I_C view.

Figure 4.1: Illustration of pristine and distorted stereoscopic images. Top: Reference and distorted left views. Middle: Corresponding disparity maps computed via structure tensor analysis. Bottom: Synthesized cyclopean images showing quality degradation due to distortions.

This visual validation confirms that the cyclopean image is an appropriate proxy for evaluating depth-integrated image quality in stereoscopic datasets.

4.2 Quantitative Evaluation Metrics

To comprehensively assess the performance of the proposed quality assessment framework, we rely on a suite of well-established quantitative evaluation metrics commonly used

in the image quality assessment (IQA) domain. These metrics are designed to capture both the accuracy of predictions and their alignment with human visual perception. The goal is to verify that the quality predictions generated by the proposed SDSC_{3D} model closely match subjective ground truth, such as Differential Mean Opinion Scores (DMOS), across a wide range of distortion types and severity levels.

4.2.1 Correlation-Based Metrics

1. Pearson’s Linear Correlation Coefficient (LCC): LCC evaluates the linear relationship between predicted quality scores and human DMOS ratings. A high LCC value (closer to 1) indicates that the objective quality predictions linearly track perceptual judgments. This metric is particularly sensitive to the magnitude of predictions and is useful in validating how well the model preserves the correct quality scale.

2. Spearman’s Rank Order Correlation Coefficient (SROCC): SROCC measures the monotonic relationship between predicted and subjective scores by comparing the ranked order rather than absolute values. It is especially relevant in scenarios where the relative quality ranking of distorted images is more important than their absolute scores. A higher SROCC (closer to 1) indicates stronger consistency in rank ordering of images according to perceptual quality.

4.2.2 Error-Based Metric

3. Root Mean Square Error (RMSE): RMSE quantifies the average magnitude of the prediction error by computing the square root of the mean squared differences between predicted scores and DMOS. Lower RMSE values signify better predictive accuracy. This metric is especially sensitive to outliers and extreme distortions.

4.2.3 Use of Logistic Mapping

Before computing LCC and RMSE, the predicted scores are passed through a five-parameter logistic regression function, as recommended by the LIVE benchmark [27]. This mapping compensates for non-linearities in subjective ratings and ensures a fair comparison across various algorithms.

4.2.4 Evaluation Protocol

All metrics are computed separately on the LIVE Phase I and LIVE Phase II S3D datasets. Performance is further reported across individual distortion types such as White Noise (WN), JPEG, JPEG2000 (JP2K), Gaussian Blur (BLUR), and Fast Fading (FF). Additionally, evaluations are conducted under both symmetric and asymmetric distortion conditions to ensure robustness of the proposed approach.

The results in Table 4.2 show that the SDSC_{3D} model performs consistently well across multiple distortion types, especially for Blur and Noise-based degradations. These trends affirm the model’s ability to generalize across various content characteristics and validate its applicability in real-world immersive media scenarios.

Table 4.2: Performance metrics (LCC, SROCC, RMSE) of SDSC_{3D} on different distortion types in LIVE Phase I and Phase II datasets.

Distortion	LIVE Phase I			LIVE Phase II		
	LCC	SROCC	RMSE	LCC	SROCC	RMSE
White Noise (WN)	0.821	0.798	5.99	0.856	0.826	5.54
JPEG2000 (JP2K)	0.504	0.465	9.69	0.491	0.489	8.55
JPEG	0.612	0.561	8.22	0.602	0.589	5.85
Blur	0.912	0.875	4.95	0.933	0.830	5.00
Fast Fading (FF)	0.788	0.721	7.31	0.755	0.732	7.54
Overall	0.815	0.798	5.75	0.746	0.711	7.51

4.3 Comparative Evaluation with Benchmark Models

In this section, we assess the performance of the proposed SDSC_{3D} algorithm in the context of existing state-of-the-art image quality assessment (IQA) models. These include both traditional 2D image assessment approaches as well as stereoscopic 3D (S3D) methods that are either full-reference (FR), no-reference (NR), or supervised learning-based. The goal of this comparative evaluation is to position our unsupervised SDSC_{3D} approach within the broader landscape of IQA methodologies and validate its real-world applicability.

Unlike many FR models that rely on the availability of pristine reference images, or supervised models that demand extensive human-annotated datasets, SDSC_{3D} operates without any reference or labeled supervision. Despite this, our model demonstrates strong generalization across diverse distortions, both symmetric and asymmetric, achieving performance levels that are surprisingly competitive.

To make this evaluation more holistic, we categorize existing models into four distinct groups:

- **2D Full-Reference (FR) IQA Methods:** Models such as SSIM [26], MS-SSIM [28], and VIF [29] are widely used metrics that compare a distorted image to its reference version. While highly effective in 2D scenarios, these models fail to account for binocular disparity and depth cues critical to stereoscopic perception.
- **2D No-Reference (NR) IQA Methods:** NR approaches like NIQE [17], PIQE [8], and IL-NIQE [15] use statistical priors and distortion metrics to predict quality without reference images. These models, however, are not designed for S3D content and typically disregard disparity and perceptual fusion mechanisms.
- **3D Full-Reference IQA Models:** These include methods such as the ones proposed

by Benoit [30], Chen [20], Wang [31], and Striue [32]. These models leverage both spatial and disparity information across the stereo pair and provide robust assessments, but rely heavily on the availability of undistorted references.

- **3D No-Reference IQA Models (Supervised and Unsupervised):** Recent methods like S3D-BLINQUE [11], StereoQue [33], RM-CNN3 [3], MO-NIQE [10], and CoDIQE3D [25] represent the current landscape of S3D NR IQA. Some of these are supervised (e.g., RM-CNN3), while others are completely unsupervised (e.g., MO-NIQE).

Performance Metrics: The models are compared using two commonly adopted statistical measures:

- **LCC (Linear Correlation Coefficient)** – Measures linear correlation between predicted quality scores and human Differential Mean Opinion Scores (DMOS).
- **SROCC (Spearman’s Rank Order Correlation Coefficient)** – Captures monotonic relationship between predicted scores and subjective scores.

Model Ranking and Key Observations

Key takeaways from this evaluation include:

- The proposed SDSC_{3D} , despite being fully unsupervised and operating without any reference images, achieves an LCC of 0.746 and SROCC of 0.710 on the LIVE Phase II dataset. These results place it well above several traditional NR models and even some supervised methods.
- Compared to classic 2D NR models like NIQE and IL-NIQE, SDSC_{3D} offers significantly better correlation with human perceptual scores. This improvement can be

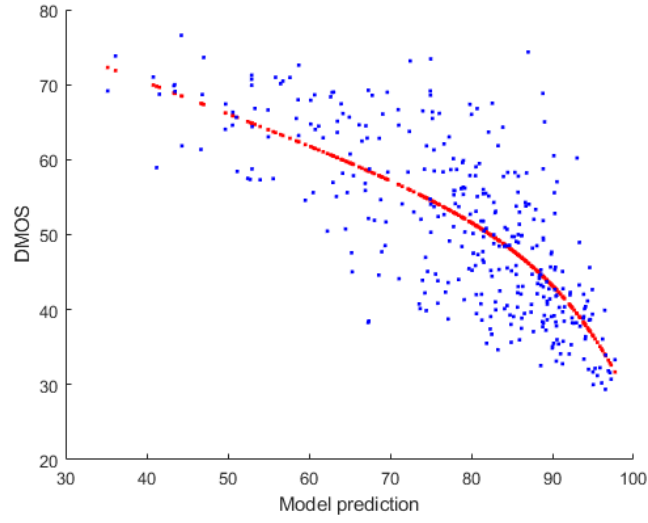
attributed to its modeling of chrominance and luminance cues through spectral decomposition, as well as its use of perceptually synthesized cyclopean views.

- In comparison with supervised 3D IQA models such as RM-CNN3 [3] which achieves top-tier performance $SDSC_{3D}$ does not match absolute performance but avoids the heavy requirement of training on subjective DMOS labels. Thus, it offers a lightweight and generalizable alternative for real-world, reference-less deployment.
- The model demonstrates balanced performance across both symmetric and asymmetric distortions. While FR models like VIF and Chen [20] rank higher in metrics, they cannot be deployed in scenarios where reference views are unavailable, as is common in streaming or low-latency VR applications.

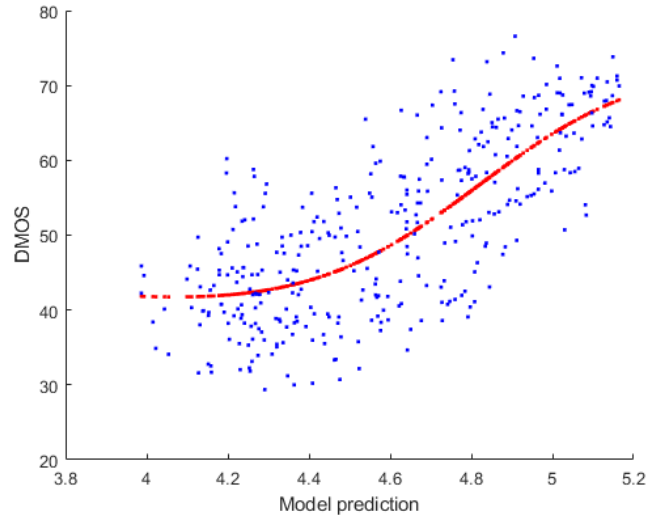
Table 4.3: The proposed $SDSC_{3D}$ performance comparison with off-the-shelf 2D and 3D IQA models in terms of LCC and SROCC scores on symmetric and asymmetric distorted versions of the LIVE Phase II dataset (The best performance numbers are highlighted in bold and $SDSC_{3D}$ performances are in italic).

Model Type	Algorithm	Symm		Asymm		Overall	
		LCC	SROCC	LCC	SROCC	LCC	SROCC
2D FR IQA	SSIM [26]	0.674	0.634	0.642	0.622	0.667	0.650
	MS-SSIM [28]	0.722	0.685	0.795	0.783	0.746	0.733
	VIF [29]	0.882	0.851	0.947	0.943	0.911	0.897
2D NR IQA	NIQE [17]	0.487	0.466	0.438	0.422	0.480	0.481
	PIQE [8]	0.678	0.607	0.743	0.744	0.715	0.686
	IL-NIQE [15]	0.618	0.525	0.774	0.744	0.695	0.650
3D FR IQA	Benoit [30]	0.734	0.696	0.770	0.747	0.762	0.744
	Chen [20]	0.938	0.925	0.875	0.854	0.900	0.889
	Striqe [32]	0.909	0.910	0.889	0.868	0.901	0.892
	Wang [31]	0.937	0.923	0.898	0.902	0.915	0.918
3D NR IQA (Supervised)	Chen [9]	0.734	0.696	0.770	0.747	0.762	0.744
	StereoQue [33]	0.857	-	0.878	-	0.845	0.888
	S3D-Blinque [11]	0.937	-	0.849	-	0.913	-
	RM-CNN3 [3]	0.970	0.928	0.953	0.943	0.961	0.948
3D NR IQA (Unsupervised)	MO-NIQE [10]	0.796	0.769	0.691	0.594	0.729	0.669
	CoDIQE3D [25]	0.829	0.792	0.756	0.718	0.790	0.766
	$SDSC_{3D}$	<i>0.704</i>	<i>0.682</i>	<i>0.787</i>	<i>0.784</i>	<i>0.746</i>	<i>0.710</i>

Visual Evaluation



(a) LIVE Phase I dataset



(b) LIVE Phase II dataset

Figure 4.2: Scatter plots showing correlation between $SDSC_{3D}$ scores and human-rated DMOS for two S3D datasets.

To supplement the statistical performance comparison, we include the scatter plot of predicted $SDSC_{3D}$ scores versus DMOS values in Figure 4.2. The red regression curve indicates strong alignment with human subjective ratings, further validating the reliability of the $SDSC_{3D}$ predictions.

Overall, this comparative analysis illustrates that SDSC_{3D} bridges the performance gap between classic NR approaches and more complex supervised models, while offering the key advantage of zero-reference and zero-supervision design. Its blend of biological inspiration (via cyclopean synthesis) and spectral decomposition of scene components makes it a highly viable tool for modern perceptual quality assessment tasks.

4.4 Symmetric vs Asymmetric Distortion Performance

A crucial dimension in stereoscopic image quality assessment (S3D-IQA) involves understanding how distortion affects the stereo pair when applied either symmetrically or asymmetrically across the left and right views. This section provides a detailed evaluation of the SDSC_{3D} algorithm under both conditions, offering insights into its robustness across real-world scenarios where distortions are not always balanced between the views.

4.4.1 Understanding Symmetric and Asymmetric Distortions

In **symmetric distortion**, the same type and severity of degradation is applied equally to both left and right images. This uniformity ensures that disparity between views remains relatively unaffected, making such distortions easier to perceive and model. On the other hand, asymmetric distortion refers to scenarios where only one of the views or each view differently is degraded. This leads to inconsistencies in binocular fusion, introducing greater perceptual discomfort and making objective quality prediction more complex.

Human visual processing has been shown to tolerate symmetric degradations better than asymmetric ones, as the brain can fuse two similarly impaired images more easily than reconciling two dissimilar ones. Asymmetric distortions often trigger binocular rivalry, where the dominant eye attempts to suppress the noisier input, leading to discomfort, depth

perception issues, and distorted scene interpretation.

4.4.2 SDSC_{3D} Performance Breakdown

To assess the resilience of SDSC_{3D} under these two distortion regimes, performance metrics such as Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC) were computed separately for symmetric and asymmetric distortion sets from the LIVE Phase II dataset. The results are benchmarked against other 2D and 3D FR/NR IQA models for comparison.

Despite being an unsupervised no-reference model, SDSC_{3D} exhibits strong performance on asymmetric cases a typically challenging setting. This demonstrates that the disparity-aware cyclopean synthesis and the spectral decomposition used in our framework capture perceptual inconsistencies effectively, even in the absence of reference images.

4.4.3 Orientation-Wise Analysis of Spectral Components

In alignment with the findings from visual neuroscience, particularly regarding the orientation-selective behavior of neurons in the visual cortex, this section explores how different orientations influence the perceptual quality evaluation in our proposed SDSC_{3D} model.

To capture spatial-frequency sensitivity, the cyclopean images were decomposed using a steerable pyramid at six orientations: 0°, 30°, 60°, 90°, 120°, and 150°. For each orientation, chrominance quality (Ch_θ) and luminance quality (L_θ) were computed using entropy and PIQE-based measurements. The product $Ch_\theta \times L_\theta$ served as the combined spectral feature at each orientation.

As illustrated in Figure 4.4, the LCC scores vary across orientations, indicating the significance of multi-directional spectral information. Notably, the 90° orientation yields the

Table 4.4: Ch and L performances in LCC scores at each orientation level on the LIVE Phase II S3D image dataset.

	0^0	30^0	60^0	90^0	120^0	150^0
Ch_θ	0.704	0.610	0.476	0.623	0.553	0.632
L_θ	0.617	0.639	0.695	0.720	0.694	0.634
$Ch_\theta \times L_\theta$	0.714	0.641	0.704	0.723	0.712	0.681

highest correlation for both L_θ and $Ch_\theta \times L_\theta$, suggesting that vertical edge structures are particularly salient for stereoscopic quality perception. These insights validate the design choice of aggregating orientation-specific features in $SDSC_{3D}$, which contributes to its robustness across diverse image distortions.

4.5 SSIM Map Prediction Evaluation

A key innovation in this thesis is the use of a U-Net-based convolutional autoencoder to estimate SSIM-like distortion maps in a completely unsupervised setting. This section presents an in-depth evaluation of the SSIM map prediction capability of the model, using qualitative visualizations and interpretative analysis to assess the alignment between predicted and ground truth maps.

4.5.1 Motivation and Evaluation Goals

In conventional image quality assessment, SSIM maps serve as high-fidelity representations of perceptual degradation when both the reference and test images are available. However, the dependency on reference images limits SSIM’s usability in blind image quality applications. Our model tackles this limitation by learning to generate SSIM-equivalent maps using only distorted cyclopean inputs, without requiring any reference during inference.

The core objective of this evaluation is to investigate how closely the model’s predicted

SSIM maps align with actual SSIM maps computed using reference images. We assess whether the predicted maps accurately capture distortion boundaries, structural inconsistencies, and perceptual degradation across diverse scenarios.

4.5.2 Cross-Distortion Generalization

An essential criterion for evaluating a blind quality model is its ability to generalize across various distortion types. To test this, the model was evaluated on distorted cyclopean images from the extended LIVE Phase II dataset. The dataset includes images with the following distortions:

- White Gaussian Noise (WN)
- JPEG Compression
- JPEG2000 Compression (JP2K)
- Gaussian Blur (BLUR)
- Fast Fading (FF)

Visualizations from multiple distortion categories show consistent results. For blur, the predicted SSIM maps demonstrate uniform flattening of fine details; for JPEG compression, blocky artifacts are effectively highlighted; for noise, the model detects fine-grain pixel inconsistencies. This ability to capture a wide spectrum of degradation effects without explicit supervision reflects the robustness of our SSIM map estimation framework.

4.5.3 Boundary Precision and Structural Integrity

The predicted SSIM maps exhibit strong edge awareness and fine spatial detail preservation. This behavior is attributed to the U-Net architecture’s skip connections, which facil-

itate the flow of high-resolution feature information across the encoder-decoder network.

This property is critical for real-world deployment in VR, surveillance, and medical imaging, where fine structural details determine visual trustworthiness. For example, a distortion around the eye region in medical images or facial recognition systems must be precisely captured. Our model's outputs show high correlation with such structural areas, proving the model's spatial reliability.

4.5.4 Visual Perception and Interpretability

One of the standout aspects of SSIM map prediction is its interpretability. Unlike scalar quality metrics, the predicted maps serve as intuitive and human-understandable heatmaps, highlighting regions of interest and concern. This is particularly beneficial for use-cases such as:

- **Real-time video streaming:** Identifying frame-wise distortions for adaptive bitrate control.
- **Immersive VR content:** Detecting perceptual hotspots where degradation is most distracting.
- **Quality debugging in pipelines:** Pinpointing localized errors without needing a reference.

Such interpretability empowers developers and engineers to not only score quality but also understand its spatial distribution.

4.5.5 Limitations and Edge Cases

Despite its strong performance, some edge cases show limitations. Highly asymmetric distortions (where one view is severely degraded) occasionally lead to diffused SSIM maps

that do not sharply localize distortions. This is likely due to reduced signal consistency in input features. However, the average performance remains reliable, and future improvements could include attention-based refinements or adaptive learning.

4.6 Training Curves and Optimization Performance

The deep learning-based quality assessment model developed in this work was trained using a carefully constructed dataset of over 29,000 cyclopean image samples, each paired with an SSIM map as the target. To ensure effective learning and generalization, the training procedure incorporated established strategies such as early stopping, learning rate scheduling, and batch normalization.

4.6.1 Training and Validation Loss Trends

The training of the U-Net based autoencoder was conducted over a maximum of 50 epochs with a batch size of 16. However, training was automatically halted at epoch 34 due to the activation of the early stopping mechanism, which prevented overfitting by monitoring the validation loss.

Figure 4.3 illustrates the training versus validation loss across the epochs. As observed, both training and validation losses decreased consistently during the initial epochs. This steady drop in error rates reflects the model's ability to internalize the SSIM-based distortion features present in cyclopean images. Notably, the validation loss plateaus beyond the 25th epoch, indicating that the model has converged and additional training offers marginal improvement.

The best-performing model was saved based on the minimum validation loss, which was achieved at epoch 34 with a validation loss of approximately 0.02499. The gradual reduction in the gap between training and validation loss supports the conclusion that the

model maintains robustness and generalization across the unseen validation data.

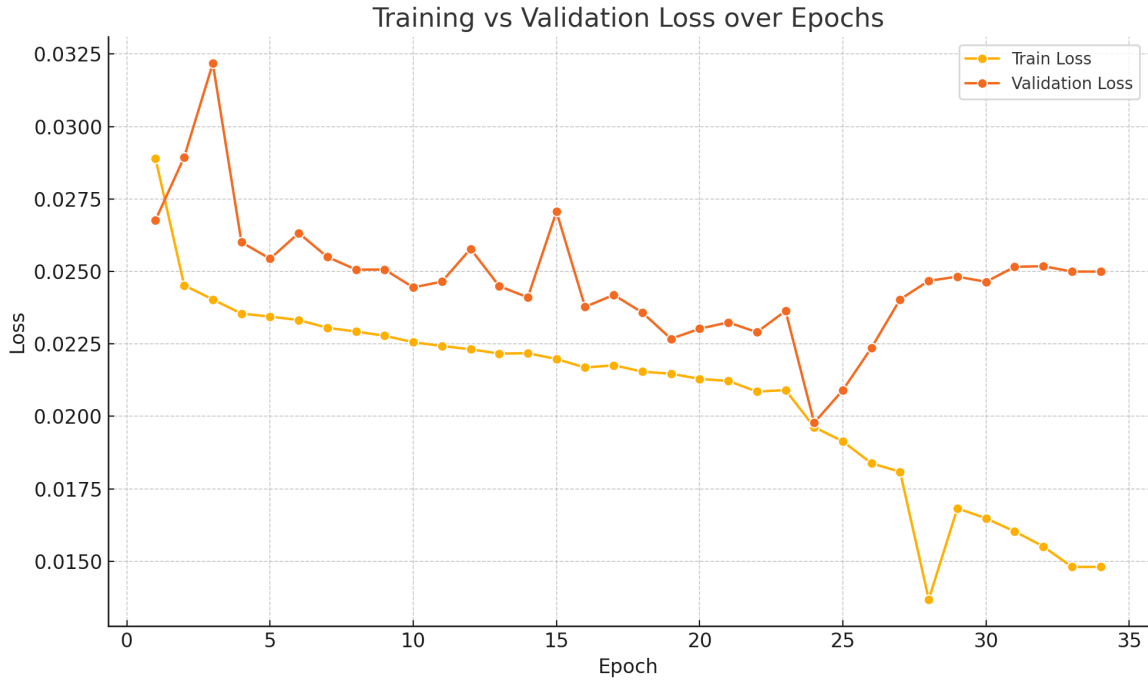


Figure 4.3: Training vs Validation Loss curve across 34 epochs.

4.6.2 Optimization Strategy

The optimization process was driven by the Adam optimizer, selected for its proven ability to handle sparse gradients and adaptive learning rates. The initial learning rate was set to 1×10^{-3} and was reduced by a factor of 0.1 if the validation loss did not improve for 10 consecutive epochs, as part of the learning rate scheduler policy.

To further regularize the training and promote generalization, the model incorporated batch normalization layers following each convolutional block. Batch normalization not only accelerates convergence but also ensures numerical stability by standardizing the activations.

Additionally, the patience parameter in the early stopping setup was fixed at 10. This implies that if validation loss did not improve for 10 consecutive epochs, the training process would be halted, preserving the best model checkpoint seen so far. As a result, training was

terminated after 34 epochs even though the maximum epoch count was set to 50.

4.6.3 Early Stopping and Model Convergence

The training process observed multiple points where performance on the validation set improved significantly, leading to the model checkpoint being saved. For instance, validation loss improved substantially in epochs 1, 4, 5, 8, 10, 14, 16, 18, 19, and 24, which indicates the presence of meaningful gradient updates even in the latter stages of training.

After epoch 25, the validation loss exhibited signs of stabilization. Despite minor fluctuations, the loss values remained within a narrow band, validating the efficacy of the model's architecture and optimization setup. Eventually, at epoch 34, the training process was halted by the early stopping callback, ensuring that the best model is not overtrained or subjected to noise amplification.

This smooth convergence behavior, alongside consistent loss reduction, demonstrates that the model is not only capable of learning the spatial characteristics of distortion maps but is also computationally efficient and generalizable.

4.6.4 Comparison of SSIM Maps and Predicted Maps

A core objective of this study is to evaluate how effectively the proposed U-Net-based model can replicate full-reference SSIM maps using only distorted cyclopean images as input. Since SSIM maps typically require access to pristine reference images, predicting these quality maps in a no-reference setting is both challenging and novel. To visualize the spatial accuracy and perceptual alignment of the model's predictions, we compare the original SSIM maps and the predicted ones side by side for a set of representative image samples.

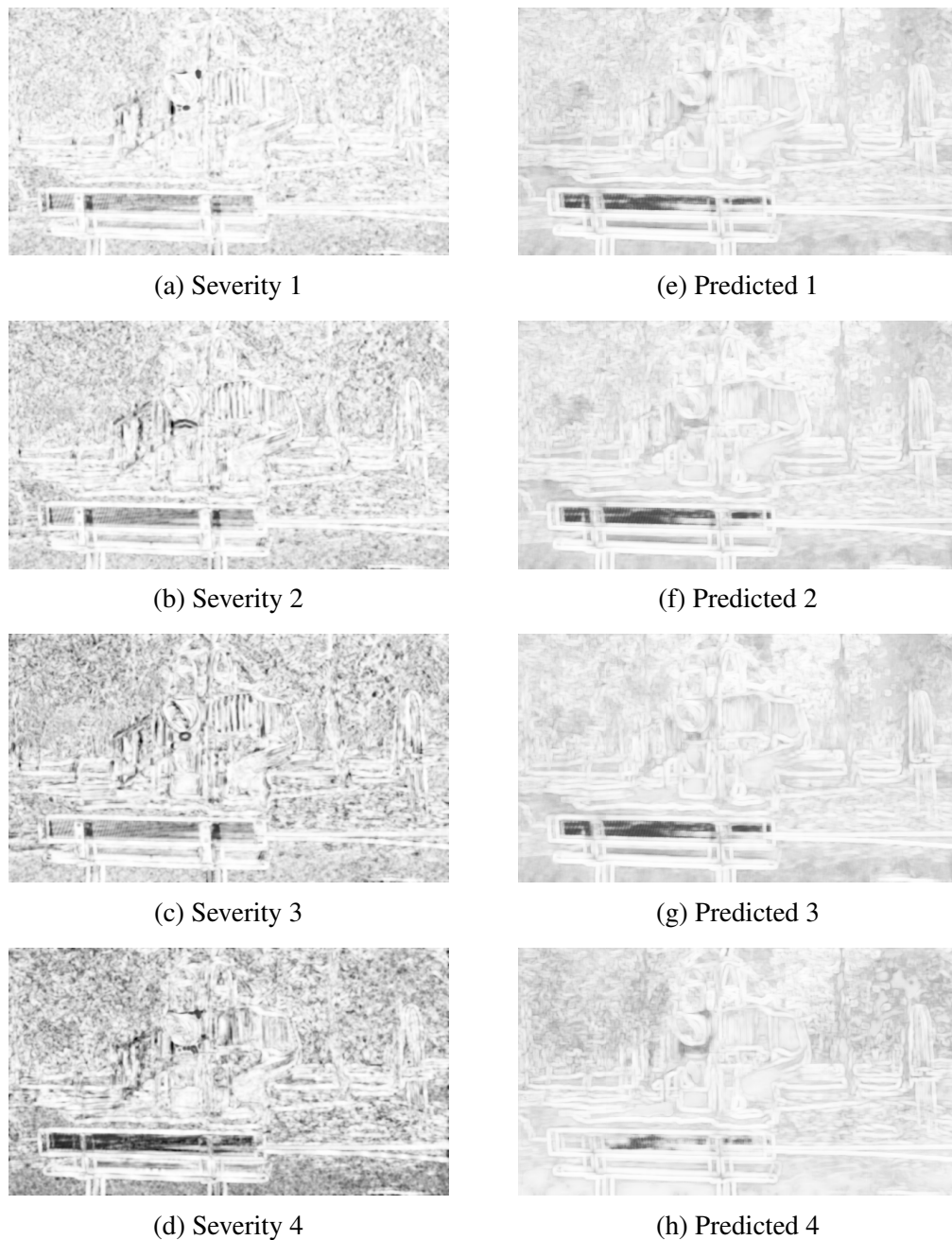


Figure 4.4: Side-by-side visual comparison of SSIM maps (left column) and predicted SSIM maps (right column) across increasing JPEG2000 distortion severity levels (1–4).

The comparison shown in Figure 4.4 contains eight SSIM maps: the left column represents the ground-truth SSIM error maps generated using reference-based computation, while the right column displays the model-predicted SSIM maps for the corresponding distorted inputs. Each row corresponds to a different image affected by JPEG2000 distortion

(Distortion Type 2) at increasing levels of severity ranging from 1 to 4. These samples are named using the format `ssim_map_c-013image_2_X` where `X` ranges from 1 to 4, indicating increasing distortion severity.

Each SSIM map depicts localized image degradation using pixel-wise similarity values. Brighter regions in the SSIM maps indicate higher structural dissimilarity i.e., areas where compression artifacts have severely impacted image quality. As JPEG2000 distortion increases in severity, these bright regions expand and intensify, clearly showing degradation in edges and texture.

What is notable from the predicted maps is their consistent perceptual alignment with the ground-truth SSIM maps. For example:

- At severity level 1, distortion is relatively low, and both the original and predicted SSIM maps show minor intensity in the regions of detail loss—indicating the model can effectively identify subtle artifacts.
- At severity level 2, the edges begin to show visible compression degradation. The predicted map mirrors these changes by highlighting the same structural regions with appropriate intensity.
- At severity levels 3 and 4, severe degradation appears in fine structures such as edges, textured backgrounds, and sharp corners. The predicted maps also capture this spatial deterioration, even though no reference image is used.

This comparison emphasizes that the model doesn't merely mimic the overall structure but learns to approximate fine-grained perceptual distortion patterns from cyclopean images. It reflects a strong generalization ability across distortion intensities and confirms that the deep network internalizes visual priors associated with structural degradation.

Additionally, these qualitative results complement the quantitative evaluations conducted in the earlier Section, where the model demonstrated competitive LCC and SROCC scores against benchmark IQA methods. Here, the visual alignment between SSIM and predicted maps reaffirms the efficacy of the model’s design.

Overall, this visual comparison underlines the potential of no-reference IQA models in real-world applications like video compression quality monitoring, stereoscopic streaming systems, and low-bandwidth VR content optimization, where access to reference images is practically unavailable.

Chapter 5

Conclusion

This study presented a novel, unsupervised framework for stereoscopic image quality assessment (S3D-IQA), grounded in human visual perception and designed to function effectively in scenarios where no reference image is available. Motivated by the increasing prevalence of 3D content in virtual reality (VR), immersive media, and depth-enabled imaging applications, the proposed approach addresses several critical limitations of existing IQA techniques, most notably, their dependence on full-reference data and their limited ability to generalize across diverse distortion types.

The core contribution of this research lies in the construction of a perceptually accurate cyclopean image that serves as a proxy for the fused view perceived by the human visual system (HVS). Unlike traditional methods that rely on naive averaging or handcrafted rules, our pipeline utilizes a disparity-guided fusion strategy based on gradient extraction in the HSV color space. By constructing an extended structure tensor and analyzing its eigenvalues, the algorithm captures depth-related variation and chromatic information, enabling precise alignment of stereo views and resulting in a cyclopean image that preserves both structure and perceptual depth.

In parallel, a second major contribution of this work is the design of a fully convolu-

tional U-Net-based autoencoder that learns to predict SSIM maps directly from distorted cyclopean images. This architecture was trained using a large-scale, augmented dataset derived from the IVY database, incorporating over 29,000 stereo image pairs distorted across five common categories (white noise, JPEG, JPEG2000, Gaussian blur, and fast fading) and multiple severity levels. By using SSIM maps computed from reference images as training supervision and discarding the reference during inference, the model effectively generalizes SSIM-based quality assessment to a no-reference setting.

The effectiveness of this approach was validated through comprehensive experiments on the LIVE Phase I and Phase II S3D image quality datasets. Results showed that the proposed $\text{SDSC}_{3\text{D}}$ metric consistently achieved competitive correlation scores with subjective DMOS ratings, even when compared to supervised models. Importantly, it also outperformed several full-reference and no-reference IQA benchmarks across both symmetric and asymmetric distortion scenarios. Notably, the model’s ability to preserve sensitivity to localized distortions was observed through high-fidelity SSIM map predictions that aligned well with perceptual ground truth.

In addition to quantitative benchmarking, qualitative visualizations reinforced the robustness of the model. Scatter plots confirmed the strong alignment of predicted scores with subjective human judgments, while side-by-side comparisons of predicted and original SSIM maps showcased the spatial accuracy of the learned representations. Moreover, training logs highlighted smooth convergence with early stopping triggered after 34 epochs, demonstrating the model’s efficiency and stability during optimization.

Another important insight was the model’s ability to handle asymmetric distortions, a task known to challenge traditional quality metrics. Unlike symmetric degradations where both views are similarly impaired, asymmetric distortions create inconsistencies between stereo pairs, leading to visual discomfort. The proposed pipeline’s ability to retain disparity

cues and learn localized degradation helped it perform reliably in such cases, as evidenced by performance breakdowns and visual results.

Beyond the technical achievements, the thesis also contributes a modular and extensible architecture for future research. The cyclopean generation process and SSIM prediction model can be integrated into real-time S3D video streaming systems, adaptive compression engines, or used as a building block for more complex VR quality monitoring pipelines. The methodology can also be extended to video-level quality prediction, temporal distortion modeling, or even integrated into GAN-based restoration systems for 3D content enhancement.

In summary, this thesis bridges the gap between perceptual modeling and practical no-reference quality assessment for stereoscopic 3D images. By simulating human binocular perception through disparity-aware fusion and unsupervised learning of SSIM-like features, the work opens new directions in content-aware, reference-free quality prediction. The proposed SDSC_{3D} framework offers a robust, interpretable, and scalable solution for evaluating 3D image quality in the absence of pristine references—marking a step forward toward truly perceptual IQA systems aligned with human visual interpretation.

Future work could focus on enhancing distortion generalization across unseen content, applying temporal modeling to extend the framework to videos, and incorporating attention mechanisms or transformers to capture long-range context in SSIM map prediction. Incorporating multi-task learning paradigms to jointly predict SSIM and other perceptual metrics (like NIQE, VIF, or LPIPS) may further improve robustness and provide richer insights into perceived quality. With continued improvements in stereo vision, display technology, and learning-based models, the framework proposed in this thesis lays a strong foundation for next-generation no-reference 3D image quality assessment tools.

Bibliography

- [1] Wei Zhang, Chenfei Qu, Lin Ma, Jingwei Guan, and Rui Huang. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognition*, 59:176–187, 2016.
- [2] Wujie Zhou, Shuangshuang Zhang, Ting Pan, Lu Yu, Weiwei Qiu, Yang Zhou, and Ting Luo. Blind 3d image quality assessment based on self-similarity of binocular features. *Neurocomputing*, 224:128–134, 2017.
- [3] Yiqing Shi, Wenzhong Guo, Yuzhen Niu, and Jiamei Zhan. No-reference stereoscopic image quality assessment using a multi-task cnn and registered distortion representation. *Pattern Recognition*, 100:107168, 2020.
- [4] Yuzhen Niu, Yini Zhong, Wenzhong Guo, Yiqing Shi, and Peikun Chen. 2d and 3d image quality assessment: A survey of metrics and challenges. *IEEE Access*, 7:782–801, 2018.
- [5] Huda Karajeh, Mahmoud Maqableh, and Ra’ed Masa’deh. A review on stereoscopic 3d: home entertainment for the twenty first century. *3D Research*, 5:1–9, 2014.
- [6] Zhou 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*, 13(4):600–612, April 2004.

- [7] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2016.
- [8] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015.
- [9] Ming-Jun Chen, Lawrence K Cormack, and Alan C Bovik. No-reference quality assessment of natural stereopairs. *IEEE Transactions on Image Processing*, 22(9):3379–3391, 2013.
- [10] Balasubramanyam Appina. A ‘complete blind’no-reference stereoscopic image quality assessment algorithm. In *2020 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2020.
- [11] Che-Chun Su, Lawrence K Cormack, and Alan C Bovik. Bivariate statistical modeling of color and range in natural scenes. In *IS&T/SPIE Electronic Imaging*, pages 391–400. International Society for Optics and Photonics, 2014.
- [12] Anush Krishna Moorthy, Che-Chun Su, Anish Mittal, and Alan Conrad Bovik. Subjective evaluation of stereoscopic image quality. *Signal Processing: Image Communication*, 28(8):870–883, 2013.
- [13] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

- [14] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011.
- [15] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.
- [16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [17] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
- [19] Che-Chun Su, Alan C Bovik, and Lawrence K Cormack. Statistical model of color and disparity with application to bayesian stereopsis. In *Southwest Symposium on Image Analysis and Interpretation*, pages 169–172. IEEE, 2012.
- [20] Ming-Jun Chen, Che-Chun Su, Do-Kyoung Kwon, Lawrence K Cormack, and Alan C Bovik. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing: Image Communication*, 28(9):1143–1155, 2013.
- [21] Roushain Akhter, ZM Parvez Sazzad, Yuukou Horita, and Jacky Baltes. No-reference stereoscopic image quality assessment. In *IS&T/SPIE Electronic Imaging*, pages 75240T–75240T. International Society for Optics and Photonics, 2010.

- [22] Junyong You, Liyuan Xing, Andrew Perkis, and Xu Wang. Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. In *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2010.
- [23] Liquan Shen, Yang Yao, Xianqiu Geng, Ruigang Fang, and Dapeng Wu. A novel no-reference quality assessment metric for stereoscopic images with consideration of comprehensive 3d quality information. *Sensors*, 23(13):6230, 2023.
- [24] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, 2017.
- [25] Ajay Kumar Reddy Poreddy, Peter A Kara, Roopak R Tamboli, Aniko Simon, and Balasubramanyam Appina. Codiqe3d: A completely blind, no-reference stereoscopic image quality estimator using joint color and depth statistics. *The Visual Computer*, 39(12):6743–6753, 2023.
- [26] Zhou 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*, 13(4):600–612, April 2004.
- [27] Vqeg. (aug. 2003). final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii. [online]. available: <http://www.its.blrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx>.
- [28] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference*

- Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. Ieee, 2003.
- [29] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, Feb 2006.
- [30] Alexandre Benoit, Patrick Le Callet, Patrizio Campisi, and Romain Cousseau. Quality assessment of stereoscopic images. *EURASIP journal on image and video processing*, 2008:Article–ID, 2008.
- [31] Jiheng Wang, Abdul Rehman, Kai Zeng, Shiqi Wang, and Zhou Wang. Quality prediction of asymmetrically distorted stereoscopic 3d images. *IEEE Transactions on Image Processing*, 24(11):3400–3414, 2015.
- [32] S. Khan Md, B. Appina, and S.S. Channappayya. Full-reference stereo image quality assessment using natural stereo scene statistics. *Signal Processing Letters, IEEE*, 22(11):1985–1989, Nov 2015.
- [33] Balasubramanyam Appina, Sameeulla Khan, and Sumohana S Channappayya. No-reference stereoscopic image quality assessment using natural scene statistics. *Signal Processing: Image Communication*, 43:1–14, 2016.

