Human Perception Based Multimedia Quality Assessment

Thesis submitted for the award of the Degree
of

Doctor of Philosophy

in the Department of Computer Science and Engineering

by

Deebha Mumtaz
(2018RCS0011)

Under the supervision of

Dr. Vinit Jakhetiya

and

Dr. Karan Nathwani

भारतीय प्रौद्योगिकी
संस्थान जम्मू
INDIAN INSTITUTE OF
TECHNOLOGY JAMMU

Indian Institute of Technology Jammu
Jammu 181221

July 2023

Declaration

I hereby declare that the matter embodied in this thesis entitled **"Human Perception Based Multimedia Quality Assessment"** is the result of investigations carried out by me in the Department of Computer Science and Engineering, Indian Institute of Technology Jammu, India, under the supervision of **Dr. Vinit Jakhetiya and Dr. Karan Nathwani** (IIT Jammu) and it has not been submitted elsewhere for the award of any degree or diploma, membership etc. In keeping with the general practice in reporting scientific observations, due acknowledgements have been made whenever the work described is based on the findings of other investigators. Any omission that might have occurred due to oversight or error in judgment is regretted. A complete bibliography of the books and journals referred in this thesis is given at the end of the thesis.

July 2023
Indian Institute of Technology Jammu

Deebha Mumtaz
(2018RCS0011)

"Dedicated to the Spirit of Courage, Hard work, and Perseverance that keeps us going."

Abstract

Multimedia plays a crucial role in today's digital age due to its versatility and effectiveness in conveying information, entertainment, and communication. Multimedia enriches learning experiences through interactive and engaging educational materials. However, the quality of multimedia can suffer due to various reasons such as low bandwidth, poor-quality capturing devices, transmission errors, data compression, inefficient rendering techniques, etc. These cause the occurrence of degradation like artifacts, blurring, noise, video freezing, etc. On the other hand, there is an increasing demand from the end users for better Quality of Experience (QoE). This has driven the service providers and researchers to cater to market needs and develop algorithms that can effectively assess and enhance user experience.

The domain of perceptual multimedia quality assessment (QA) aims to evaluate the subjective quality of multimedia content, such as audio, video, image, or speech, from the perspective of human perception. Moreover, Deep Learning (DL) techniques are now widely used due to their efficiency and ability to handle complex problems. Perceptual loss, a result of advances in Perceptual Quality Assessment (PQA) techniques, has replaced conventional metrics like Mean Square Error in training DL models, capturing high-level features to optimize results based on human perception. Consequently, ongoing research is increasingly directed towards perception-based loss functions, enabling more precise assessments of user experiences and driving advancements in multimedia technologies and applications.

Broadly speaking the quality assessment techniques are divided into two main categories. First is the subjective evaluation, where human beings or subjects perceive the data and annotate its quality. Though this technique encompasses human judgment it is time-consuming, and expensive, and can't be used in real-world scenarios. On the other hand, in objective quality assessment techniques automation of quality assessment is done using different metrics. These techniques are usually used since they are cheap and easier to use. Further, the objective quality assessment techniques are divided into three sub-categories, namely Full-Reference (FR) or Intrusive, Reduced-Reference (RR) and No-Reference (NR) or Non-Intrusive, based on the information used from the reference or pristine sample. In this work, we propose a series of FR and NR metrics for objectively assessing the quality of various multimedia data (3D synthesized views and

User Generated Audio) and try to mimic the human perception system.

The field of Image Quality Assessment (IQA) involves analyzing and quantifying various distortions present in images. In recent times, Depth Index-Based Rendered (DIBR) views or 3D views have gained popularity due to their widespread application in Virtual Reality, Free-viewpoint Televisions, etc., offering viewers an immersive experience. Apart from the conventional degradations present in natural images, these DIBR views suffer from geometric degradations caused due do poor rendering and inpainting techniques. Thus the conventional IQAs are not effective in the quality assessment of the DIBR view. In this regard, the researchers have come up with metrics that take the unique characteristics of DIBR images into consideration and tailor the QA metric accordingly.

User Generated Multimedia (UGM) encompasses the multimedia data created, captured, uploaded and shared by naive or non-professional users in in-the-wild scenarios. Such data is prone to various types of distortions caused by poor capturing devices, low bandwidth for sharing, background noise, low bit-rate etc. Further, the UGM audio has different acoustic characteristics as compared to plain speech content, and consequently, the quality assessment algorithm designed for speech signals can not be directly employed on UGM. Thus, formulating an efficient quality assessment metric for UGM is a vital task.

As mentioned above, in this thesis, we propose different metrics for the quality assessment of the two multimedia types i.e. DIBR views and UGM audio content. A brief summary of these metrics is given below;

1. **Full Reference Quality Assessment Metric for DIBR Views:**

   Many of the existing QA metrics make use of feature maps (such as the Laplacian pyramid, LBP maps, and saliency maps) to highlight certain attributes in the image for effective quality assessment. However, in DIBR images, texture and edge information may be lost during the acquisition of these maps, which is crucial for their quality assessment. In the first chapter of the thesis, a novel quality assessment approach for DIBR images is proposed, utilizing Non-Subsampled Contourlet Transform (NSCT) maps. The NSCT employs a non-subsampled pyramid structure, maintaining multi-scale characteristics, and a non-subsampled directional filter bank for directionality. Thus, effectively capturing visually significant contours near object edges or occlusions, which are highly perceptible to the human visual sys-

tem. Additionally, the NSCT decomposition does not involve any down-sampling or up-sampling operations, thereby preventing frequency aliasing in low-frequency submaps. To further refine the features, a pre-trained CNN model is used to generate compact and representative features. The final step involves finding the difference between the feature vectors of the reference and synthesized views. The results demonstrate high performance compared to existing quality assessment techniques for DIBR images.

2. **No Reference Quality Assessment Metric for DIBR Views:**
   From the literature, it was analysed that some of the existing block-based NR metrics typically divide an image into blocks and assign the same subjective quality scores to each block for training a deep learning model. However, this approach is not suitable for DIBR synthesized views, as distortions are often localized in specific areas rather than affecting the entire view. Consequently, the performance of existing block-based deep-learning algorithms suffers due to the absence of accurate ground truth scores for each image block. To address this limitation, this work proposes an innovative method for determining ground truth scores for individual image blocks. Firstly, we obtain the deep features of NSCT map of an image block and the quality score for each block is calculated using its and the reference block's feature vector. These block-wise ground truth scores are used to train a deep learning model which serves as a NR metric for estimating the quality of a given test block. Finally, the predicted block-level quality values are aggregated to determine the overall quality of the entire image. Experimental results demonstrate that the proposed NR algorithm outperforms existing NR objective metrics for DIBR synthesized views.

3. **Non-Intrusive Audio Quality Assessment Metric for User-Generated Multimedia Using Deep Learning:**
   This work begins by conducting a comprehensive analysis of the existing audio databases for quality evaluation and identifying their limitations. To fill the research gap in assessing UGM audio content, we create a benchmark audio repository called the IIT-JMU-UGM Audio Dataset along with human annotations. This dataset comprises diverse content, context, and degrees of distortions typically present in various real-time multimedia applications. Then, human subjective testing is con-

ducted on the dataset to obtain ground truth on quality. Moreover, a robust end-to-end non-intrusive metric is proposed for estimating the quality of audio. The proposed metric is based on stacked Gated Recurrent Unit (GRU) architecture. The "gated" mechanism in GRUs enables them to control the flow of information through the network, making them well-suited for tasks that involve sequential data with long-range dependencies. The proposed metric outperformed the existing state-of-the-art intrusive and non-intrusive methods applied to the dataset.

Acknowledgement

The term Ph.D. translates to 'lover of wisdom', which refers to the immense knowledge a student gains when earning the degree. Before embarking on this journey, I thought that this degree would just specialize me in the subject and increase my chances of employment. However, during the course of time, I discovered that Ph.D. transformed my personality in ways I never thought possible. This degree not only attests that I have made a (very small) contribution to research, but it has also allowed me to experience significant personal growth. It has taught me how to learn and also unlearn, the power of hard work and dedication. I am incredibly thankful to everyone around me who provided support and guidance along the way.

First of all, I express my deepest gratitude to my supervisor, Dr. Vinit Jakhetiya, for being a wonderful mentor. His teachings and dedication have shaped me as a researcher, and I owe my progress to his constant efforts. He has taught me the value of diligence and sincerity in work. I would also like to thank my co-supervisor, Dr. Karan Nathwani, for his support and concern during my Ph.D. I express my thanks to my SRC committee members, Dr. Sumit Pandey and Dr. Ashok Bera for their encouragement. I am also grateful to Dr. Badri Narayan Subudhi and Dr. Sharath Chandra Guntuku for their feedback and mentorship. Furthermore, I am thankful to IIT Jammu, MHRD, and Microsoft for providing the necessary resources and financial assistance, which allowed me to carry out research and travel.

My family has been an integral part of my academic life and has sacrificed many pivotal life moments in this pursuit. I am highly grateful to my parents Mr. Firdous Ahmad Bhat and Mrs. Shahzada Akhter. Their love, support, and trust have been the driving force behind my achievements so far. I am deeply thankful to my husband, Sajjad, for his patience and for being my pillar of strength. His presence during the ups and downs of life has been invaluable. I convey deep gratitude to my sister, Saiqa, for always being there to listen to my worries and cheering me. Additionally, I am thankful to my brother-in-law, Audil Hussain, for being always at my side. Both of them have been my go-to people whenever I encountered difficulties. I cherish the constant love and encouragement of my brother, Ubaier, in shaping my journey. I express my gratitude to my sister-in-law, Snowbar, for instilling in me the belief that everything will be alright. I would like to express my gratitude to my Aba and Ami, Mr. Abdul Rashid Sofi, and

Mrs. Yasmeen Akhter, for their relentless care and patience. I feel overwhelmed by my bunch of munchkins, Usmaan, Khadija, Habib, Venkie, and Ali, who brought love and innocence into my life. Spending time with them and watching their videos on a loop has been a source of calm and joy.

I consider myself fortunate to have a second family in the form of friends and colleagues who have been there for me through both the good and the challenging times. I consider myself blessed to have met Sadbhwana who has become my soul sister. She introduced me to a different outlook on life, empowering me to think out of the box, be positive and have the courage to believe in myself. I want to express my heartfelt gratitude to Ambreen for her care and support. Her thoughtfulness in detail and pursuit of perfection has always amazed me. Our endless conversations and shared understanding of life's journey will be cherished as my fondest memories. I am also thankful to Mrs. Annu Maheshwari, who has been like an elder sister, opening her heart and home to us. Her kindness and warmth during stressful times have been a source of comfort.

I want to extend my thanks to Akshita and Aanchna for being wonderful friends. I would also like to thank Insha Amin, Insha Wani, Mehran, Suhaib, Ayoub, Burhan, Haseeb, Tawseef, Sapna, Ajaz, Saima, and Gazenfer for being a part of my journey. I want to thank and wish luck to my juniors Ajeet, Rishika, Poonam and all other lab mates. Additionally, I want to thank the M.Tech students and interns who worked with me, bringing in their valuable understanding and skills. I also want to thank my childhood friends Bisma, Ifham, and Heeba for being around. I want to thank Dr. Arooj for encouraging me to take admission in Ph.D. at IIT Jammu.

I extend my gratitude to all the service providers who have worked diligently behind the scenes, unknowingly making my life easier. I also want to express my gratitude to my aunts and uncles for their concern and care for me. A special mention to Aunt Mrs. Hamida Bhat, who named me after a famous lady doctor and takes pride in saying that I have become one (though not a medical doctor!).

Above all, I want to thank Almighty Allah for giving me the opportunity, capability and resources to finish this journey.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
ACR	Absolute Category Rating
APT	Auto Regression Plus Thresholding
AQA	Audio Quality Assessment
AR	Auto Regression
BIQI	Blind Image Quality Index
BRISQUE	Blind Referenceless Image Spatial Quality Evaluator
BRISK	Binary Robust Invariant Scalable Keypoints
CLGM	Combining Local and Global Measures
CL	Convolution Layer
CNN	Convolution Neural Network
CODIF	COlor-Depth Image Fusion
DCT	Discrete Cosine Transform
DF-CS	Deep Features Cosine Similarity
DIBR	Depth Image Based Rendered
DMOS	Differential Mean Opinion Score
DSCB	Distortion Specific Contrast Based
FCL	Fully Connected Layer
FR	Full Reference
FVV	Free View-point Video
GANs	Generative Adversarial Networks
GRU	Gated Recurrent Unit
HHF	Hierarchical Hole Filling
HPS	Human Perceptual System
HVS	Human Visual System
IDEA	Instance DEgradation and global Appearance
IQA	Image Quality Assessment
KRR	Kernel Ridge Regression
KRCC	Kendall Rank Correlation Coefficient
LDI	Layered Depth Image
LBP	Local Binary Pattern
LPIPS	Learned Perceptual Image Patch Similarity
LSTM	Long Short Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
MHA	Multi-Head Attention
MOS	Mean Opinion Score

Abbreviation	Description
MP	Morphological Pyramids
MPL	Max Pooling Layer
MQA	Multimedia Quality Assessment
MSE	Mean Square Error
MW	Morphological Wavelet
NIQE	No Reference Image Quality Evaluator
NR	No Reference
NSS	Natural Scene Statistics
NSCT	Non-Subsampled Contourlet Transform
NSDFB	Non-Subsampled Directional Filter Bank
NSPFB	Non-subsampled Pyramid Structure Filter Banks
NSP	Non-subsampled pyramid
PLCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal to Noise Ratio
PQA	Perceptual Quality Assessment
PU-IR	Perceptually Unimportant Information Reduction
QA	Quality Assessment
QoE	Quality of Experience
RANSAC	RANdom SAmple Consensus
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RR	Reduced Reference
RV	Reference View
SAMVIQ	Subjective Assessment Methodology for Video Quality
SIFT	Scale Invariant Feature Transform
SI-DL	Stretching Identification Deep Learning
SNR	Signal-to-Noise Ratio
SOTA	State-Of-The-Art
SQA	Speech Quality Assessment
SR	Super Resolution
SROCC	Spearman Rank Order Correlation Coefficient
SSIM	Structural SIMilarity
SS	Statistical Significance
SURF	Speed Up Robust Features
STFT	Short Time Frequency Transform
SV	Synthesized View
UGM	User-Generated Multimedia
VR	Virtual Reality
VGG	Visual Geometry Group
VSRS	View Synthesis Reference Software

List of Symbols

Symbol	Description
r	Pearson Linear Correlation Coefficient
ρ	Spearman Rank Correlation Coefficient
τ	Kendall Rank Correlation Coefficient

Chapter 1

Introduction

1.1 Multimedia Quality Assessment

In recent years, the rapid growth of communication techniques, affordable capture devices, social media, and digital technology has led to an increase in the creation and transmission of multimedia content. This has impacted various areas, including online learning, entertainment, information sharing, healthcare services, etc. However, challenges like storage, limited bandwidth, device diversity, and network issues can degrade the quality of transmitted data. Consequently, there is a need to prioritize the end user's quality of experience. High-quality multimedia content enhances user satisfaction, engagement, and retention. Quality assessment (QA) techniques play a vital role in identifying and resolving issues such as compression artifacts, distortion, color accuracy, and audio clarity, thereby improving the user experience. These techniques also assist in optimizing encoding parameters, compression algorithms, and multimedia processing methods to achieve the best quality possible within given limitations.

Multimedia quality assessment deals with quantifying the degree of degradation that affects the quality of multimedia data such as images, audio, video, etc. Specifically, the perceptual quality assessment (PQA) considers the human perceptual system's (HPS) response to content, leading to a more realistic evaluation of multimedia quality compared to traditional objective metrics. In this context, Wang et al. [9] investigated the limitations of metrics like "mean square error (MSE)" in assessing image quality, as they fail to align effectively with human perception. Instead, they introduced SSIM, a perception-based quality metric that incorporates human perceptual characteristics, resulting in a

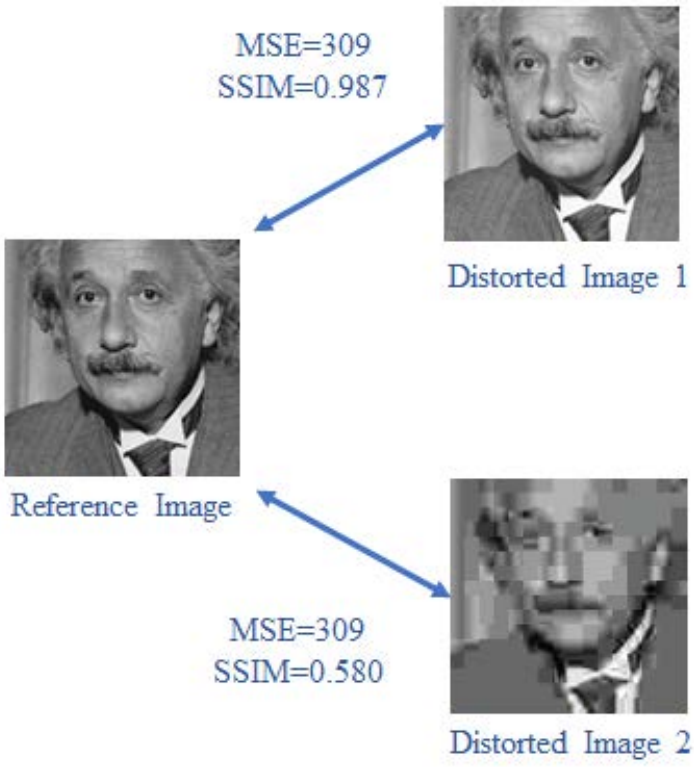


Figure 1.1: Example highlighting the importance of perception-based quality assessment metric [1].

more accurate representation of perceived image quality. To emphasize the significance of perceptual intuition in quality assessment tasks, an example image is presented in Fig 1.1, a clean image is altered by various degradations. Both images have significantly different perceptual qualities, however, the same MSE of these images is the same. On the contrary, this quality difference is efficiently captured by the SSIM metric which is sensitive to HPS, thus giving more realistic results.

Moreover, the efficiency and ability to address complex problems have led to the use of Deep Learning (DL) techniques for a wide range of applications. During training a DL Model, the loss function is crucial for measuring how well a model's predictions match the actual target values. Initially, common metrics like MSE and Mean Absolute Error (MAE) were used as loss functions. However, advances in PQA techniques have resulted in the adoption of perceptual loss in training the DL models [10–14]. Perceptual loss captures high-level features to improve the overall quality of generated data. This helps to optimize results based on human perception. Thus, current research is inclined towards perception-based loss functions that offer better evaluation of the user's experience, leading to improved multimedia technologies and applications.

Two primary techniques are used to determine the quality of multimedia data. The

first technique is subjective testing, where human subjects assess and annotate the quality of the data. While this approach relies on human perceptual judgment, it is time-taking, labor-intensive, expensive, and not effective in real-world scenarios. In contrast, objective quality assessment techniques automate the assessment process and try to mimic the HPS. These techniques are preferred due to their cost-effectiveness and ease of use, although they may not perfectly model human perception. Objective QA techniques are classified into:

1. **Full-Reference (FR)/ Intrusive:** This technique requires both a reference (clean) sample and the distorted/test sample for comparison. While FR QA metrics provide high accuracy, it may not be feasible when a clean reference is unavailable or in real-life situations.

2. **Reduced Reference (RR):** In this technique, only selected parameters from the reference signal are utilized for quality assessment.

3. **No-Reference (NR)/ Non-Intrusive:** These metrics assess quality using only the features of the test/degraded signal, making it more realistic for practical use.

These objective QA techniques offer different trade-offs in terms of accuracy and practicality, allowing for efficient assessment of multimedia quality in various scenarios.

In literature, a lot of work has been done on various types of multimedia data such as audio, images, video, text, etc. Their quality is objectively assessed along various dimensions, such as degradation type, degree, location of distortions, impact on HPS, availability of reference data, sample size (pixel, block, overall), applicability, etc. With this view, this thesis aims to propose approaches for assessing the quality of the following two important types of multimedia;

1. **Quality Assessment of Depth Image Based Rendered Views:**
   The 3D synthesized or Depth Image Based Rendered (DIBR) views/images represent the synthesized views (SV) of a scene or object generated using depth information to enable virtual camera movements or produce stereo/multi-view displays. These have recently gained popularity in various domains like augmented reality, free-viewpoint televisions, virtual reality, and more. Conversely, current 3D synthesis algorithms often generate distortions like stretching artifacts, misalignment,

and a range of geometric and structural errors. Therefore, evaluating the quality of these synthesized views is a significant research aspect of computer vision.

2. **Audio Quality Assessment of User-Generated Multimedia**
   User Generated Multimedia (UGM) refers to multimedia data created, captured, uploaded, and shared by naive/non-professional users in real-world scenarios. Such data is often subject to various distortions caused by low-quality capturing devices, low bandwidth, background noise, low bit rate, etc. Furthermore, the existing metrics mainly quantify speech quality and intelligibility which has different acoustic properties as compared to UGM audio. Thus, quality assessment algorithms for speech signals cannot be directly applied to UGM, and developing an efficient quality assessment metric for UGM is essential.

This thesis aims to conduct a systematic study of existing work on quality assessment of DIBR views and UGM audio, addressing the limitations, and devising novel QA techniques for improved performance. The detailed literature survey is given in the proceeding section.

1.2 Quality Assessment of DIBR Views

Depth Image Based Rendering is a technique to create new/novel views of a scene from existing images and their corresponding per-pixel depth information. The depth information provides the distance of each point in the scene from the camera's viewpoint. By utilizing this, the rendering technique allows for the synthesis of additional views of the scene, creating a sense of depth and perspective for the viewer. In the Free-viewpoint video, the DIBR views can be interactively explored by changing the viewpoint which gives an immersive effect. This helps in the creation of realistic visual experiences by offering multiple perspectives of the same scene [15].

The DIBR process is based on the concept of 3D warping wherein the transformation of points between 2D and 3D space is done [16]. In this process, the given image points are reprojected into a 3D world based on their depth information. This means each pixel's 2D position is transformed into a 3D position in space, accounting for its distance from the camera. Next, the 3D points in space are projected onto a virtual camera at the desired viewpoint, generating a new view of the scene [17].

In many cases, certain objects or parts of the scene that were hidden from the camera in the actual view might be evident/visible from the virtual camera's viewpoint in the synthesized view. These newly visible areas are termed dis-occlusion. Handling dis-occlusion areas is a crucial aspect of DIBR. In this regard, image inpainting techniques are often used to fill in the missing information in these areas.

Natural images often experience uniform, non-localized, and structural distortions like Gaussian noise, compression artifacts, and blur. On the other hand, the process of synthesizing DIBR views introduces distinct geometric distortions for example "flickering", "black-holes", "cracks", "ghosting", "stretching", and "crumbling" [15]. These artifacts are different from the conventional distortions typically encountered in regular natural images. As a result, DIBR views suffer from both geometric and traditional structural distortions. Consequently, various studies have demonstrated that conventional quality assessment techniques used for natural images are not efficient when applied to DIBR views [18]. Therefore, many researchers have attempted to develop specialized techniques to handle the unique challenges in the QA of DIBR views. Figure 1.2 illustrates several instances of the artifacts found in DIBR views. These artifacts can be classified into distinct categories according to their perceptual characteristics.

1. **Black-hole:** Presence of black patches or holes is caused due to filling of disoccluded areas by black pixels.

2. **Crumbling:** The edges of the object seem to be degraded in the synthesized view, primarily caused by artifacts in the depth maps surrounding strong discontinuities, resembling erosion effects.

3. **Stretching:** Due to imperfect inpainting techniques, the textures may be repeated or stretched over the holes causing stretching artifacts.

4. **Blurry regions:** In-painting methods employed to fill disoccluded areas may generate some blurry regions.

5. **Incorrect rendering:** Complex textured areas may exhibit rendering errors such as information loss, which cannot be adequately reconstructed by in-painting methods.

6. **Object shiftings:** In the synthesized view, object regions may experience slight shifts due to depth preprocessing techniques such as low-pass filtering or depth

encoding methods used to smoothen object edges.

1.2.1 Benchmark QA Datasets For DIBR Views

Benchmark datasets are extensively used to provide standardized data, enabling fair comparisons and evaluation of techniques. Additionally, the dataset should include subjective scores as ground truth for evaluation. Numerous benchmark datasets have been proposed in the literature, offering comprehensive resources for evaluating QA techniques, some of which are described below.

1. **The IRCCyN/IVC Dataset: [19]**

   The IRCCyN/IVC Dataset was proposed in 2011 and consists of three reference views (RV), which are used to render 84 corresponding synthesized views. Four distinct techniques are used for image rendering. However, the resultant views suffer from different degradations such as blur, stretching, etc. Among these, "black holes" are the most prominent degradation characterized by empty black pixels caused due to ineffective hole-filling methods. Initially, many QA metrics focused on detecting and quantifying these black-hole degradations as they have a high impact on perceptual quality.

2. **The IVY Dataset [20]:**

   The IVY dataset proposed by Jung et al. in year 2016, consists of stereoscopic DIBR views for QA. It consists of a total of eighty-four stereo image pairs generated from seven reference views which were obtained from MVD sequences and Middlebury datasets. Further, four 3D view synthesis algorithms were employed in the synthesis namely Ahn's [21], Criminisi's [22], inter-view consistent inpainting method, and MPEG VSRS. The synthesized views generated are dominated by ghosting artifacts.

3. **The IETR Dataset [18]:**

   The IETR dataset comprises one hundred forty synthesized views generated using ten reference views. Each image is associated with its subjective quality score. Seven reference views are Natural, while the remaining three are Synthetic. Seven synthesis techniques were employed in developing this dataset. The Single-view 3D methods use only one image along with its depth map to synthesise images.

(a) Black-hole

(b) Crumbling

(c) Stretching

(d) Blurring

(e) Information loss

(f) Ghosting

Figure 1.2: Examples of different types of artifacts and distortions present in the DIBR views.

Five Single-view 3D techniques (HHF [23], Luo's [24], Ahn [21], Criminisi's [22], and LDI [25]) have been employed for synthesis. Furthermore, an Inter-view DIBR synthesis method proposed by Zhu [26] and an additional (single-view and inter-view) method VSRS [27] is also included for the image synthesis. Some samples of the reference and synthesized views from the dataset are given in Fig. 1.3. The rendered views suffer from distortions such as blur, stretching, ghosting, cracks, etc. The subjective score is also provided for each of the views.

4. **MCL-3D Dataset [28]:**
   In MCL-3D Dataset comprises nine reference image sequences along with their depth maps. In this dataset, pre-DIBR synthesis distortions are explored by manually adding various distortions to the depth and RGB images. These include "JPEG and JPEG-2000 compression", "additive white noise", "down-sampling blur", "Gaussian blur", and transmission error. The data comprises more than six hundred images synthesized from the reference views along with their subjective scores.

From analyzing these repositories it was found that the datasets such as MCL-3D and IRCCyN/IVC employ outdated synthesis methods, leading to the prevalence of distortions such as black holes. However, the advancement of inpainting techniques has resulted in the almost elimination of these black-hole distortions, making it advisable to employ more recent benchmark datasets for evaluation [18]. Therefore, this thesis focuses on using IETR and IVY datasets for assessment as they represent the current scenario in rendering the DIBR views.

1.2.2 Existing Techniques for Quality Assessment

The DIBR views suffer from various degradations, this has prompted investigations in QA in the field. Presented below are some of the existing QA methods which are sub-categorized as FR and NR techniques.

1. **Full-Reference Quality Assessment Techniques**
   The details for some of the well-known full reference image QA techniques are given below.

   - **The MW-PSNR metric [29]** makes use of the "morphological wavelet" transform to get the low-level features of the images at various scales. It

(a)

(b)

(c)

(d)

Figure 1.3: Examples of reference and synthesized DIBR views from the IETR Dataset

calculates the PSNR between these maps and pools the values to calculate the quality of the image.

- **Tian** [30] presented a method for shift compensation which is caused during the synthesis of DIBR views. They also employed the "dis-occlusion mask" technique to qualify the distortions present in the dis-occlusion regions using corresponding depth maps.
- **The MP-PSNR metric [31]** makes use of "morphological pyramids" to extract perceptually important geometric details such as edges from the images. Next, it calculates the PSNR between the maps of clean and the synthesized images to calculate the quality score.
- **The "LOcal Geometric distortions in dis-occluded regions and global Sharpness (LOGS) metric" [32]** is based on determining the location of dis-occluded areas in the DIBR views and then quantifying their size and intensity. Furthermore, it obtains global sharpness by employing a "re-blurring-based strategy" and later merges the scores together to obtain the image quality.
- **The "Instance DEgradation and global Appearance (IDEA)" metric [33]** is proposed by Li *et al.* The metric integrates the global as well as local distortions measured using "super-pixel" representations and "discrete orthogonal moments" respectively. To get the quality of an image, both the scores are integrated together.
- **The Shift Compensation (SC-IQA) metric [34]** explores the occurrence of shift in the synthesized views which is caused during the rendering process. In this regard, they employed the existing SURF and RANSAC homography techniques for "shift compensation" globally as well as in a block-wise manner. Additionally, for determining the weighting function, a "visual saliency map" was employed.
- **The SuperPixel Difference metric (SSPD) [35]** proposed by Mahmoudpour *et al.*, is based on calculating the changes in global contrast and appearance along with determining the local geometrical distortions in the images. It employs the Speed Up Robust Features techniques to equate the feature points of the synthesized and reference views and determine the distortions based on

the corresponding descriptors.

- **The Saliency-Deep Feature (Sal-DF) metric [36]** employs saliency maps to highlight the perceptually vital features such as edges and textures in the DIBR views. To further refine the feature maps, convolution layers of a pre-trained model are used. The aggregated quality score is obtained by calculating the cosine similarity between the clean and degraded view.
- **The "Perceptual Representations of Structural Information" (PRSI) metric [37]** is inspired by the human perceptual system. It pools task-oriented non-natural structure descriptors along with the mid and low-level features for quality assessment. Further, the concepts of dictionary learning, sparse representation, and rank pooling are employed to determine the quality.
- **The PUIR-DFCS [38]** proposed by Sadbhawna *et al.* make use of Laplacian pyramid maps as they provide a multi-scale representation of images. Difference in the deep features of the clean and distorted maps provided intuition for the image quality. Furthermore, they made use of various morphological operations to highlight the regions where the distortions are mainly located and determined the degree of distortions in such regions.
- **The "Elastic metric" (EM) [39]** proposed by Ling *et al.* explores the fact that geometric distortions in DIBR views are characterized by bending or stretching along the object edges. This metric determines the regions where local distortions are present and then measures the degree of stretching in the curves to quantify the distortions.
- **The"Sparse Representation" (SR-VQA) [40]** by Zhang *et al.* proposes a metric which is mainly used to assess the flicker distortions present in 3D synthesized videos. The constituent spatial and temporal domains of the video are first decomposed along two planes. Further, the edges of the depth map and the gradient features are analysed for detecting flicker. Finally, sparse representation, dictionary learning, and rank-pooling techniques are applied to determine the quality score.
- **The "Context Identification" metric [41]** relies on an observation that the perceptual quality of DIBR views is greatly dependent on the context

area i.e. foreground or background from which the inpainting for hole-filling is applied. With this motivation, the authors first predict the locations where the geometric distortions are mainly present and then quantifies the distortions present in the region.

- **The "Local and Global Geometric Distortions" metric (LGGD)** [42] proposed by Peng *et al.* aims at determining the geometric distortions at both the local as well as global levels. The metric makes use of a "sketch token-based local edge descriptor", as most of the distortions lie along the edges of the objects. Moreover, the spatial pyramid-improved Bag-of-Token and "pixel-level backward registration" techniques are explored for determining the global distortions in the images.
- **The "SEmantic- and QUality-aware feature Similarity measures plus a Salient-region detection metric (SEQUSS)"** [43] is proposed by Mahmoudpour *et al.*. This work makes use of CNN for shift compensation and the generation of visual saliency masks. For quality assessment, the metric computes and pools together the structural and semantic similarity of the clean and distorted views.
- **The "Adaptive- Deep Image Structure and Texture Similarity (A-DISTS) metric"** [44] proposes a technique for determining the "locally adaptive structure and texture similarity index" between the reference and synthesized image. It makes use of the dispersion index for obtaining single statistical features at various scales. The metric is used in the QA of super-resolution as well as synthesized images.
- **The Wang's metric** [45] proposes to quantify the geometric and texture degradations by extracting relevant features from course to fine level. The techniques used in this metric involve DCT-based texture similarity, color co-occurrence matrix, edge-based region similarity, etc. The features are then learned using a machine-learning model to determine the quality of the synthesized view.

2. **Reduced-Reference Quality Assessment Techniques**

These techniques require some information from the clean sample. Some of the RR

IQA are given below:

- **The "Reduced Reference Entropic Differencing (RRED)" metric** [46] is based on deriving the " Wavelet Coefficients (WC)" of the given images. The quality value is obtained by determining the mean difference between the "scaled entropies" of the WC. Further, they proposed a series of measures based on the quality subband taken into consideration and the amount of information required from the clean signal. The metric is parameter-free and is used in applications related to quality monitoring in networks for visual multimedia.
- **Mahmoudpour *et al.*** [47] proposed a technique to determine the quality of images corrupted by various types of degradations. This technique employs the "internal generative mechanism" for decomposition along with the entropy calculation. The difference calculation in the clean and distorted feature vectors is followed by training a regression model for the final quality prediction.
- **The "Spatial Efficient Entropic Differencing (SpEED)" metric** [48] is based upon the techniques of NSS modelling. For quality prediction, the difference in the local entropy of the two images in the spatial domain is taken into account. Its usability has been extended to the domain of video QA also.
- **Jinjian *et al.*** [49] proposed to make use of the concept of change in the saliency of an image for QA. In this process, initially, the saliency regions are identified which is followed by obtaining the vital features using the "local saliency weighted histogram" and "global saliency-based histogram". These features are then combined together to determine the quality.

3. **No-Reference Quality Assessment Techniques:**

The details of some of the SOTA NR IQA metrics available are given below.

- **The "Auto-regression Plus Thresholding (APT) metric"** [50] analysed that though the existing "NSS models" are capable of quantifying the distortions in natural images, these may not be usable for DIBR views. In this regard, the work proposes to exploit the "local image description" in the images to describe the similarity among the pixel in a patch. It further determines the difference in the DIBR view and the "auto-regression-based map" to represent the image quality.

- **"Blind Image Quality Indices (BIQI)" metric** [**51**] is a two-step technique for QA. Initially, it measures the probability of a particular distortion (like "Gaussian Blur", "compression" etc.) in the image using a machine learning model. This is followed by mapping these scores to their individual quality. The final quality of the images is calculated as a weighted sum of these individual quality values.
- **The "Kernel-Ridge Regression" metric (KRR)** [**52**] is based on using the "global kernel ridge regression" technique for quality prediction. The metric determines the boundaries of the areas affected by geometric distortions followed by quantifying the degree of degradations in these regions. It also makes use of the NSS model to determine structural degradations. The image quality score is calculated by aggregating these values together.
- **The "No reference Image Quality assessment method for 3D Synthesized Views (NIQSV+)"** [**53**] proposed by Tian *et al.* It relies on quantifying the most prominent distortions in the DIBR views like; stretching, blur, crumbling, and black holes. These are determined by gradient calculation, luminance measurement, difference calculation in colour components and finally pooling these scores together.
- **The "Multiscale Natural Scene Statistical (MNSS)" metric** [**54**] is a combination of a number of NSS models specifically designed for 3D views. It measures the damage caused by geometric distortions to the basic characteristic of natural images and quantifies the statistical irregularity at different scales of the image.
- **The "No-Reference Morphological Wavelet with Threshold (NR-MWT)" metric** [**55**] relies on the assumption that due to distortions, the amount of "high-frequency" content gets increased in 3D synthesized views. It identifies the regions in the high-high wavelet subband using morphological wavelet transformation and thresholds these to determine the areas with the highest distortion sensitivity.
- **The "Geometric Distortions and Image Complexity (GDIC)" metric** [**56**] suggested by Wang et al. employs "discrete wavelet transform" to decompose the images into wavelet sub-bands. Next, the edges detection tech-

nique is used followed by measuring the similarity between high and low frequency subbands. The quality is measured by normalizing geometric distortions through image complexity.

- **The Wang's metric [57]** proposes a QA metric based on pooling the image complexity, global sharpness, and geometric distortion values. This is done by decomposing the image by discrete wavelet transform, followed by edge detection and measurement of geometric distortions. The sharpness is calculated by "log-energies of wavelet subbands" and image complexity is computed by bilateral and autoregressive filters.
- **The "COlor Depth Image Fusion (CODIF)" metric [58]** proposed by Li *et al.* explores the color and depth representations of the DIBR views. For determining the boundaries of the image, its color information is utilized followed by a Wavelet-based technique to represent the fusion between depth and color image. The quality score is determined by training a model using statistical features of natural and interaction areas of the image.
- **The "Local Variation and Global Change (LVGC)" metric [59]** by Yan *et al.* employs the "Gaussian derivatives" and LBP technique for extracting and quantifying the local degradations in terms of chromatic and structure features. The global degradations are sensed by assessing the degree of naturalness in the views. Finally, by training a regression model using these features, predictions about the quality are obtained.
- **The "Generative Adversarial Networks based No Reference Quality Assessment Metric (GANs-NRM)" [60]** proposed by Suiyi *et al.*, employs "Generative Adversarial Network" to render mask regions in an image. After training the network, it is assumed that the discriminator of the trained GAN possesses the capability to assess the quality of the masked area. Thus, the features obtained by the discriminator to learn the "Bag-of-Distortion-Word codebook" are then used for quality measurement.
- **The Yue's metric [61]** attributes the quality of the image to geometric distortions and sharpness. This is measured by analyzing local similarity while sharpness is quantized by the deviation between the distorted and the downsampled image. Later the linear pooling technique is used for aggregating the

individual scores.

- **The "Synthesized views using DoG-based Edge statistics and Texture naturalness (SET)" metric [62]** make use of the multiple scale Difference-of-Gaussian representations to obtain the features pertaining to the degradations caused in the texture and the edges. These features are then learned using the random forest regression technique.
- **The "Synthesized Image Quality Assessment with Contextual Multi-Level Feature Pooling (SIQA-CFP)" metric [63]** extracts features of the DIBR views from multiple levels of a CNN model. These high and low-level features are then aggregated using a warping technique. This is followed by contextual pooling using a deep learning model for determining the quality scores.

From the literature analysis, some of the shortcomings of the existing work are summarized below.

- The DIBR views exhibit distinct characteristics as compared to natural 2D images. This can be validated from the performance of the existing metrics (e.g., Lao's [64], SSIM [9], Cheon's [65], GMSD [66], PSNR, FSIM [67]), which work well for natural images but exhibit low performance when applied to DIBR views. This highlights the importance of developing a dedicated metric specifically tailored for evaluating the quality of DIBR views.
- Many of the existing QA metrics such as [32, 37, 50, 52] primarily focus on detecting black-hole distortions in DIBR views (IRCCYN/IVY dataset). However, due to technological advancements, black-hole distortions have nearly eradicated from current DIBR views. Consequently, these metrics are no longer able to achieve high accuracy when applied to the latest benchmark DIBR datasets, such as the IETR and IVY which exhibit other types of distortions.
- In 3D synthesized views, there is a misalignment between the synthesized image and the reference image, which occurs due to the rendering processes. As a consequence, many FR techniques struggle to perform effectively, as they rely on pixel-level comparisons. Some metrics have attempted to mitigate the problem, for instance, the

SSPD algorithm [35] involves feature matching using the SURF technique. Other approaches such as SC-IQA [34] utilize SURF and RANSAC homography while [68] uses BRISK for shift compensation. However, incorporating an explicit shift compensation step introduces additional computational complexity, potentially posing challenges to their efficiency and overall effectiveness.

- Despite the advancements in inpainting techniques that have removed degradations like black holes, imperfect rendering processes still lead to the introduction of various distortions in images. These include stretching artifacts, blurring, blockiness, shifting, and more. Several existing techniques [6, 61] have been developed to address specific distortion types, for instance, focusing on detecting stretching artifacts. While these metrics perform reasonably well for their intended purpose, there remains a need for an algorithm that would provide a holistic quality score, taking into account the diverse range of distortions affecting the images.

- Moreover, the restricted availability of a sufficiently large dataset with subjective annotations has hindered the adoption of direct deep-learning techniques for the QA of DIBR images.

- Additionally, the existing no-reference quality assessment methods such as DSCB [69], Wang's [70], Yan's [59], NRMWT [55] still struggle to perform well on DIBR views as they are not able to mimic the HPS efficiently. An efficient NR QA metric could have otherwise been an ideal solution for real-world scenarios where reference data is unavailable.

Based on this analysis, it becomes evident that the development of a new metric specifically designed for the QA of DIBR views is very important. This metric should possess simplicity and efficiency while addressing the challenges posed by the distortions present in existing databases.

1.3 Audio Quality Assessment

In recent years, there has been an increase in digital technology, leading to the rapid generation and distribution of User Generated Multimedia content. Given this context, the assessment of UGM data quality becomes crucial as it empowers service providers,

device manufacturers, and streaming platforms to deliver a better quality of experience to end-users. This section provides detailed insights into the available datasets and existing metrics used for audio quality assessment.

1.3.1 Existing Audio Datasets

This section presents several datasets that have been utilized in diverse audio applications. The details of these repositories are presented below.

- **The International Telecommunication Union (ITU)** recommended the [71] database for speech assessment, it contains clear and distorted signals along with their subjective quality scores. The distortions are caused by narrowband speech degradation, environmental noise, audio encoding, and channel degradation. The dataset is available under licence.

- **Fazenda *et al.*** in their work [72], explored UGM audio quality estimation, wherein clean samples were acquired from YouTube, which was later corrupted by background noise. However, the repository is not openly available for use, and the number of samples is 128.

- **The NOIZEUS database [2]** designed for the assessment of speech enhancement algorithms includes speech samples with thirty standard sentences [73], distorted by different real-world noises at different Signal to Noise Ratios.

- **Creusere *et al.*** in their work [74], compiled a data set with clean samples and distorted them using various bit rates. The data was obtained from multiple sequences ranging from rock to classical music. Each sample is of 9 to 24-sec duration.

- **The TIMIT database** [4] is designed to provide samples for the automatic speech recognition task. It comprises recordings from 630 American speakers reading English sentences in numerous dialects. This corpus has been used in [75], for providing clean speech signals.

- **Li *et al.* [76]** developed a database that contains clean and synthetically distorted recordings belonging to various music genres and is used for quality assessment of live music. It contains live music recordings belonging to four categories (i.e., country, rock, electronic, and pop).

- **Avila *et al.*** in their work [77] estimated speech quality using a database containing recorded speech. Home, office, and other background noises, along with reverberations, were convolved with the clean samples. Online crowd-sourcing was used to determine the subjective scores wherein ten subjects rated each sample. However, the database is not openly available for assessment.

- **The Speech in Noisy Environments Evaluation database (SPINE)** [3] is developed to facilitate robust speech recognition systems in noisy military scenarios. The repository consists of recorded conversations between two participants modelling the surrounding which depicts battleground environments.

- **The CoreSV14 [78]** was developed to evaluate various codecs such as AAC, Opus, Ogg Vorbis, and MP3 and find out which among these produces the best sound quality. It contains five speeches and thirty-five music excerpts.

- **The EQ-SQAM database** [79] contains about seventy high-quality sounds from music, speech, orchestra, etc. However, no subjective scores and degradations are present in the database.

- **The ACE Challenge [80]** was developed as a blind metric for the determination of acoustic parameters i.e. Direct-to-Reverberant Ratio and Reverberation Time from speech. The data set consists of clear speech sample recordings convolved with noise and acoustic impulse response measures from rooms of various sizes.

- **Bs1387Conform dataset [81]** is a small dataset consisting of about thirty-two instrumental sounds and speech samples. It has been used to validate the implementation PEAQ QA metric.

- **The UnbAvq2013 dataset [82]** comprises a total of twenty-four audio clips extracted from six audio-video samples. The clean files are corrupted by various codecs and bit rate compressions which are subjectively evaluated.

- Besides the datasets mentioned, researchers have utilized several other audio databases [83–87] spanning various sound domains.

From the literature survey above, we can iterate that most of the repositories consist of speech data. However, the UGM audio signals have different acoustic characteristics

as compared to plain speech content, and consequently, the quality assessment algorithm designed for speech signals can not be directly employed on UGM. In order to verify this argument, we carried out a spectrogram analysis of speech signals and UGM clips. The spectrogram aids in analyzing the range of frequency and amplitude present in a given signal. It visualizes a signal on two perpendicular axes representing the time and frequency, while the colour intensity represents the amplitude. Given a signal $s[m]$ with window w, such that Ω is the frequency and m is the discrete-time index. The Short-Time Fourier Transform (STFT) is expressed as:

$$\mathbf{STFT}\{s[m]\}(n,\Omega) = S(n,\Omega) = \sum_{m=-\infty}^{\infty} s[m]w[m-n]e^{-j\Omega m}. \tag{1.1}$$

Squaring the magnitude of the resultant STFT renders the spectrogram display of the power spectral density of the function as given by;

$$\mathbf{spectogram}\{s(t)\}(\tau,\Omega) = |S(\tau,\Omega)|^2. \tag{1.2}$$

where τ is the time axis.

Figure 1.4 represents the spectrogram of samples obtained from the three datasets NOIZUS [2], SPINE [3], and TIMIT [4] containing speech samples, respectively. It can be observed from the spectrograms that speech is represented by a restrained spectrum along with precisely defined predictable perceptual components. Additionally, the presence of phonemes is quite evident from the discontinuities (dark areas) present in the plots. Further, it can be observed that the spectrograms depict a restricted range of frequencies, due to the presence of the human voice, which encompasses a limited frequency span. The heat map depicting amplitude variation is quite constrained for the given samples. Thus, giving us the intuition that speech samples have limited diversity in frequency and amplitude. On the other hand, we find that the spectrograms of UGM audio 1.4 (d) and (e), are highly scattered. They do not contain only speech but there is music or other background sounds also present. Which makes the spectrograms different from those of the speech.

In consideration of the elaborate discussion, the following analysis can be derived:

1. Most of the existing datasets are designed for speech quality assessment and intelligibility analysis. These datasets lack other sounds that are dominant in UGM.

2. The existing datasets lack context diversity. Even though the datasets proposed in [72, 78, 82] consists of a combination of a few audio types, the amount of diversity in all these repositories is minimal.

3. The limited availability of a sufficiently large UGM dataset with subjective annotations has impeded the adoption of direct deep-learning techniques for UGM AQA.

1.3.2 Existing Audio Quality Assessment Techniques

Audio quality techniques can be divided into two main types: intrusive and non-intrusive. Intrusive metrics analyze the differences in the degraded and reference samples to calculate the perceptually weighted distance or error. On the other hand, non-intrusive metrics do not require the reference signal. Instead, they analyze only the degraded audio signal to estimate the quality score. A detailed study of the existing intrusive and non-intrusive quality assessment metrics is discussed below.

1. **Intrusive Quality Assessment Metric:**

   - **"Perceptual Evaluation of Speech Quality (PESQ)"** [88] is an intrusive metric used to estimate the quality of narrow-band speech samples. It involves aligning corresponding excerpts of the reference and test signals temporally and then analyzing them sample by sample. PESQ is used as a QA metric for a network or to evaluate the performance of specific network components.
   - **"Perceptual Objective Listening Quality Assessment (POLQA)"** [89] metric used to evaluate the quality of both narrow-band and super-wide band signals. This metric applies masking functions to the signals and conducts frequency domain analysis. The distortions, represented by the unmasked differences between the two signals, are aggregated and mapped to a quality score varying from 1 to 5.
   - **"The Virtual Speech Quality Objective Listener (ViSQOL)"** metric, proposed in the [90], relies on spectro-temporal analysis to assess the similarity between reference and distorted signals. This metric is specifically designed to handle quality issues commonly encountered in Voice over IP (VoIP) transmission and ensures robustness in evaluating VoIP-related quality problems.

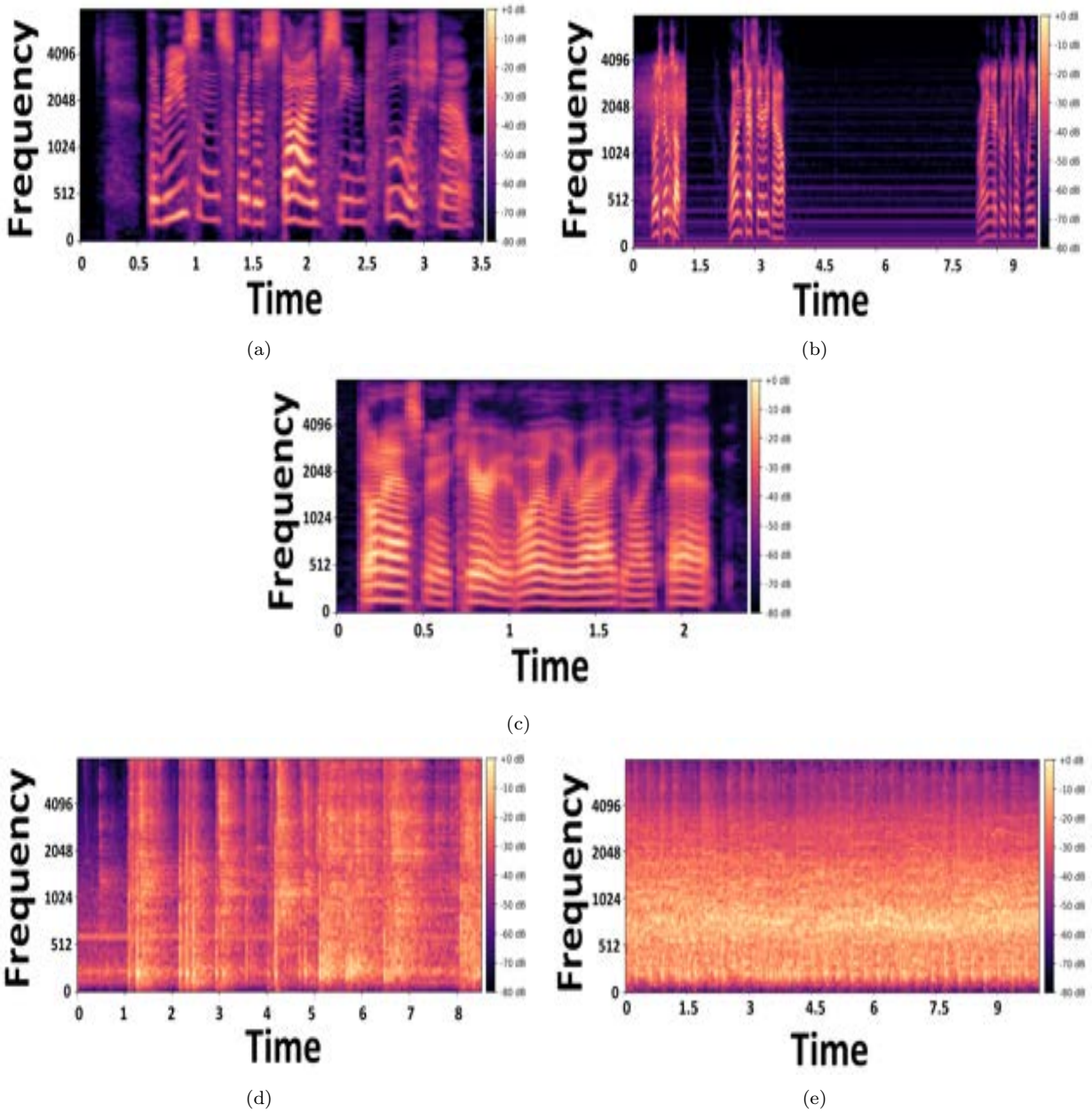


Figure 1.4: Spectrogram of the samples from (a) NOIZUS [2] (b) SPINE [3] (c) TIMIT [4], and (d), (e) are two random instances from IIT-JMU-UGM Audio Dataset.

- **"The Short-time Objective Intelligibility (STOI)"** [**91**] measure, is an intelligibility technique that utilizes DFT-based time-frequency decomposition to calculate an intermediate intelligibility measure. This approach allows for the assessment of intelligibility in a short-term context, enabling a more detailed analysis of speech signals.
- **"The Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)"** technique proposed by Jonathan *et al.* [92], introduced an enhanced technique for quality assessment in single-channel scenarios. Their approach is based on the "scale-invariant signal-to-distortion ratio" and is mainly developed for evaluating speech enhancement and source separation algorithms.
- **"The NISQA metric"** [**93**], introduced by Gabriel *et al.*, utilizes a mix of Convolutional Neural Networks with self-attention networks for speech quality assessment. This metric employs a multi-task neural network to get both the overall subjective score and the four speech quality dimensions.
- **The ViSQOL-v3** [**94**] metric utilizes gammatone spectrogram and neurogram similarity index measure for determining audio quality.
- In [**95**] the authors proposed a metric for identifying the just-noticeable difference between a given pair of audio samples. This metric is then mapped to provide a value of the perceptual distance between the samples. The audio samples considered in this metric encompass perturbations, including noise, reverberation, and compression artifacts.
- **The SESQA metric** [**96**] is a semi-supervised speech QA technique. It learns from existing labelled data, along with unlabeled or programmatically generated data to predict the quality of speech.

2. **Non-Intrusive Quality Assessment Metrics:**

- **"The Speech-to-Reverberation Modulation Energy Ratio (SRMR)"** is proposed by Falk *et al.* [97] for assessing the quality of de-reverberated and reverberant speech. SRMR utilizes spectral modulation analysis to quantify the "energy ratio" between the speech and reverberation components.
- **"The MOSA-Net metric "** [**98**] evaluates speech quality and intelligibility

using deep learning architecture together with cross-domain features to estimate PESQ, STOI, and SDI scores.

- **In the "Deep Noise Suppression Mean Opinion Score (DNSMOS)" metric [99]**, Log power Mel spectrogram features are utilized to train a "multi-stage self-teaching-based" metric. This technique is specifically developed to assess the performance of noise suppression algorithms.
- **"The NIST Signal to Noise Estimation Utility metric " [100]** computes the energy histogram to determine noise elements in a given signal.
- **The SNRVAD metric [101]** proposes to estimate the signal-to-noise ratio of speech signals by evaluating their amplitude distribution.
- **The MOSNet metric [102]** proposed by Chen *et al.* utilizes magnitude spectrogram of the voice samples as inputs to train convolution and Recurrent Neural Network inspired model for assessing voice conversion.
- **In the HPSSN [103] metric**, the authors investigate hierarchical and convolutional neural network approaches for predicting speech quality. These approaches are specifically applied to determine both the utterance-level and system-level quality scores for synthetic speech.
- **The MTF metric [104]**, suggested by Zhang *et al.* proposed a non-intrusive metric wherein they trained a CNN along with a pyramid BiLSTM-based architecture to perform the assessment. This architecture enables the model to capture temporal dependencies and learn relevant features directly from the time-domain speech signals.
- **In MBNet [105]**, a combination of a bias subnet and a mean subnet is proposed for accessing the opinion scores at an individual judge level and their average predictions. The model makes use of convolution layers with batch normalization to train the network.
- **The S3PRL metric [106]** , employs self-supervised pre-trained based models for predicting system and utterance-level quality. It also explores the use of range clipping, attention pooling, and segmental embedding modules in an end-to-end fashion to predict quality in data related to voice conversion.

- **The MetricNet [107]** employed label distribution learning along with speech reconstruction learning for speech quality assessment. It makes use of an Encoder architecture to obtain spectrograms using 1-D convolution layers. It further consists of blocks containing a series of dilated convolutional layers. The model predicts the PESQ as the ground-truth quality score.
- Apart from the above-described work, other quality metrics include [108–115].

Based on the literature analysis, it can be inferred that significant efforts have been dedicated to speech quality assessment and speech intelligibility. However, there has been considerably less focus on UGM audio quality assessment. As discussed earlier the UGM audio has different characteristics than speech data, thus the metrics do not perform well on the UGM samples.

Moreover, some existing metrics [93, 102] rely on spectrograms as input for the CNN model. However, due to the diverse nature of UGM audio, spectrograms are not suitable for UGM Quality Assessment. As a result, many of these techniques do not perform effectively when applied to UGM audio data. This can be inferred from the result analysis presented in Chapter 4.

Based on the analysis presented, it can be concluded that there is indeed a necessity for a QA metric specifically designed for UGM audio. This will enable better evaluation and enhancement of UGM audio content to meet the demands and expectations of users within the domain.

1.4 Performance Evaluation Metrics

In order to evaluate the performance of QA metrics, mainly four methods are used; Root Mean Square Error, Kendall rank correlation coefficient, Spearman Rank Correlation Coefficient, and Pearson Linear Correlation Coefficient. Each of these techniques is described below.

1. **"Pearson Linear Correlation Coefficient (r)":**

   It is a statistical technique to measure the direction and strength of the linear relationship between two variables. The range of the "r" coefficient is from -1 to 1. The 1 and -1 values represent a perfectly positive and negative correlation while zero represents no correlation. It is calculated as:

$$r = \frac{\sum(a_i - \bar{a})(b_i - \bar{b})}{\sum(a_i - \bar{a})^2(b_i - \bar{b})^2} \tag{1.3}$$

where, a_i and b_i are the individual data points of a and b, respectively.
$\bar{a}$ and $\bar{b}$ are the means of a and b, respectively.
$\sum$ denotes the summation operator.

2. **"Spearman Rank Correlation Coefficient ":**
The Spearman Rank Correlation Coefficient (ρ) is a statistical measure used to assess the strength and direction of the monotonic relationship between two variables. It is particularly useful when the relationship between variables is not necessarily linear but can be described by a monotonic function. The value of the coefficient ranges from -1 to 1, where 1 and -1 represent a perfect positive monotonic and perfect negative monotonic relationship respectively, and zero represents no monotonic relationship. SRCC is measured as:

$$\rho = 1 - \frac{6\sum(d_i^2)}{n(n^2 - 1)} \tag{1.4}$$

where d_i is the rank difference for each pair of data points.
n is the number of data points.

3. **"Kendall Rank Correlation Coefficient":**
Also called Kendall's tau (τ), is a statistical technique used to determine the direction and strength of the ordinal relationship between two variables. It measures the similarity of the orderings of the data points between the two variables instead of their specific numerical values. The coefficient ranges from -1 to 1, where 1 and -1 indicate perfect positive and negative concordance respectively, and zero denotes no concordance.

$$\tau = \frac{(A - B)}{(A + B)} \tag{1.5}$$

Here, B is the number of discordant pairs and A is the number of concordant pairs.

4. **"Root Mean Square Error" (RMSE):**
RMSE is a widely used metric to evaluate the accuracy of a model by calculating the difference between predicted values and true values. It provides an indication

of how well the model's predictions align with the true values. RMSE value can be calculated using the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N-1}(x_i - \hat{x}_i)^2}{N}} \tag{1.6}$$

Here, x_i are true values.

$\hat{x}_i$ are predicted values.

N is the number of data points.

A lower value of RMSE indicates that the predicted values of the model are closer to the actual values. In other words, it signifies that the model's predictions have smaller errors or discrepancies compared to the true values.

1.5 Outline of The Thesis

Considering the motivation from the above studies, this thesis proposed various quality assessment metrics for DIBR views and UGM Audio, which are discussed in the proceeding chapters. The overall outline of this thesis is given below;

- Chapter 2: Full Reference Quality Assessment Metric for DIBR Views
- Chapter 3: No Reference Quality Assessment Metric for DIBR Views
- Chapter 4: Non-Intrusive Audio Quality Assessment Metric for User-Generated Multimedia Using Deep Learning
- Chapter 5: Conclusion and Future Work

Chapter 2

Full Reference Quality Assessment Metric for DIBR Views

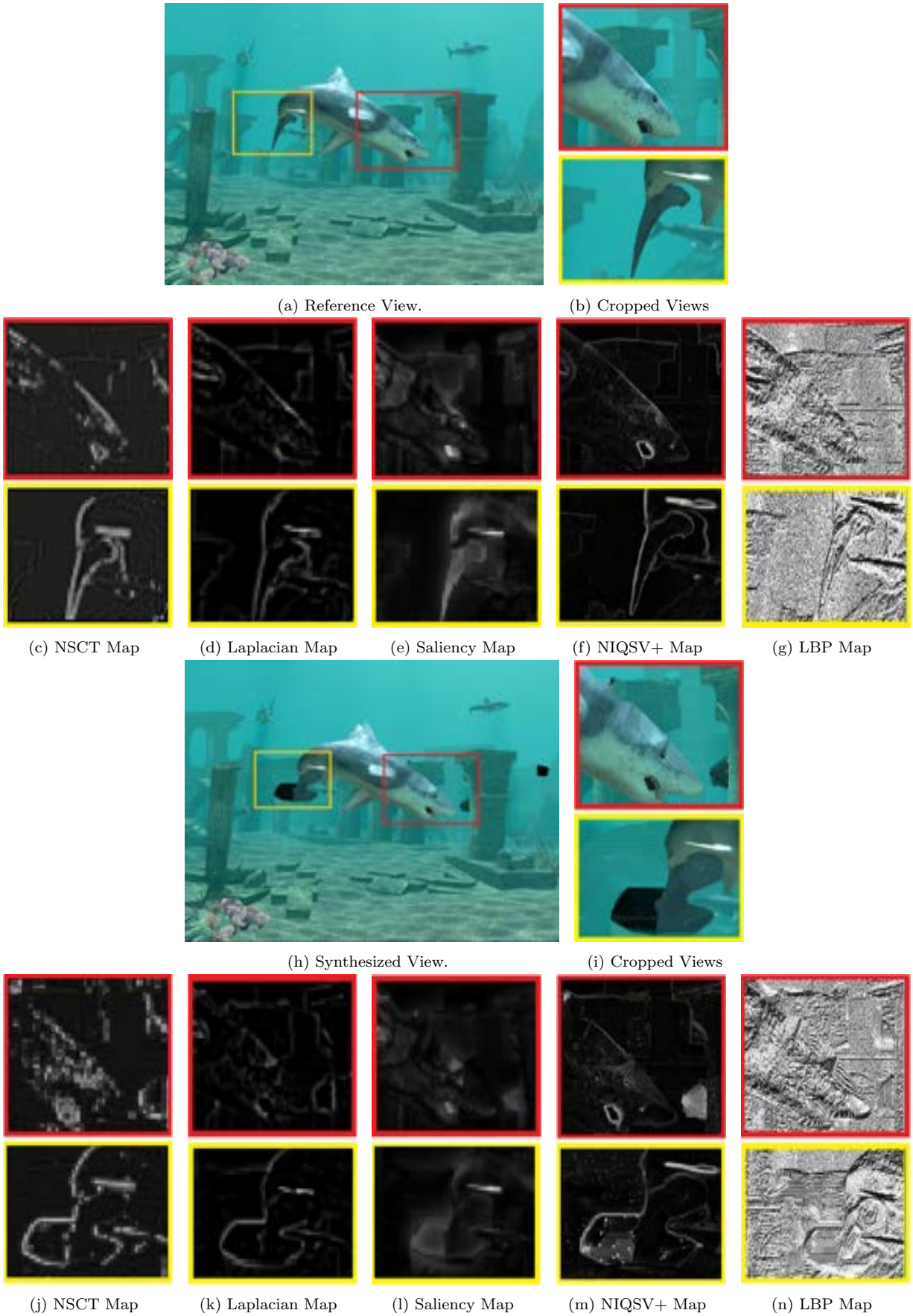
2.1 Introduction

In Chapter 1, the literature analysis points out that even though substantial research has been undertaken in the domain of DIBR QA, existing techniques are still unable to achieve high levels of accuracy for the benchmark DIBR datasets [18, 20]. The emergence of new inpainting techniques has effectively addressed certain degradation types like black holes, however, the imperfect rendering techniques introduce other distortions which significantly impact the quality of the images. Furthermore, because of the rendering processes involved in DIBR views, image misalignment occurs, posing a challenge for traditional 2D quality metrics. Thus, many full-reference techniques to deliver satisfactory performance. Given these considerations, there is a need for quality assessment algorithms that can effectively handle and quantify the various types of distortions present in 3D synthesized views. In this chapter, a comprehensive technique for quality assessment is presented which improves upon the drawbacks of the existing work.

2.2 Motivation

The motivation behind the proposed work is elaborated below:

1. The DF-CS metric [116] proposes a full reference QA metric wherein the deep fea-



(a) Reference View. (b) Cropped Views

(c) NSCT Map (d) Laplacian Map (e) Saliency Map (f) NIQSV+ Map (g) LBP Map

(h) Synthesized View. (i) Cropped Views

(j) NSCT Map (k) Laplacian Map (l) Saliency Map (m) NIQSV+ Map (n) LBP Map

Figure 2.1: Example Images from IETR Dataset, (a), (h) are the RGB DIBR reference and synthesized views, (b) and (i) are the cropped and zoomed-in patches of these views. Fig (d), (e), (f), and (g) show the reference's Laplacian Map, NSCT maps, NIQSV+ Map, and the LBP Map while, (j), (k), (l), (m) and (n) show the synthesized view's counterparts.

tures obtained from the Laplacian pyramids of the synthesized views and its reference view were compared to obtain the quality score. Various levels of the Laplacian pyramid are estimated using low-pass filtering and then down-sampling of the image. However, these two operations cause the loss of perceptually important information. To demonstrate this, Figure 2.1 (a) and (h) represent the DIBR reference and synthesized views, and the rectangular boxes 1(b) and 1(i) represent the zoomed-in view of the images for better visualization. Figures 1(d) and (k), show the zoomed-in patches obtained from the third level of the Laplacian pyramid of the synthesized, and reference view. From these figures, it can be observed that due to the low-pass filtering, texture information is lost in the synthesized view's Laplacian image which is vital for the identification of stretching artifacts. Thus, this information loss causes the low performance of the DF-CS metric.

2. The Saliency and Deep feature-based (Sal-DF) metric proposed in [117] is a full reference QA method that utilizes deep features obtained from the saliency maps of both the synthesized and reference views. The similarity value between their respective deep features is then calculated to determine the quality score of the view. Saliency maps highlight the most prominent or salient regions in an image [118]. However, these maps are not effective in capturing texture information in the synthesized views. As a result, they are inadequate in accurately predicting perceptual quality. To demonstrate this point, Figure 2.1 (e) and (l) showcase the saliency maps of the DIBR reference and synthesized views. As depicted in the figures, although saliency maps can identify prominent areas and edges, they do not provide substantial information about the texture in regions where distortions are present. Consequently, the inability of saliency maps to include all distortion-prone areas and low-frequency information in the distorted regions leads to a decrease in performance.

3. Furthermore, quality assessment techniques such as NIQSV+ highlight the textures present in an image as shown in Figures 1(f) and 1(m). Also, quality assessment metric [119] makes use of Local Binary Pattern (LBP) to extract features. Figures 1(g) and 1(n) represent LBP feature maps that highlight textures in the images. As depicted from these figures, extensive details are present in the feature maps, however, all these details are not necessary for the QA of DIBR views as distortions

are only present in a few areas. Thus, resulting in the low performance of these metrics.

4. Conversely, the analysis reveals that the Non-Subsampled Contourlet Transform (NSCT) maps offer feature maps of DIBR views that are rich in information. Figures 1(c) and (j) illustrate the NSCT maps, which, in comparison to other maps, effectively preserve the texture details. These NSCT maps thus provide information-rich feature maps that are crucial for the HVS.

Motivated by the limitations observed in the existing QA metrics, the following proposed metric introduces a FR technique that takes advantage of the NSCT coefficients for QA.

2.3 Proposed Full-Reference Quality Assessment Metric

In this work, a technique is proposed for estimating the quality of DIBR/3D views by taking into consideration the implicit frequency domain characteristics of the image using NSCT. The capabilities of existing deep-learning models are leveraged to extract deep features, which are then used to determine the quality of the synthesized view. The proposed model is based on two parts: the extraction of NSCT maps and the subsequent extraction and comparison of deep features. The detailed architecture of the model is depicted in Figure 2.2.

2.3.1 Extraction of NSCT Maps

In recent years, various image transformation techniques have emerged to highlight different image features in different transform domains. Among these techniques, the Non-Subsampled Contourlet Transform has proven to be a powerful method for 2D image representations [5]. In the context of 3D synthesized images, distortions primarily occur near object edges or contours due to occlusions, and these distortions are highly perceptible to the human visual system. The NSCT employs a dual filter bank structure that effectively captures these visually significant contours present in an image. Unlike the Laplacian Pyramid used in other methods, the NSCT utilizes a non-subsampled pyramid

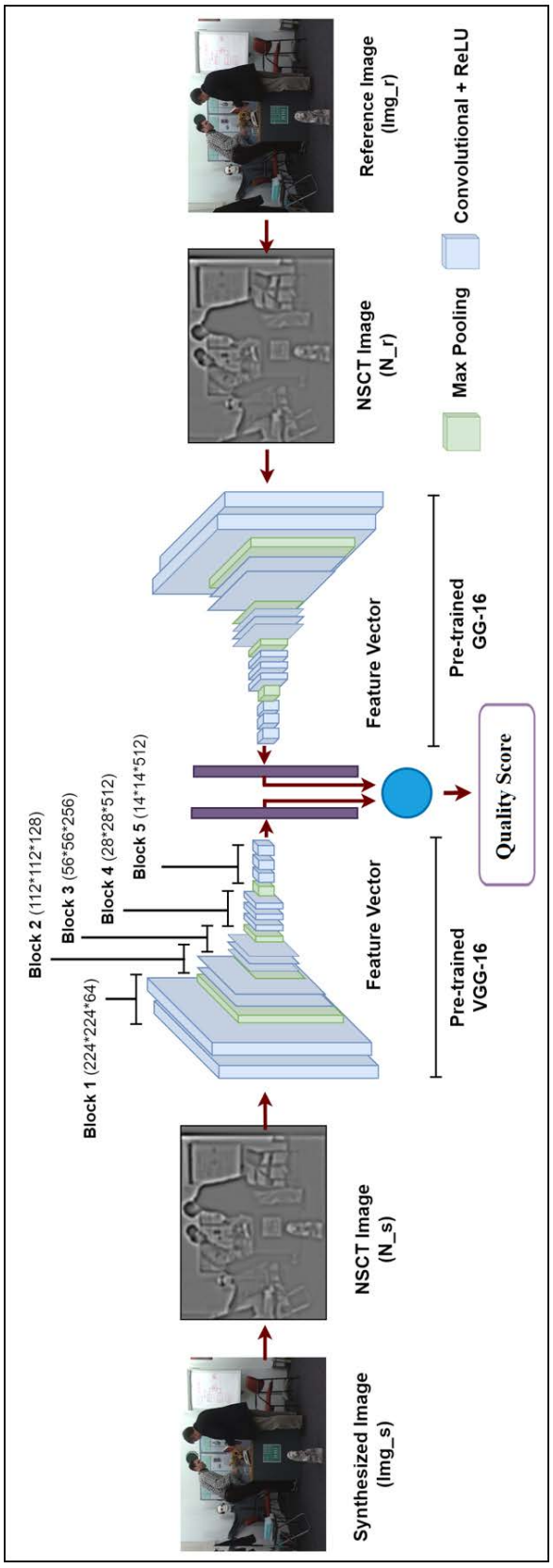


Figure 2.2: Workflow of the proposed full reference quality assessment model.

structure to maintain multi-scale characteristics, while a non-subsampled directional filter bank (NSDFB) is employed for directionality. Consequently, the NSCT exhibits properties such as multi-resolution, directionality, and shift-invariance. These transforms have utility in various domains, such as texture classification and image enhancement. Additionally, the NSCT decomposition does not involve any down-sampling or up-sampling operations, thereby preventing frequency aliasing in low-frequency sub-maps.

Non-subsampled Pyramid Structure Filter Banks (NSPFB) followed by NSDFB, form the basis of NSCT decomposition (Fig. 2.3) [5]. For i number of decomposition levels, $i+1$ sub-bands consisting of one low and i high-frequency sub-bands are obtained by Non-subsampled pyramid (NSP). Further, these high-frequency sub-bands at each level are then decomposed by NSDFB into directional sub-bands. These resultant sub-bands have the size same as that of the original image. Finally, the resultant filters of the i^{th} level cascading NSP are represented as [120]:

$$X_n(z) = \begin{cases} X_1(z^{2^{n-1}}) \prod_{j=0}^{n-2} X_0(z^{2^j}) & 1 \leq n < 2^i \\ \\ \prod_{j=0}^{n-1} X_0(z^{2^j}) & n = 2^i \end{cases} \tag{2.1}$$

where, z^j stands for $[z_1^j, z_2^j]$. In the proposed model, the NSCT maps corresponding to Scale 2 and directional decomposition equal to two are used. The depiction of various NSCT levels along with their directional decompositions for an image in the IETR dataset is given in Fig. 2.4.

2.3.2 Deep-feature Extraction

Following the acquisition of NSCT decompositions, these features are subsequently inputted into a pre-trained VGG-16 neural network [121]. The use of VGG-16 offers several advantages over other pre-trained networks, as demonstrated in various existing quality assessment studies [116, 117, 122]. These deep neural networks consist of cascading convolution layers capable of learning representative features from input images. Furthermore, the DIBR synthesized views often suffer from image misalignment when compared to reference views. In the state-of-the-art (SOTA) IQA techniques, explicit methods are employed during image pre-processing to compensate for this misalignment. However, CNN models also possess the capability to provide robust feature representation implic-

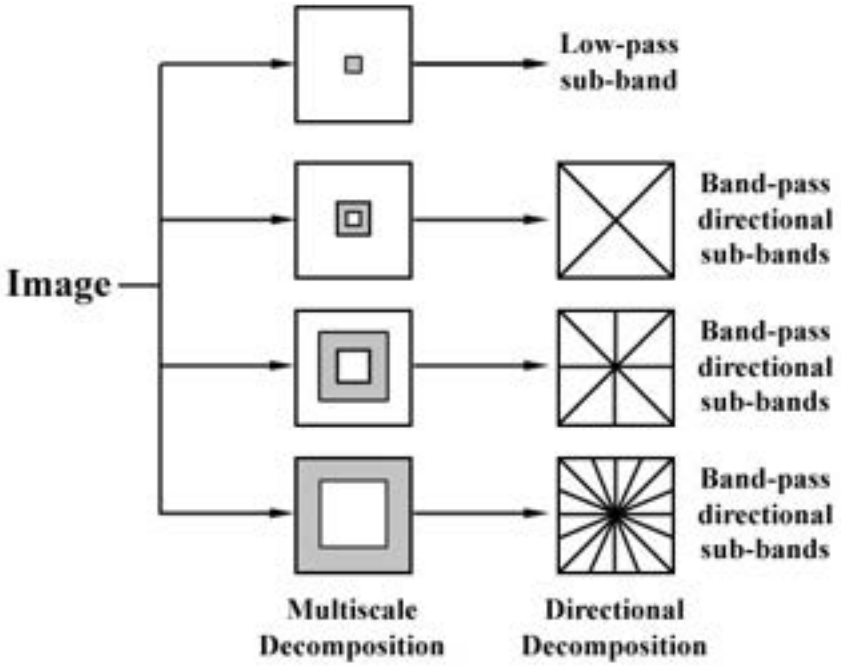


Figure 2.3: NSCT Decomposition [5].

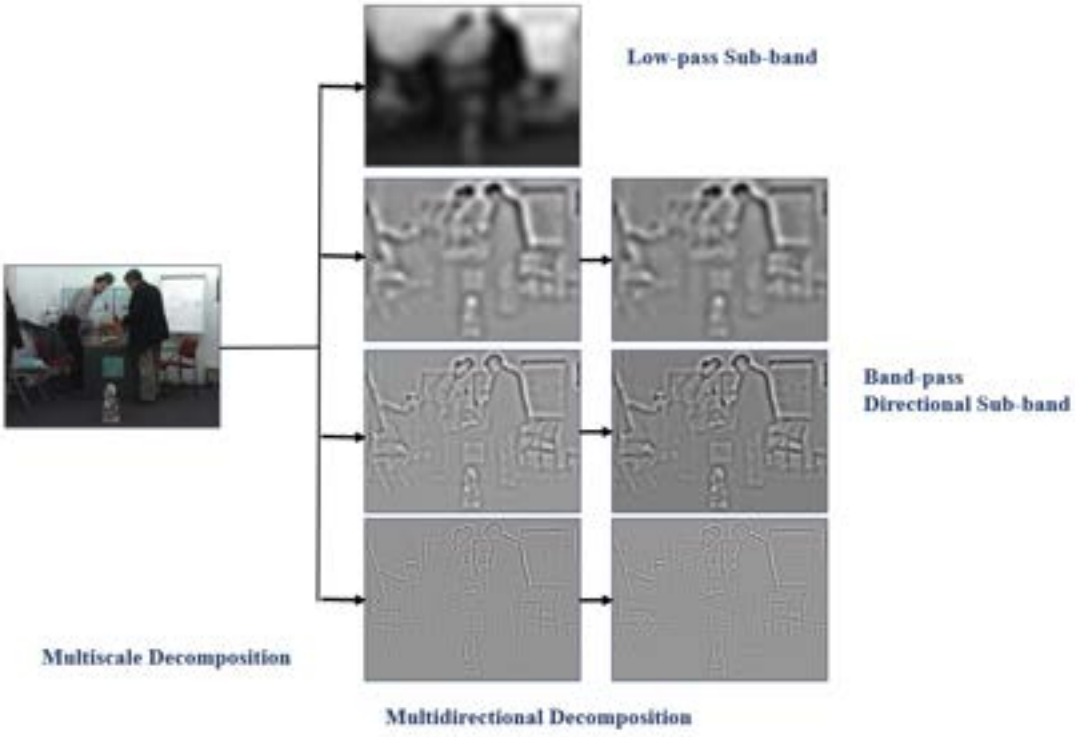


Figure 2.4: Example of NSCT Decomposition of a DIBR view.

itly, thereby exhibiting resilience to misalignment between the synthesized and reference views. Consequently, CNN models serve a dual purpose: efficient extraction of quality-aware features and addressing image misalignment.

The VGG-16 architecture has gained popularity and achieved remarkable performance in the field of computer vision. Its architecture consists of five blocks with the number of filters within each block corresponding to 64, 128, 256, 512, and 512, respectively. The main characteristic of VGG-16 is the repeated use of small-sized convolutional filters (3x3) throughout the network. This design choice allows for deeper and more efficient feature representation. An overview of the architecture is given below:

1. Input Layer: The input layer of VGG-16 takes an image as input.

2. Convolutional Layers (CL): It consists of 13 (CL) with a Rectified Linear Unit as the activation function. These layers are trained for learning hierarchical feature representations from the given training images.

3. Max Pooling Layers (MPL): There are 5 MPLs, which help to compress the spatial dimensions of the feature maps and preserve vital information.

4. Fully Connected Layers (FCL): The CL and MPLs are followed by three fully connected layers with the ReLU activation function.

5. Softmax Layer: The last layer is a softmax layer of 1000 units in size, reflecting the number of classes in the ImageNet dataset [123].

In the proposed approach, the feature vector is acquired from the third convolution layer of Block 5 within the VGG16 model. This feature vector is subsequently flattened into a 1-dimensional representation, providing a comprehensive and quality-aware feature vector of the image. In order to calculate the quality score, the feature vectors of both the reference and synthesized views are normalized and then the difference between the two vectors is determined by using the following equation:

$$Q' = 0.5 \sum_{k=1}^{K} R(k) \ln \frac{R(k)}{T(k)} + 0.5 \sum_{k=1}^{K} S(k) \ln \frac{S(k)}{T(k)} \tag{2.2}$$

where, $S(k)$ and $R(k)$ represent the k^{th} element of the feature vectors of the synthesized and reference view, respectively. Also, K represents the length of the feature vector and

$T(k) = \frac{1}{2}(R(k) + S(k))$. A quality score denoted as Q', close to zero indicates high similarity between the feature vectors of the synthesized and reference views, indicating better perceptual quality. Whereas, a value close to 1 suggests dissimilarity between the feature vectors, indicating poorer image quality.

Table 2.1 presents an initial quantitative performance analysis of different algorithms applied to extract feature maps. As previously mentioned, the DF-CS and Sal-DF metrics employ features from the Laplacian image and Saliency Maps, respectively. In the proposed approach, deep features from NSCT images for quality assessment are utilized. Similarly, we make use of the NIQSV+ and LBP maps as input to VGG-16 for deep feature extraction, followed by the calculation of quality scores using Equation 3.1. The table demonstrates that the proposed method surpasses the other techniques in terms of Root Mean Square Error, Pearson Linear Correlation Coefficients, Spearman's Rank Order Correlation Coefficients, and Kendall rank correlation coefficient. This outcome validates our initial motivation that NSCT maps provide a rich quality-aware feature representation for DIBR views. Furthermore, to improve the performance of our proposed FR metric, a pooling technique with the BIQI metric [51], is also used which is explained in detail in Chapter 3.

2.4 Result Analysis

To analyze the performance of the proposed work, various datasets, tools, and techniques are employed. An elaborate analysis of our work is discussed in this section.

2.4.1 Evaluation Criteria

As described in Chapter 1, to assess the performance of the proposed method in comparison to SOTA QA algorithms, the standard correlation evaluation metrics are used, which are; r, ρ, τ, and RMSE. An efficient metric has higher values for r, ρ, and τ, indicating stronger correlations, while lower values close to zero are desirable for RMSE.

To decrease the non-linearity of objective prediction scores, for efficiently mapping them to subjective ratings, a nonlinear logistic function is used [50]:

$$k(s) = \alpha_1(\frac{1}{2} - \frac{1}{1 + e^{\alpha_2(s-\alpha_3)}}) + \alpha_4 s + \alpha_5 \tag{2.3}$$

Table 2.1: Performance comparison of the Proposed NSCT-FR metric when different feature extraction algorithms are used (such as Saliency, Laplacian, NIQSV+, LBP, and NSCT maps).

S. No.	Technique	r	ρ	τ	RMSE
1.	**Proposed NSCT-FR**	**0.8207**	**0.8187**	**0.6203**	**0.1417**
2.	DF-CS [116]	0.7848	0.7676	0.5753	0.1537
3.	Sal-DF [117]	0.7620	0.7513	0.5542	0.1605
4.	Deep NIQSV+	0.7238	0.7176	0.5272	0.2255
5.	Deep LBP	0.3629	0.3606	0.2357	0.2325

here $\alpha_y, (y \in 1, 2, 3, 4, 5)$ represents the parameters to be fitted, while $k(s)$ depicts the subjective score to which the objective score s is mapped.

2.4.2 Evaluation Dataset

To evaluate the proposed metric two benchmark DIBR datasets are used i.e. the IETR dataset [18] and the IVY dataset [20] are used. The IETR dataset consists of ten reference views, from which 140 synthesized images are generated using different rendering techniques. For each view, its Differential Mean Opinion Score (DMOS) is available, which serves as the ground truth quality score. These DMOS values provide a reliable basis for performance evaluation of the proposed model against subjective evaluations.

The publicly available IVY dataset [20] contains stereo images and consists of seven pairs of left and right reference views. These are accompanied by their respective synthesized views, generated using various DIBR synthesizing techniques. To evaluate the proposed metric, the quality score for both the right and left views is determined. Subsequently, these scores are averaged to get the quality value for each view.

2.4.3 Performance Analysis

To assess the performance of the proposed full reference metrics on the IETR dataset, a comparative analysis was conducted with various SOTA IQA metrics. The results are presented in Table 2.2. These include QA metrics for 3D views as well as those applicable to natural images. The proposed FR metric, achieved impressive results, with r, ρ, τ, and RMSE values of 0.8207, 0.8187, 0.6203, and 0.1417, respectively. In comparison,

Table 2.2: Performance comparison of the proposed FR metric with various FR objective quality metrics on the IETR dataset. The '-' symbol depicts that the data is not available and the ' " ' symbol denotes "same as above".

S. No	Technique	Aimed For	r	ρ	τ	RMSE
1.	**Proposed NSCT-FR**	**3D views**	**0.8207**	**0.8187**	**0.6203**	**0.1417**
2.	SEQUSS [43]	3D views	0.8030	0.8020	0.6000	0.1470
3.	PUIR-DFCS [116]	"	0.7965	0.7909	0.5992	0.1499
4.	CII [41]	"	0.7707	0.7572	0.5700	0.1580
5.	Sal-DF [117]	"	0.7620	0.7513	0.5542	0.1605
6.	Wang's [45]	"	0.7446	0.7223	-	0.1610
7.	MLFA [124]	"	0.7378	0.7036	-	0.1899
8.	A-DISTS [44]	"	0.705	0.686	0.499	-
9.	SSPD [35]	"	0.7020	0.6850	-	0.1790
10.	Tian's [30]	"	0.6685	0.5903	-	0.1844
11.	LGGD+ [42]	"	0.6650	0.6580	-	0.1930
12.	LPIPS [122]	Natural images	0.6659	0.6144	0.4386	0.1850
13.	SC-IQA [34]	3D views	0.6620	0.5960	-	0.1850
14.	LOGS [32]	"	0.6280	0.6160	-	0.1930
15.	Sui's [125]	"	0.673	0.593	-	0.183
16.	MP-PSNR [31]	"	0.6190	0.5809	0.3802	0.1947
17.	PSNR	Natural images	0.6012	0.5809	0.4024	0.1985
18.	FSIM [67]	Natural images	0.6052	0.4755	0.3243	0.1973
19.	GMSD [66]	Natural images	0.5560	0.4787	0.3257	0.2070
20.	MW-PSNR [29]	3D views	0.5389	0.4875	0.3364	0.2088
21.	Cheon's [65]	Natural images	0.4644	0.4416	0.3151	0.2882
22.	Li's [126]	Depth images	0.4584	0.4304	0.3009	0.4155
23.	Lao's [64]	Natural images	0.4266	0.4458	0.3040	0.3062
24.	SSIM [9]	Natural images	0.4016	0.2395	0.2647	0.2275
25.	SIQE [127]	3D views	0.3144	0.3418	-	0.2353
26.	DSQM [128]	"	0.2977	0.2369	-	0.2367

the best-performing existing metric, SEQUSS [43], attained r, ρ, τ, and RMSE values of 0.8030, 0.8020, 0.6000, and 0.1470, respectively. This indicates that the proposed FR metric outperforms SEQUSS, with improvements of approximately 2.2%, 2.08%, and 3.3% in terms of r, ρ, and τ respectively. These results depict the superior performance of the proposed FR metric, highlighting its ability to achieve higher correlation and lower error in quality assessment compared to the SOTA metric.

The results of the proposed FR metric for the IVY dataset [20] is provided in Table 2.3. According to the table, the proposed FR metric achieves r, ρ, τ, and RMSE values of 0.7580, 0.7375, 0.5418, and 9.4090, respectively. On the other hand, the corresponding values for the best-performing FR metric i.e. SSPD are 0.6892, 0.6814, 0.4872, and 10.3210, respectively. Therefore, the proposed FR metric demonstrates notable performance improvement compared to SSPD. Specifically, there is an increase of 9.9%, 8.2%, and 11.2% in terms of r, ρ, and τ respectively, highlighting the superiority of the proposed metric.

2.4.4 The Statistical Significance (SS) Test

Further, the SS-Test or F-Test is adopted for analyzing the performance of the proposed model. The test is conducted between the quality scores acquired from the proposed method and those acquired by employing various IQA metrics. The F score is calculated as [116];

$$F = \frac{g_{\beta_1}^2}{g_{\beta_2}^2} \tag{2.4}$$

where, (β_1, β_2) are the scores obtained by the objective metric and the proposed algorithm that are tested and g_{β_1}, g_{β_2} represent their respective RMSE. This test works on the variance-based hypothesis, wherein, the value '+1' depicts that the proposed model is statistically better than the other metric, 'zero' represents that the two methods are similar, and '-1' represents the inferiority of the proposed metric to other metrics. Table 2.4 depicts the F-score between the proposed FR and six IQA metrics. These scores are +1 for all the methods, thus depicting that our method is statistically better in comparison to other IQA metrics (confidence interval equal to 90%).

Table 2.3: Performance comparison of the proposed NSCT-FR metric with various FR objective quality metrics on the IVY dataset. The '-' symbol depicts that the data is not available.

S. No	Technique	Aimed For	r	ρ	τ	RMSE
1.	**Proposed NSCT-FR**	**3D views**	**0.7580**	**0.7375**	**0.5418**	**9.409**
2.	SSPD [35]	"	0.6892	0.6814	0.4872	10.3210
2.	CII [41]	"	0.6726	0.6547	-	-
3.	LOGS [32]	"	0.6442	0.6385	0.4509	18.8549
4.	LTG [129]	Natural images	0.6214	0.6072	0.4337	19.3139
5.	IDEA [33]	3D views	0.6311	0.6132	0.4405	19.0379
6.	MP-PSNR [31]	"	0.6114	0.5954	0.4217	19.0379
7.	FSIM [67]	Natural images	0.6118	0.5975	0.4223	19.4998
8.	Bosc [19]	3D views	0.6196	0.6046	0.4246	19.3497
9.	GSM [130]	Natural images	0.5736	0.5805	0.4068	20.1934
10.	SSIM [9]	Natural images	0.5684	0.5662	0.4068	20.2826
11.	MW-PSNR [29]	3D views	0.5240	0.5051	0.3528	20.9969
12.	RMW-PSNR [29]	"	0.5224	0.5008	0.3540	21.0200
13.	VSQA [131]	"	0.5012	0.5034	0.3118	22.5648
14.	VIF [132]	Natural images	0.4013	0.3958	0.2685	22.5802
15.	3DSwIM [133]	3D views	0.3338	0.33764	0.2249	23.2375

Table 2.4: Results of the F-Test conducted between the proposed NSCT-FR metric and the various SOTA metrics on the IETR dataset.

Metric	**PUIR-DFCS**	**SSPD**	**LPIPS**	**SI-DL**	**DSCB**	**APT**
Score	+1	+1	+1	+1	+1	+1

2.4.5 Scatterplot Analysis

To enhance the visual interpretation of the results, scatterplots are generated to illustrate the correlation between the DMOS and the quality scores obtained from various SOTA methods. These methods include FR metrics such as SSPD [134], PUIR-DFCS, and LPIPS [122]. As well as NR IQA metrics such as NIQSV+ [53], DSCB [69], KRR [52], NIQE [135], BRISQUE [136], BIQI [51], Highgrade [137], and HyperIQA [138]. Figure 2.5 showcases these scatterplots. The scatterplots clearly demonstrate a strong linear relationship between the proposed full-reference metric and the objective scores, surpassing the performance of the other techniques. This observation indicates that the proposed model exhibits a higher consistency with the HVS. The visual representation provided by the scatterplots reinforces the effectiveness and reliability of the proposed metric in accurately assessing image quality.

2.4.6 Ablation Study

Furthermore, a series of experiments are conducted by varying the parameters and decomposition types to get an elaborate ablation study of the proposed model.

2.4.6.1 Analysis of NSCT Level

Firstly, the impact of using different NSCT multi-scale maps and directional decompositions on the performance of the proposed FR technique is analyzed. The results are presented in Table 2.5. The performance analysis reveals that the NSCT maps of the second level produced the best results, indicating the significance of both high and low-frequency components in extracting efficient features. Furthermore, there is only a slight variation in performance across different levels, suggesting that the model is not heavily reliant on the specific NSCT level number. Thus, the ablation study sheds light on the importance of the NSCT multi-scale maps and the effectiveness of the proposed FR tech-

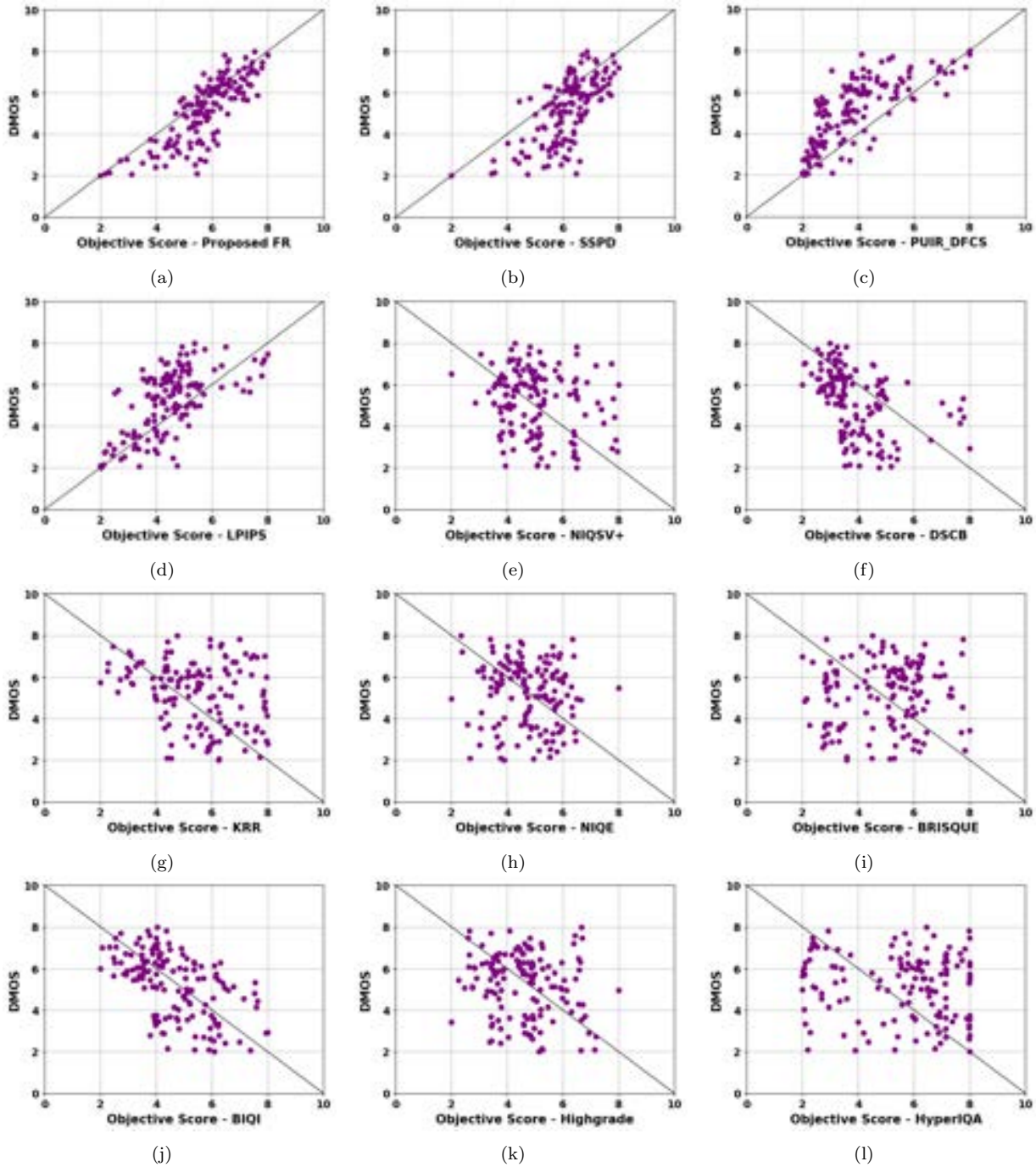


Figure 2.5: Scatter Plot of subjective score/DMOS values and objective scores of SOTA metrics on IETR dataset.

Table 2.5: Performance analysis of the proposed NSCT-FR model by varying the NSCT scales and orientations on the IETR database.

S. No.	NSCT layer	Orientation	r	ρ	τ	RMSE
1.	Layer 1	1	0.7866	0.7687	0.5770	0.1531
2.	Layer 1	2	0.8014	0.7921	0.5924	0.1443
3.	Layer 2	1	0.8131	0.8095	0.6132	0.1443
4.	**Layer 2- Proposed**	2	**0.8207**	**0.8187**	**0.6203**	**0.1417**
5.	Layer 3	1	0.7920	0.7919	0.5940	0.1513
6.	Layer 3	2	0.7701	0.7667	0.5743	0.1581

nique. It demonstrates that the model can achieve robust performance while exhibiting flexibility in handling different NSCT scales.

2.4.6.2 Analysis of The Backbone Deep Learning Model

An ablation study is conducted wherein, different pre-trained models are used as the backbone neural network for the extraction of features in the proposed metric. As presented in Table 2.6, a range of well-known CNN models including VGG-16 [121], Inception V3 [139], MobileNet [140], SqueezeNet [141], AlexNet [142], and Inception V4 [143] were utilized. Additionally, the performance of various transformer-based models such as Vision Transformer [144], BEIT [145], Swin Transformer [146], BIT [147], and DEIT [148] were assessed for quality assessment purposes. A brief summary of these deep learning models is given below:

1. **Inception V3 [139]:**
   It is a CNN proposed by Google in 2015 for image classification, it has a total of 42 layers. It features inception modules with parallel convolutions of varying sizes, for multi-scale feature extraction.

2. **MobileNet [140]:**
   This architecture is used for real-world applications. It leverages depthwise separable convolutions, a technique that replaces standard convolutions used in earlier architectures, to construct lightweight models.

3. **SqueezeNet [141]:**

It is a compact CNN architecture designed to have a small model size while maintaining competitive accuracy. It achieves this by employing "fire modules" that use a combination of 1x1 and 3x3 convolutions, reducing the number of parameters.

4. **AlexNet [142]:**

   This model includes convolution and fully connected layers utilizing ReLU activation, local response normalization, and dropout regularization.

5. **Inception V4 [143]:**

   It is a CNN architecture that improves upon its predecessor, Inception V3. It introduces enhancements such as varied kernel sizes, additional convolutional layers, and residual connections for improved performance and gradient flow.

6. **Vision Transformer [144]:**

   It is a DL architecture that applies the transformer model to image recognition tasks. It treats images as sequences of patches and uses self-attention mechanisms to capture global and local dependencies. By leveraging the transformer's capability to model long-range dependencies, it has achieved competitive performance on various image classification benchmarks.

7. **BEIT [145]:**

   Big Transfer (BEIT) is a vision transformer with improved performance and scalability. It introduces the concept of "big transfer" by pre-training the model on a large repository, similar to BERT in natural language processing.

8. **Swin Transformer [146]:**

   Its architecture addresses the limitations of traditional vision transformers in handling large-scale image recognition tasks. It introduces a hierarchical structure that divides the input image into smaller windows, enabling efficient processing of large images. It utilizes shifted windows and shift-based position encoding to capture spatial relationships effectively.

9. **BIT [147]:**

   Big Transfer with Transformers (BIT) is a vision transformer architecture that combines the hierarchical structure of CNNs with the self-attention mechanism of transformers to capture global and local features.

Table 2.6: Performance evaluation of varying the pre-trained deep learning model in the proposed NSCT-FR metric.

S. No.	Model	r	ρ	τ	RMSE
1.	**VGG-16 (Proposed NSCT-FR)**	**0.820**	**0.818**	**0.620**	**0.141**
2.	DEIT	0.721	0.721	0.532	0.423
3.	Xception	0.711	0.697	0.509	0.174
4.	SqueezeNet	0.687	0.742	0.559	0.376
5.	Alexnet	0.615	0.708	0.527	0.462
6.	Inception V3	0.703	0.686	0.512	0.176
7.	Inception V4	0.702	0.682	0.499	0.176
8.	Swin Transformer	0.699	0.700	0.517	0.387
9.	MobileNet	0.678	0.678	0.495	0.182
10.	Vision Transformer	0.668	0.647	0.471	0.387
11.	BIT	0.662	0.660	0.493	0.320
12.	BEIT	0.656	0.68	0.493	0.483

10. **DEIT [148]:**

    Data-efficient Image Transformer (DEIT) is a vision transformer architecture that introduces distillation techniques to leverage the knowledge from large-scale pre-training datasets and transfer it to smaller datasets.

Table 2.6 clearly illustrates that VGG-16 achieves the highest performance among both CNN architectures and transformer architectures for quality assessment. Its superior performance aligns with existing literature [121] and further underscores its effectiveness in feature extraction for quality assessment tasks.

2.5 Conclusion

This work introduces some of the techniques for image quality assessment, specifically focusing on DIBR views. In this regard, a full reference QA technique is proposed. Through the experimental analysis, it is analyzed that the maps obtained by applying the Non-Subsampled Contour Transform offer a rich representation of quality-aware features in DIBR views. To enhance the quality assessment process, further refinement of these

maps is done by extracting their deep features. The results obtained from the proposed algorithms demonstrated significant improvements over existing DIBR quality assessment algorithms.

Chapter 3

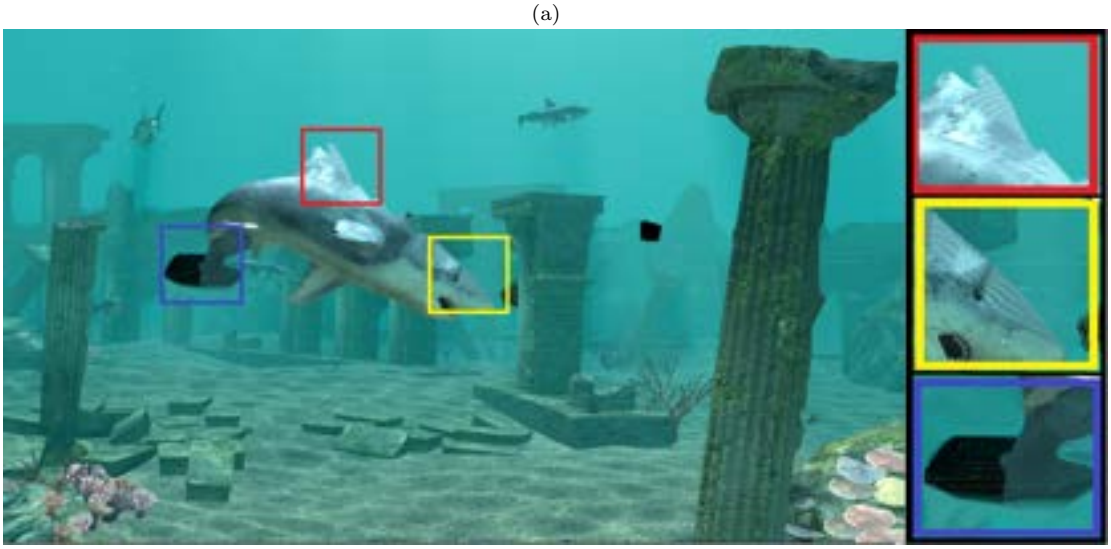
No Reference Quality Assessment Metric for DIBR Views

3.1 Introduction

In Chapter 2, a full reference technique for quality assessment was introduced. While FR techniques are known for their efficiency, in certain scenarios like Free Viewpoint Video, the reference image may not be available. In such scenarios, no-reference (NR) techniques are often employed as an alternative approach to quality assessment. Thus, leveraging the capability of the earlier proposed FR model, this work extends its usage to a no-reference metric. The motivation for the proposed NR metric is listed below:

1. From the literature, it is analyzed that many of the previous studies such as [149], [127], and [150] employ the concept of block-based deep-learning models. In this technique, the deep learning model is trained on some patches cropped from the image for QA. One of the limitations of such techniques is that they consider the subjective score of the entire image as the ground truth for individual image blocks. This assumption is valid for natural images where distortions such as blur, noise, etc. are evenly distributed across the image. However, in the case of DIBR views, this assumption is not applicable. This is due to the fact that the distortions in DIBR views predominantly occur along object boundaries due to imperfect rendering and inpainting techniques. Thus the quality of the image blocks is not uniform and the MOS of the whole image may not be representative of the quality of each individual block. This can be analyzed from some examples for the IETR dataset depicted

(a)



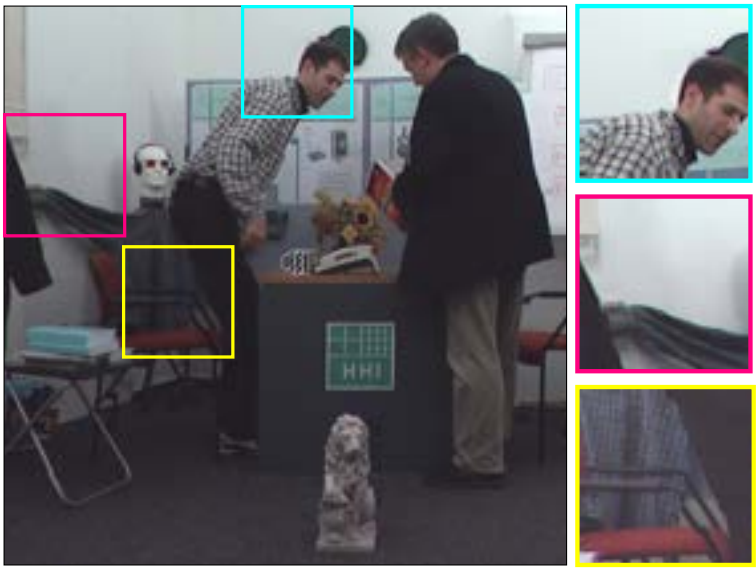
(b)

Figure 3.1: Examples of a Reference and synthesized view from IETR dataset, highlighting blocks with distortions.

in Fig. 3.1 (a) and (b). The blocks represent the zoomed-in view of some of the areas of the images with distortions. As depicted in these illustrations, the geometric distortions are localized and concentrated near object boundaries, rather than being universally present throughout the image. Therefore, this reinforces our argument that using the direct MOS of the entire image may not accurately represent the quality score at the block level.

2. Furthermore, in SI-DL metric [6], the authors explored the relationship between the DIBR view's quality and the occurrence of patches containing stretching artifacts within that view. They made the assumption that different levels of geometric distortions have a uniform impact on image quality. To identify such blocks, they

(a) Example View 1



(b) Example View 2

Figure 3.2: Examples of the image blocks detected as 'distorted' blocks by the SI-DL algorithm [6].

employed a binary classifier based on a convolutional neural network. While this metric successfully detects blocks with geometric distortions, it does not quantify the severity of distortions in each block. Additionally, it is analysed this assumption does not hold true, as the effect of geometric distortions on perceptual quality depends on the magnitude of the distortions. To support this argument, examples of blocks detected with geometric distortions using the SI-DL metric are illustrated in Fig. 3.2. It is evident that these blocks exhibit varying degrees of distortion, yet they are all classified just as "distorted" without considering the severity of the distortions, which significantly influences the overall perceptual quality.

From the aforementioned study, it can be inferred that block-based learning can be explored for DIBR view QA. Furthermore, estimating the frequency and intensity of distorted blocks in the image can give valuable insights into its perceptual quality. The proposed approach provides a promising method for accurately calculating quality by considering the distortions of individual blocks within an image.

3.2 Proposed No-Reference Quality Assessment Metric

This section explains in detail the proposed NR QA metric for DIBR views based on determining the block-level quality of the given image. The main steps in the model are given below:

3.2.1 Block Level Ground-truth Quality Score Estimation

In Chapter 2, the NSCT-FR quality metric is introduced, which demonstrates the capability to efficiently predict the quality score of DIBR views. Given the effectiveness of the NSCT-FR metric in assessing overall quality, it can be inferred that the algorithm also possesses the capability to evaluate the perceptual quality of individual blocks within these views. Thus, in the proposed work, similar to the NSCT-FR model, the non-subsampled contourlet transform maps are extracted from the DIBR views. Next, to determine the ground truth quality score at the block level, the DIBR view's NSCT map is cropped into blocks of size $n \times n$ using a sliding window technique, resulting in a total of M blocks. The

workflow for this step is illustrated in Figure 3.3. The blocks from both the reference and synthesized views are simultaneously inputted into the pre-trained CNN model (VGG-16) to extract features. The output from the last layer of the fifth block of each network is then flattened to obtain a 1-D feature vector. As done in the NSCT-FR, the quality score for each block/patch is obtained by calculating using the equation:

$$Q'_B = 0.5 \sum_{k=1}^{K} R_B(k) \ln \frac{R_B(k)}{T_B(k)} + 0.5 \sum_{k=1}^{K} S_B(k) \ln \frac{S_B(k)}{T_B(k)} \tag{3.1}$$

where, $S_B(k)$ and $R_B(k)$ represent the k^{th} element of the feature vectors of the synthesized and reference image block, respectively. Also, K represents the length of the feature vector and $T_B(k) = \frac{1}{2}(R_B(k) + S_B(k))$. A quality score denoted as Q'_B, close to zero indicates a high similarity between the feature vectors, indicating better perceptual quality. On the other hand, a value close to 1 suggests a dissimilarity between the feature vectors, indicating poorer quality. This score represents the level of distortion in each block and serves as the ground truth to train the subsequent deep learning (DL) model.

To demonstrate that the block-level quality score Q'_B mimics the human visual system for quality assessment, some examples from the IETR dataset are given in Fig. 3.4. As observed from the figure, the blocks which have less distortions eg. Fig. 3.4 (a)-(b) have lower Q'_B values, while more distorted blocks eg. Fig. 3.4 (c)-(i) have higher scores. This shows that there is an inverse relationship between perceptual quality and Q'_B scores. This quality score associated with each block is a significant contribution of the proposed metric, distinguishing it from the binary classification of blocks in Figure 3.2. These blocks and their respective scores are utilized to train the CNN-based deep learning model proposed in the subsequent section.

3.2.2 Training Deep Learning Model

In the previous step, the quality scores of the blocks in the DIBR views were obtained. These scores, along with their corresponding blocks, are utilized in the proposed NR algorithm. This approach allows for the determination of a ground truth score for each block, which in turn generates sufficient training data for training a DL model at the block level.

In transfer learning a pre-trained model is employed as a starting point for a new application. By leveraging knowledge learned from a large dataset and applying it to a related problem with limited data, transfer learning helps enhance the efficiency of the new model. In the proposed NR metric, the transfer learning technique is employed wherein a pre-trained model is concatenated with a few trainable layers. Figure 3.5 represents the workflow of the deep learning model, i.e., Step-2 of the proposed NR IQA. To train the model, the NSCT image blocks are fed to the proposed deep learning model along with the block-level ground truth score Q'_B (from Step 1). The final deep learning model consists of the backbone VGG-16 along with a series of consecutive dense layers. To avoid overfitting, dropout layers are included after each Dense layer with a value equal to 0.3 in the model. The ReLu activation function is employed for all the Dense layers except the last one. Moreover, the final Dense layer has a size equal to 1, depicting regression output with the Linear activation function. The loss function was taken as Mean Square Error, and the ADAM optimizer is deployed. The results were obtained for the whole dataset using the "k-fold validation technique" with k equal to five. The Early stopping strategy is also employed to avoid overfitting. After efficiently training the model, the blocks from the remaining 20% dataset were tested. Thus, the proposed NR model determines the quality of each block (m) of an image given by $Q''_B(m)$.

3.2.3 Thresholding and Pooling

In DIBR views the prominent distortions are located in just a few blocks, and the remaining blocks have less effect on the overall perceptual quality. With this context, we propose to only select blocks that have significant distortions (i.e., poor quality) for estimating the quality score. Next, we calculate the sum of the quality scores of these poor-quality blocks, $QS_1(i)$ for an image i using Eq. 3.2.

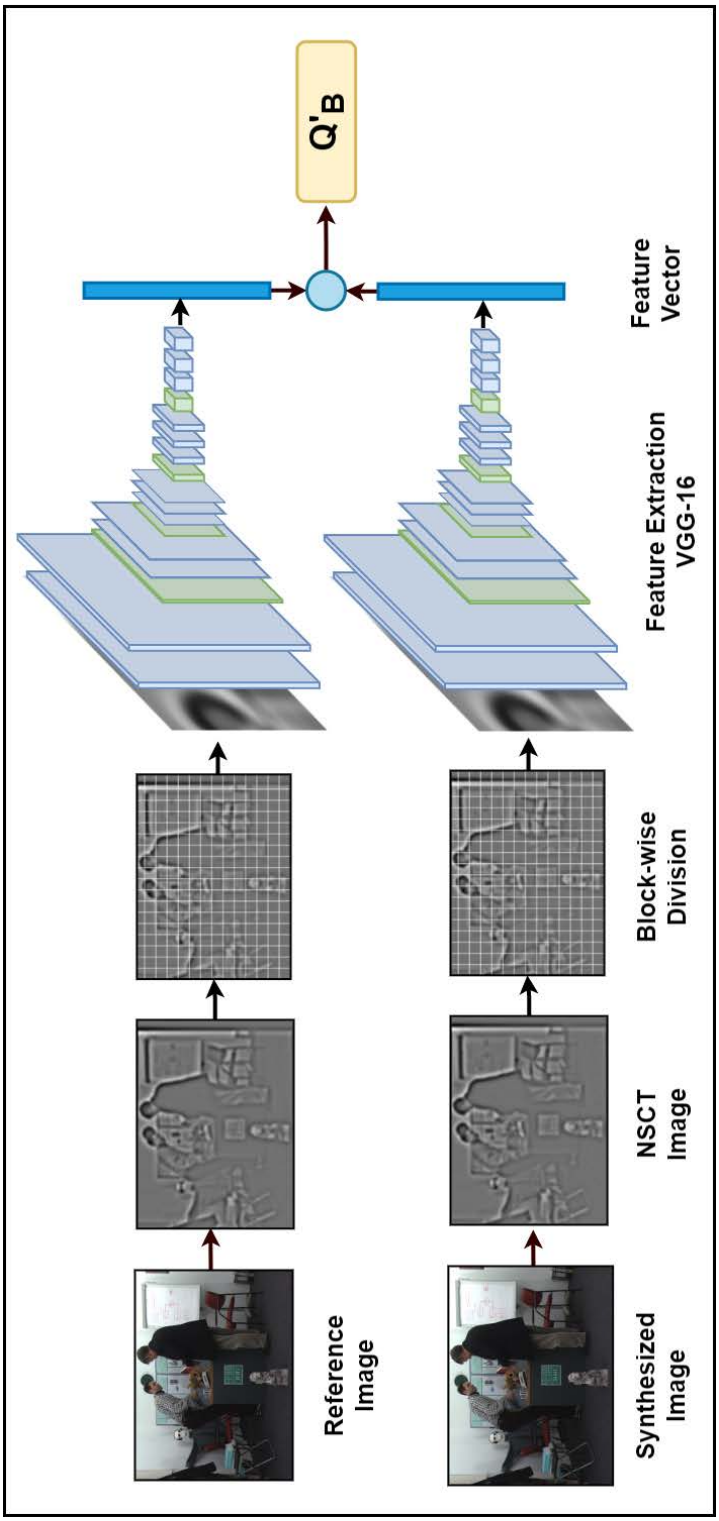


Figure 3.3: Workflow for obtaining block-level ground truth scores in the Proposed NR Metric (Step-1).

(a) $Q'_B = 0.119$ (b) $Q'_B = 0.212$ (c) $Q'_B = 0.325$

(d) $Q'_B = 0.358$ (e) $Q'_B = 0.403$ (f) $Q'_B = 0.415$

(g) $Q'_B = 0.563$ (h) $Q'_B = 0.591$ (i) $Q'_B = 0.625$

Figure 3.4: Examples of image blocks from the IETR dataset of different perceptual quality along with the quality score (Q'_B) obtained by the proposed NR model. Higher values of Score indicate poor perceptual quality.

$$QS_1(i) = \sum_{m=1}^{M} \phi(Q''_B(m)) \tag{3.2}$$

where,

$$\phi(Q''_B(m)) = \begin{cases} Q''_B(m), & \text{if } Q''_B(m) \geq Th \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

Here, Th is the threshold value. Since the size of the images present in the evaluating database is different, to normalize the quality score, divide $QS_1(i)$ by the total number of blocks M in the given image to obtain the no-reference predicted score $QS_2(i)$ as given

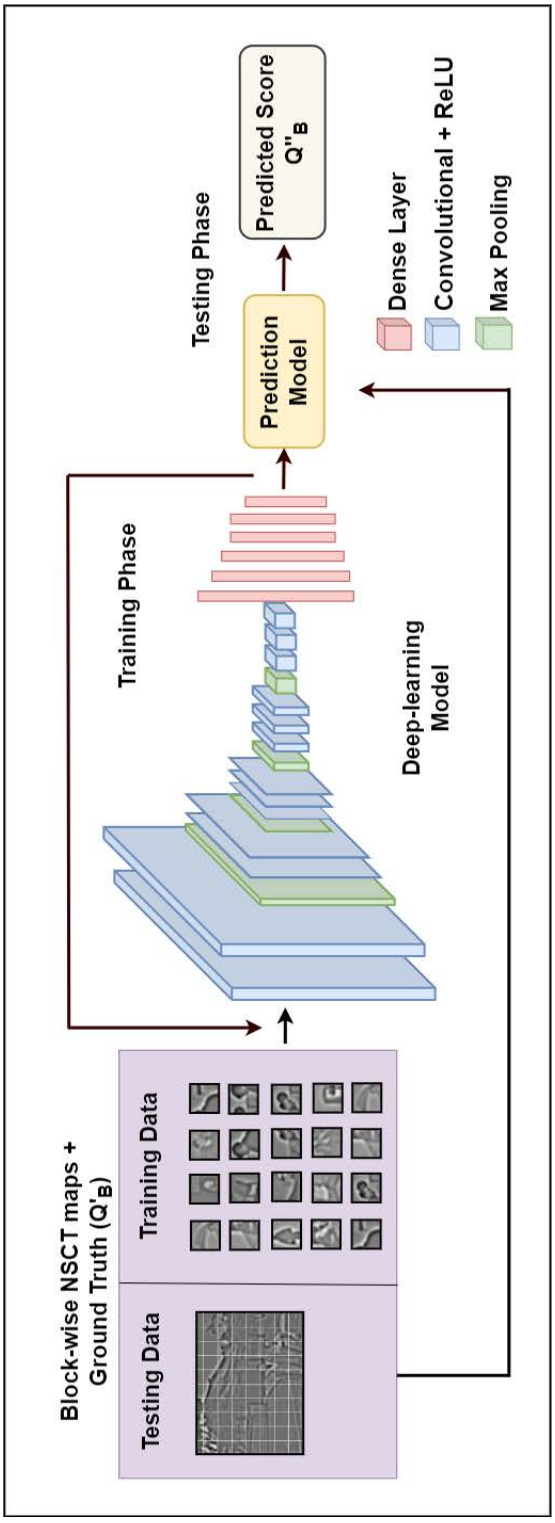


Figure 3.5: Workflow of training the block-based deep learning model (Step 2) in Proposed NR Metric.

in the equation 3.4.

$$QS_2(i) = \frac{QS_1(i)}{M} \tag{3.4}$$

In the proposed algorithm, importance is mainly given to geometric distortions. Hence, to handle structural distortions, similar to the [6], [116], we propose to pool the scores of the BIQI metric [51] with the proposed algorithm as:

$$NQ_i = QS_2(i)^p - B_i^q \tag{3.5}$$

where B_i is the quality value calculated by the BIQI metric for the image i. $QS_2(i)$ is the initial predicted quality score from the proposed metric. Moreover, it may be noted that there is a directly proportional relationship between the predicted $QS_2(i)$ score and DMOS. At the same time, there is an inverse relationship between the DMOS and the B score. A lower value of NQ_i indicates better quality and vice-versa. The same intuition is reflected in Eq. 3.5. The powers p and q are used to stabilize the scores, as the values of the proposed metric and B have different ranges. The same pooling step is also included in the earlier proposed NSCT-FR metric to enhance its performance further.

3.3 Result Analysis

For analyzing the performance of the proposed NR quality assessment metric, a series of experiments were conducted. The performance is evaluated on the basis of r, ρ, τ, and RMSE for these studies. This section presents a detailed study of the results obtained for the proposed work on two main DIBR repositories; IETR [18] and IVR dataset [20].

3.3.1 Performance Analysis

Table 3.1 shows the performance of the proposed NR metric and 15 existing no-reference IQA metrics on the IETR dataset. These metrics include SI-DL [6], Yan's [119], Tian's [30], GANs-NR [60], DSCB [69], Wang's [57], BIQI [51], APT [50], NR-MWT [55], MNSS [54], OMIQA [151], NIQE [135], Jakhetiya's [151], and Yue's [61]. As observed from Table 3.1, the proposed NR model has the values of r, ρ, τ, and RMSE equal to 0.7211, 0.7091, 0.5114, and 0.1718 for the IETR dataset [18]. In comparison to the best-performing NR

metric SI-DL with these scores equal to 0.7087, 0.6672, 0.4726, and 0.1749, there is a reasonable increase in the overall performance.

Likewise, the performance of the proposed NR metric with respect to various SOTA NR IQA which include SI-DL [6], APT [50], Jakhetiya's [151], NIQSV [152], SIQE [153], OMIQA [151], MNSS [54], and NIQSV+ [53] for IVY dataset. The results are given in Table 3.2. The proposed NR IQA has the r, ρ, τ, and RMSE equal to $0.6693, 0.6283, 0.4402$, and 10.5856. In comparison to the highest performing NR IQA, SI-DL, these scores are equal to $0.5459, 0.5396$, and 11.9349, there is a significant gain in performance.

In addition, a cross-dataset experiment has also been conducted for the proposed work, where we trained the model using the IETR dataset and evaluated its performance on the IRCCyN/IVC dataset [19]. The results were quite promising, with r, ρ, τ, equal to 0.7709, 0.7312, 0.5341, and respectively. These findings indicate that the proposed technique demonstrates good generalization ability across different datasets.

3.3.2 Statistical Significance Test

Table 3.3 depicts the results of the F-score/statistical test between the proposed NR and six IQA metrics respectively. These scores are +1 for all the methods, thus depicting that the proposed NR metric is statistically better than the various in comparison to other IQA metrics (confidence interval equal to 90%).

3.3.3 Ablation Study

In this study, an extensive ablation analysis is performed to investigate the impact of different parameters on the proposed model's performance. The ablation studies conducted are presented below:

3.3.3.1 Analysis of the Effect of Pooling

To show the effect of the pooling step, we conducted an ablation study. The step-wise comparison of the proposed NR metrics, with and without the pooling with the BIQI metric for the IETR dataset, is presented in Table 3.6. From the table, it can be perceived that the individual BIQI metric has r and ρ equal to 0.4327 and 0.4321, while, the proposed NR metric without pooling has r and ρ values equal to 0.6637 and 0.6628. However, when

Table 3.1: Performance comparison of the proposed NR metric with various NR objective quality metrics on the IETR dataset. The '-' symbol depicts that the data is not available and '"'symbol represents "same as above".

S. No	Technique	Aimed For	r	ρ	τ	RMSE
1.	**Proposed NR Metric** (NQ_I)	**3D views**	**0.7211**	**0.7091**	**0.5114**	**0.1718**
2.	SI-DL [6]	"	0.7087	0.6672	0.4726	0.1749
3.	Yan's [119]	"	0.6881	0.6261	0.4660	0.1750
4.	Tian's [30]	"	0.6685	0.5903	-	0.1844
5.	GANs-NR [60]	"	0.6460	0.5710	-	0.1980
6.	DSCB [69]	"	0.6030	0.5571	0.3677	0.1978
7.	Wang's [57]	"	0.4338	0.4254	-	0.2244
8.	BIQI [51]	Natural images	0.4327	0.4321	0.2898	0.2223
9.	APT [50]	3D views	0.4329	0.4164	0.2830	0.2235
10.	NR-MWT [55]	"	0.4720	0.4560	0.3170	0.2180
11.	MNSS [54]	"	0.2930	0.2960	0.1940	0.2370
12.	OMIQA [151]	"	0.2705	0.2331	0.1593	0.2387
13.	NIQSV+ [53]	"	0.2324	0.1545	0.1083	0.2411
14.	NIQE [135]	"	0.2240	0.1360	-	0.2420
15.	Jakhetiya's [151]	"	0.1650	0.1640	0.1120	0.2440
16.	Yue's [61]	"	0.1146	0.0860	-	0.2463

Table 3.2: Performance comparison of the proposed NR metric with various NR objective quality metrics on the IVY dataset. The '-' symbol depicts that the data is not available.

S. No	**Technique**	**Aimed For**	**r**	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$	**RMSE**
1.	**Proposed NR** (NQ_I)	**3D views**	**0.6693**	**0.6283**	**0.4402**	**10.5856**
2.	SI-DL [6]	3D views	0.5459	0.5396	-	11.9349
3.	APT [50]	3D views	0.5240	0.4748	0.3389	20.9961
4.	Jakhetiya's [52]	3D views	0.5211	0.2288	-	12.1467
5.	NIQSV [152]	3D views	0.4113	0.2717	0.1945	22.4706
6.	SIQE [153]	3D views	0.3855	0.3764	0.2524	22.746
7.	OMIQA [151]	3D views	0.2367	0.1131	-	13.7566
8.	MNSS [54]	3D views	0.2205	0.1474	-	22.756
9.	NIQSV+ [53]	3D views	0.2191	0.2990	0.2037	24.0530

Table 3.3: Results of the F-Test conducted between the proposed NR metric and the various SOTA IQAs

Metric	**SI-DL**	**LPIPS**	**DSCB**	**APT**	**NIQSV**	**SSPD**
Score	+1	+1	+1	+1	+1	+1

Table 3.4: Step-wise comparison of the performance of the proposed NR metrics on the IETR dataset.

S. No.	**Technique**	**r**	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$	**RMSE**
1.	**Proposed NR with pooling**	**0.7211**	**0.7091**	**0.5114**	**0.1718**
2.	Proposed NR without pooling	0.6637	0.6628	0.4738	0.1855
3.	BIQI [51]	0.4327	0.4321	0.2898	0.2223

these scores are pooled with the BIQI metric score, there is a gain of about 8.6% in the metric's performance (in terms of ρ). Likewise, the step-by-step analysis for the IVY dataset is given in Table 3.5. For the proposed NR technique without pooling, the r, and ρ values are 0.5682 and 0.5252, and after pooling, this value is increased by 17.7%, and 19%, respectively.

Similar studies were conducted for the NSCT-FR metric proposed in Chapter 2. The results on the benchmark datasets are given in Tables 3.6 and 3.7. The results show that the pooling technique results in an increase in performance for NSCT-FR metrics also.

The study shows that while the proposed metrics are suitable for the detection of degradation caused by geometric distortions, the BIQI metric helps in determining the quality with respect to structural distortions, and by fusing these two together, the overall perceptual quality is efficiently predicted.

3.3.3.2 Analysis of the Effect of Block Size

During the initial steps (1 and 2) of the proposed NR metric, the ground truth scores for image blocks are obtained, and these blocks are subsequently utilized to train the deep learning model. The quality score assigned to each block represents the level of distortion present within it. Determining the appropriate block size is crucial as an excessively large block may not accurately capture the location of distortion, while an overly small block may fail to adequately represent the image properties.

Table 3.5: Step-wise comparison of the performance of the proposed NR metrics on the IVY dataset.

S. No.	Technique	r	ρ	τ	RMSE
1.	**Proposed NR with pooling**	**0.6693**	**0.6283**	**0.4402**	**10.5856**
2.	Proposed NR without pooling	0.5682	0.5252	0.3547	11.3231
3.	BIQI [51]	0.5308	0.5159	0.3610	12.0724

Table 3.6: Step-wise comparison of the performance of the proposed FR metrics on the IETR dataset.

S. No.	Technique	r	ρ	τ	RMSE
1.	**Proposed FR with pooling**	**0.8207**	**0.8187**	**0.6203**	**0.1417**
2.	Propose-FR without pooling	0.8113	0.8105	0.6117	0.1450
3.	BIQI [51]	0.4327	0.4321	0.2898	0.2223

Table 3.7: Step-wise comparison of the performance of the proposed FR metrics on the IVY dataset.

S. No.	Technique	r	ρ	τ	RMSE
1.	**Proposed FR with pooling**	**0.7580**	**0.7375**	**0.5418**	**9.4090**
2.	Propose FR without pooling	0.7128	0.7113	0.5159	9.9912
3.	BIQI [51]	0.5308	0.5159	0.3610	12.0724

Table 3.8: Effect of varying the block size in the proposed NR metric on the IETR database.

S. No.	Block Size	r	ρ	τ	RMSE
1.	128×128	0.6924	0.6969	0.5044	0.1778
2.	**160×160**	**0.7211**	**0.7091**	**0.5114**	**0.1718**
3.	192×192	0.6661	0.6720	0.4740	0.1849

To assess how the size of the block impacts the effectiveness of the proposed metric, an ablation study investigating the relationship between block size and performance is presented in Table 3.8. Different block sizes were examined: 192×192, 160×160, and 128×128. As demonstrated in the table, the optimal performance is achieved when employing a block size of 160×160 compared to the other sizes within the IETR dataset.

3.3.3.3 Analysis of the Parameter Sensitivity.

In the pooling stage, the scores of the proposed NR metric were fused with those of the BIQI metric using 3.5. As discussed earlier, there is a direct and inversely proportional relationship between the DMOS with the proposed metric and BIQI respectively. Further, the range of proposed metrics is between 0 to 1 while that of the BIQI metric is much wider (-1 to 56), it is implicit that the value p will be greater than q. We employed a 3D mesh graph in Fig.3.6 to analyze the effect of change of the two-parameter values i.e. "p" and "q" on the performance, in terms of τ, of the proposed NR metric on the IETR dataset. The analysis from the 3D plot in Fig.3.6 validates our intuition and depicts good performance when the value of "p" is greater than "q".

3.3.3.4 Analysis of the Ground-truth Generator

Furthermore, to demonstrate the effectiveness of the previously proposed NSCT-FR algorithm in generating reliable ground truth values, an ablation study is conducted (see Table 3.9) using various IQA methods for generating the ground truth scores. The results showcased in the table reveal that when employing the NSCT-FR algorithm as the block-level ground truth generator, the DL model outperforms other SOTA algorithms used for the same purpose. This ablation study validates the efficiency of the earlier proposed NSCT-FR metric and its applicability within the proposed NR metric as a block-wise

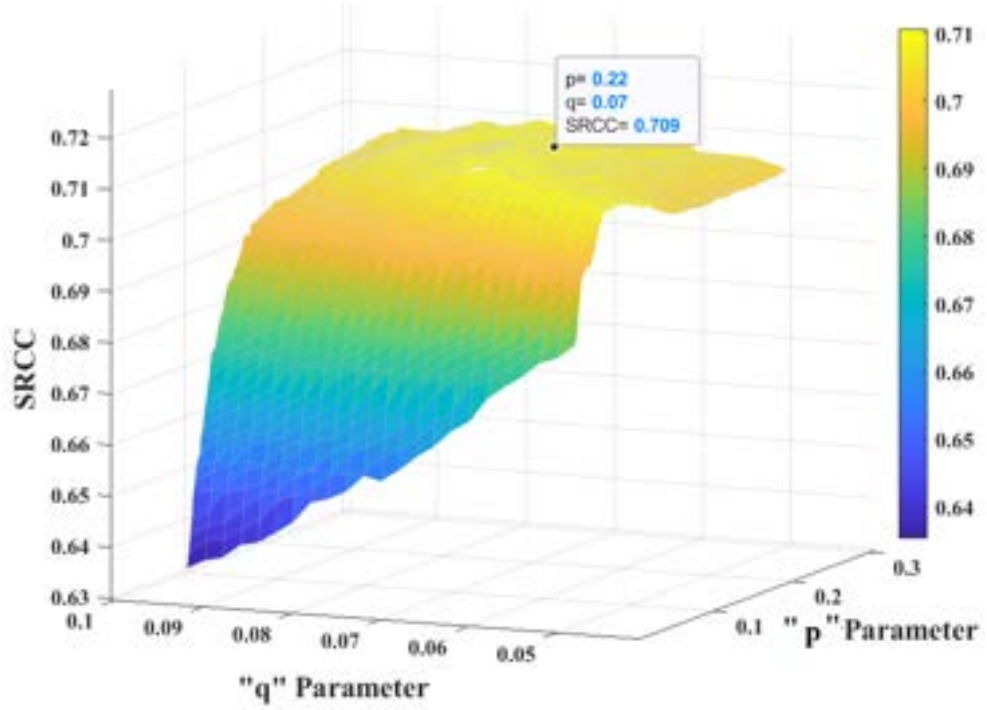


Figure 3.6: Performance dependency of the proposed NR metric with respect to the two parameters 'p' and 'q' on the IETR dataset.

ground truth generator.

3.3.4 Scatterplot Analysis

Furthermore, to facilitate a clearer visual interpretation of the outcomes, scatterplots illustrating the correlation between DMOS and the quality scores predicted by several SOTA methods are generated. The NR IQA metrics analysed include NIQSV+ [53], DSCB [69], KRR [52], NIQE [135], BRISQUE [136], BIQI [51], Highgrade [137], and HyperIQA [138]. The scatterplots, as shown in Figure 3.7, reveal a strong linear relationship between the proposed no-reference metric and the subjective scores when compared with the other techniques. These results indicate that the proposed model exhibits a high level of consistency with the human visual system.

3.4 Conclusion

This chapter introduces a novel NR quality assessment metric for DIBR views. The metric utilizes a DL model trained on block-level ground truth scores. By predicting

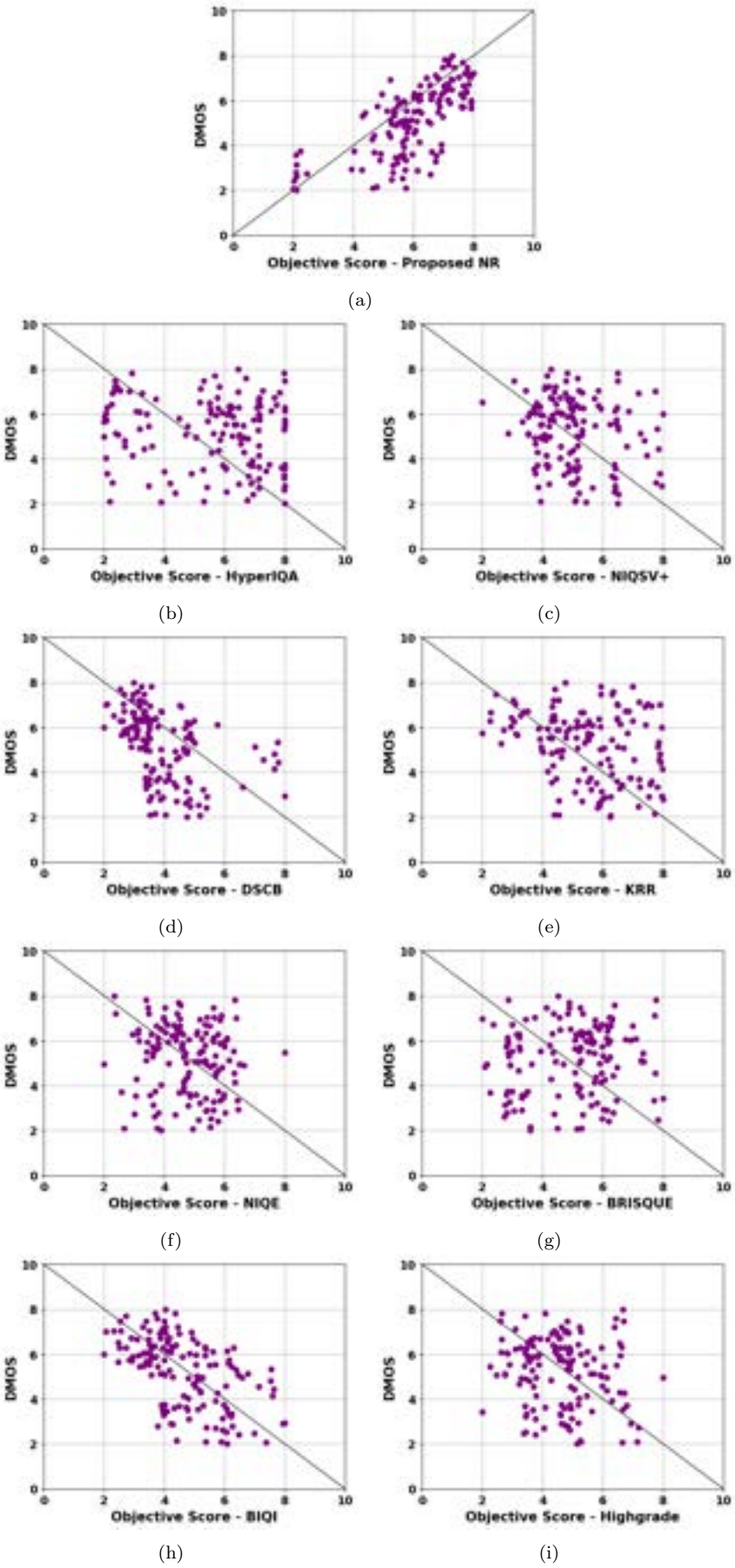


Figure 3.7: Scatter Plot of subjective score/DMOS and objective scores of SOTA metrics on IETR dataset.

Table 3.9: Ablation study when several SOTA techniques are used for the ground truth score generation in the proposed NR algorithm.

S. No.	Technique	r	ρ	τ	RMSE
1.	**Proposed-NR** (QS_2)	**0.6637**	**0.6628**	**0.4738**	**0.1855**
2.	PU-IR [116]	0.6175	0.6201	0.4488	0.1950
3.	DF-CS [116]	0.5248	0.5243	0.3690	0.2110
4.	SSPD [35]	0.4360	0.3345	0.2382	0.2231
5.	SSIM [9]	0.2220	0.1274	0.0832	0.2417
6.	NIQSV+ [53]	0.1870	0.0887	0.0547	0.2439
7.	APT [50]	0.0420	0.0220	0.0145	0.2477

quality scores for each block of an image and aggregating them, a final quality score for the entire image is obtained. The proposed method showcases a unique approach to determining the intensity of block-level quality scores, resulting in better performance compared to existing SOTA techniques.

Chapter 4

Non-Intrusive Audio Quality Assessment Metric for User-Generated Multimedia Using Deep Learning

4.1 Introduction

User Generated Multimedia refers to multimedia content created, captured, uploaded, and shared by inexperienced or non-professional users in real-world scenarios. This kind of data is susceptible to various distortions caused by factors such as poor capturing devices, limited bandwidth for sharing, background noise, and low bit-rate. In this chapter, we explore the domain of quality assessment for UGM audio and propose a novel technique for evaluation.

4.2 Motivation

Chapter 1 provides an in-depth analysis of the literature on audio quality assessment. The literature reveals that the majority of prevalent techniques are primarily designed for assessing the quality of speech data. However, upon analyzing the spectrograms of both speech and UGM audio, it becomes apparent that speech signals possess distinct acoustic characteristics compared to UGM audio. As a consequence, the application of existing

speech metrics to UGM data leads to suboptimal performance, highlighting the need for specialized quality assessment algorithms tailored for UGM content.

Moreover, in Chapter 1, an in-depth study of the existing audio datasets is given which is summarized in Table 4.1. The table encompasses information such as the purpose of dataset creation, types of distortion present, size, year of creation, availability, and whether subjective testing has been conducted or not. From the table, we can identify the following limitations associated with the existing datasets.

- Lack of diversity in terms of magnitude and distortion types.
- Lack of subjectively annotated and openly available UGM datasets.
- Lack of diversity in terms of content and context.

Thus, to address the limitations in the existing literature on UGM audio quality assessment, this chapter presents the following contributions:

1. A novel and comprehensive large-scale database called IIT-JMU-UGM Audio Dataset is designed and developed, comprising 1,150 audio clips extracted from diverse UGM sources. The dataset encompasses a vast range of contexts, content, and distortions, and each clip is annotated with subjective quality scores.
2. A non-intrusive stacked Gated Recurrent Unit model is developed for audio quality estimation. The model takes multiple audio features as input and emulates the human perceptual auditory system, achieving better performance compared to existing SOTA methods. The model attains a remarkable Pearson's Linear Correlation Coefficient value of 0.834.

4.3 Proposed Work

The proposed work consists of two main parts. The first part involves the development of the IIT-JMU-UGM Audio Dataset. This dataset aims to gather a diverse collection of audio samples from various UGM sources. The second part of the work focuses on developing a quality assessment metric to evaluate the quality of the UGM audio data.

Table 4.1: Description of the existing audio datasets. "-" indicates unavailability of relevant information. Avail. indicates the open availability of the dataset and S.T. represents Subjective testing.

Dataset	**Purpose**	**Type of Distortions**	**Year**	**Avail.**	**Sample Size**	**S.T.**
TIMIT [4]	Speech recognition applications	None	1993	Under Licence	5.4 hours	-
ITU-T P Suppl. 23 [71]	Objective voice quality assessment	Narrowband speech degradation, environmental noise, audio encoding, and channel degradation	1998	Under Licence	-	Yes
SPINE [3]	Speech recognition in military setup	Noisy military environment	2000	Under Licence	12 hours	-
NOIZUS [2]	Assessment of speech enhancement techniques	Various background noises	2006	Yes	960 samples	Yes
EBU-SQAM [79]	Sound quality assessment	-	2008	Yes	70 samples	-
Creusere's [74]	Audio quality assessment	Changing bit rates	2008	-	48 samples	Yes
Live Music dataset [76]	Live music recording quality assessment	Amplitude compression, amplification, band pass filtering, white noise, and crowd noise	2013	-	2900 samples	Yes
ACE-Challenge [80]	Estimation of acoustic parameters	Reverberations, multi-channel noise, different signal to noise ratio	2015	Yes	4500 samples	-
CoreSV14 [78]	Evaluation of various codec	Distortions due to different types of Codec used	2014	Yes	40 samples	Yes
REVERB-Challenge [86]	Evaluation of automatic speech recognition and enhancement techniques	Reverberant room responses and environmental noise	2016	Yes	-	Yes
Fazenda's [72]	Audio quality assessment	Background noise	2016	No	128 samples	Yes
Avila's [77]	Speech quality assessment	Room impulse response, background noise	2019	-	10,000 samples	Yes

4.3.1 Development of IIT-JMU-UGM Audio Dataset

The proposed IIT-JMU-UGM Audio Dataset consists of 1,150 audio clips paired with their respective subjective scores. The dataset encompasses a diverse range of content, context, and levels of distortions. The following steps were undertaken during the development of this dataset.

4.3.1.1 Dataset Creation

For the preliminary data collection, multimedia-sharing platforms such as Flickr, YouTube, CVDL [154], and Vimeo, among others, were identified. These platforms provide a vast collection of UGM data uploaded by both amateur users and professionals. Initially, about 350 audio-video samples were obtained, out of which 217 clips were identified, having appropriate copyright permission, an adequate amount of audio, and diverse content. Next, the audio part of the sample was extracted from the multimedia clips while disregarding the accompanying video. The samples were cropped to an average duration of 8 seconds, providing a suitable length for analysis and evaluation. To simulate realistic scenarios and create an extensive repository of audio instances, we deliberately introduced various distortions that perceptually impact audio quality. These distortions encompass a range of factors that can affect audio perception, ensuring a diverse and representative collection. The steps involved in the creation of the dataset are depicted in Fig. 4.1. Consequently, we modelled the following two types of impairments in the dataset:

1. **Background Noise:** To capture the realistic conditions found in UGM clips and their impact on perceptual quality, we introduced background noises and impairments into the dataset. These included Gaussian noise, pink noise, as well as specific impairments like hum, glitch, babble, microphone effects, echo, etc.

   Different types and intensities of these noises were added to the clean audio samples to create varying levels of degradation. This process allowed us to simulate the diverse range of degradations that can affect audio quality in UGM clips.

2. **Bit rate:** According to Winkler [155], the audio bit rate highly affects its perceptual quality, i.e., the lower the bit rate, the poorer the quality. Taking these findings into consideration, the audio clips in the dataset were compressed by a range of bit rates, spanning from very low to high values (approximately 10 kbps to 280 kbps). By

including this variation in bit rates, we aimed to capture the diverse quality levels that can be perceived based on the compression settings applied to the audio clips.

After a meticulous selection process and applying necessary preprocessing techniques, we successfully curated the final version of the IIT-JMU-UGM Audio Dataset. This comprehensive dataset consists of 1150 audio clips, specifically designed to encompass a wide range of generalized and realistic distortions in UGM.

4.3.1.2 Subjective Testing

To facilitate the subjective evaluation of the IIT-JMU-UGM Audio Dataset, a simple graphical user interface was designed. The Single-Stimulus Model [156] was adopted wherein the subjects were presented with individual audio samples without any information about the corresponding reference samples. A total of 26 volunteers participated in the subjective testing process. The workflow of the subjective testing consisted of 4 phases; Collection of the subject's demographic information, general instructions about the test, and a brief training session followed by the subjective testing phase, depicted in Fig. 4.2. Basic demographic information such as age, gender, expertise in audio processing, and the audio device used by each subject was collected. Furthermore, no separate hearing test was conducted however, it was ensured that participants had an average hearing ability and that their audio devices did not introduce any explicit sound degradation.

The subjects were provided with basic instructions about the test. Next, a short training session was held to acquaint the participants with both the user interface and the test procedure. During this session, a few audio samples were played to ensure participants were comfortable with the setup after which the actual subjective test was conducted. The database was thoroughly shuffled so as to avoid disinterest and information retention among the listeners. In the testing phase, the subjects were instructed to rate audio quality clips using the "5-Grade Absolute Category Rating (ACR) scale" (recommended in ITU-T P.910 [156]). In this scale, the options for assessment range from "Bad" to "Excellent," corresponding to numeric scores of 1 to 5 as shown in Fig.4.3. To mitigate the effects of listener fatigue, only 30 clips were presented to each listener at a time, followed by a short break before further evaluations. This approach aimed to maintain the concentration and accuracy of the subjective evaluations.

In order to determine the final subjective rating of each audio clip, the Mean Opinion

Figure 4.1: Workflow of the creation of the IIT-JMU-UGM Audio Dataset.

Figure 4.2: Workflow of the subjective testing

Quality Score	Distortion Level	Description
1	Bad	Very annoying and objectionable
2	Poor	Annoying, but not objectionable
3	Fair	Perceptible and slightly annoying
4	Good	Just perceptible, but not annoying
5	Excellent	Excellent and imperceptible

Figure 4.3: The ITU-T Five-point scale – ACR used during subjective testing.

Score (MOS) is computed by averaging the score given by the subjects for a particular clip, calculated by:

$$MOS_i = \frac{1}{N_j} \sum_{1}^{j} X_{ij} \tag{4.1}$$

where X_{ij} represents the score given by the jth subject to the ith audio clip and N_j denotes the total number of ratings given to the clip. This operation was conducted on the whole dataset to determine the ground truth or perceptual quality score of each audio clip.

4.3.2 Proposed Quality Assessment Metric

In this section, a non-intrusive quality assessment metric is proposed for evaluating the perceptual quality of audio samples. The metric consists of two main segments: the feature extraction module followed by the deep learning module. Figure 4.4 provides an overview of the proposed model's architecture.

4.3.2.1 Feature Extraction Module

The proposed architecture begins with the extraction of key audio features, which play a crucial role in representing the audio data efficiently. Feature extraction involves cap-

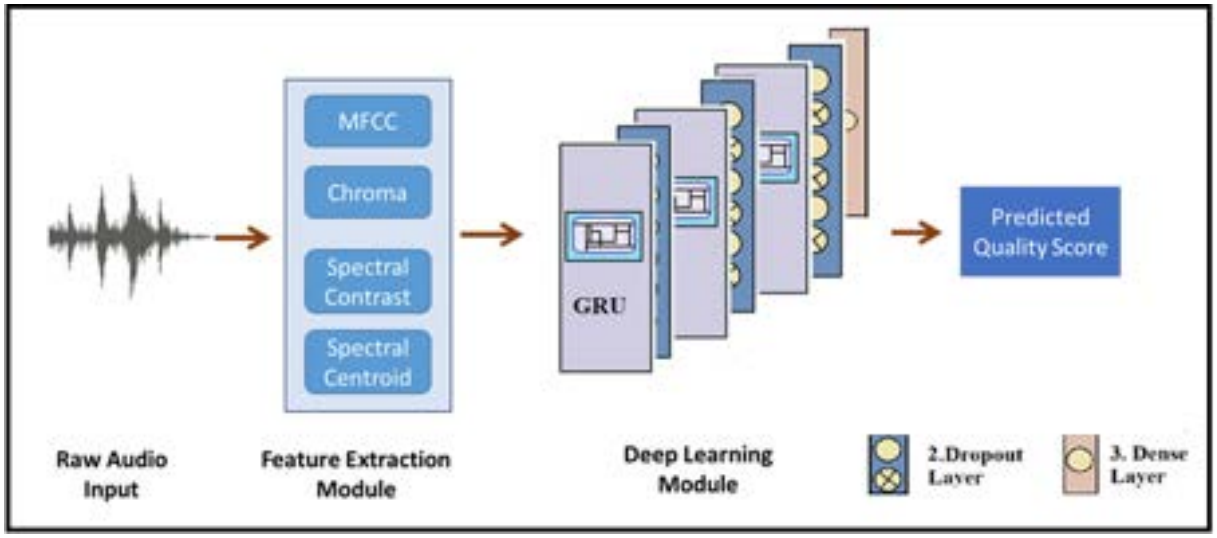


Figure 4.4: The detailed architecture of the proposed metric.

turing essential and discriminative information from the input signal while reducing its dimensionality. As a result, the subsequent learning algorithm's computational efficiency is improved, rendering it well-suited for real-time applications. In this work, the performance of various audio features was explored which include Mel-frequency Cepstral Coefficients, Spectral Centroid, Chroma, and Spectral Contrast. A description of each of these features is given below. These features have also been employed in other audio processing applications, such as speaker recognition, classification, information retrieval, speech enhancement, and more [157].

1. **Mel-frequency Cepstral Coefficient (MFCC):** The MFCCs depict the short-term power spectrum of sound instance, which is based upon a "linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency" [158]. The steps to obtain MFCC involve preprocessing the audio signal, segmenting it into frames, computing the power spectrum, applying Mel filter banks, taking the logarithm, and applying the "Discrete Cosine Transform (DCT)" [159].

   Let the outputs of an I-channel filterbank be denoted as $Z(i), i = 1, 2, ...I$, the MFCCs are calculated as:

$$c_g = \sum_{i=1}^{I} [logZ(i)] \, cos \left[\frac{\pi g}{I} \left(i - \frac{1}{2} \right) \right], \tag{4.2}$$

   where g denotes the index of the cepstral coefficient. The lowest DCT coefficient can be extracted to determine the final MFCC vector.

These "biologically inspired" MFCC coefficients effectively capture crucial acoustic characteristics in audio signals, closely mimicking human auditory perception. These have found widespread use in a number of audio-processing applications such as speaker recognition, genre classification, audio similarity measures, etc. [160–162]. Given their advantages and proven performance, we use the first twenty MFCC features in our work.

2. **Spectral Centroid:** The spectral centroid represents the point of the "center of gravity" i.e. the point where the energy of a spectrum is centered upon. It is used as an indicator of "brightness" in sound and extensively employed as an automatic estimator of timber in digital music and audio processing. Spectral Centroid (η) is obtained by calculating the weighted mean of the frequencies extracted using a Fourier transform, where the magnitudes are present as the weights [163].

$$\eta = \frac{\sum_{i=a_1}^{a_2} g_i l_i}{\sum_{i=a_1}^{a_2} l_i} \tag{4.3}$$

where a_1 and a_2 are the band edges, in bins, over which to calculate the spectral centroid, l_i is the spectral value at bin i and g_i is the frequency in Hz corresponding to bin i.

3. **Chroma:** Chroma represents the intensity value of the twelve defined musical octaves at each time frame [164]. In the chroma feature, the entire information regarding spectra corresponding to a given pitch class is accumulated into a unique coefficient.

   Let the value of chroma $c \in [0:11]$ denote the 12 pitch attributes present in music and pitch $s \in [0:127]$. Let $T_F : \mathbb{Z}[0:127] \rightarrow R_{\geq}0$ be a pitch-based log-frequency spectrogram, then the chroma representation $\mathbb{Z} \times [0:11] \rightarrow \mathbb{R}_{\geq}0$ is obtained by aggregating the coefficients of pitch which belong to the same chroma [165]:

$$C(m,c) := \sum_{s \in [0:127] | s \ mod \ 12 = c} T_F(m,s), \tag{4.4}$$

   Chroma features have been also used in a number of sound-based applications such as [166–168] and have shown good performance. In this regard, we also made use

of chroma feature is the model.

4. **Spectral Contrast:** The spectral contrast consists of the spectral valley, spectral peak, and their difference in every sub-band of the frequency. Seven of these features were used in this work. Let the Fast Fourier Transform vector of r-th sub-band be $\{y_{r,1}, y_{r,2}, ...y_{r,M}\}$. The sorted vector can be expressed as $\{y'_{r,1}, y'_{r,2}, ...y'_{r,M}\}$, such that $\{y'_{r,1} > y'_{r,2} > ... > y'_{r,M}\}$ and neighborhood factor is represented by β. Then the strength of spectral valleys (V), spectral peaks (P) and their difference (ST) is expressed as [169];

$$V_r = log\left\{\frac{1}{\alpha M}\sum_{i=1}^{\beta M} y'_{r,M-i+1}\right\}, \tag{4.5}$$

$$P_r = log\left\{\frac{1}{\alpha M}\sum_{i=1}^{\beta M} y'_{r,i}\right\}, \tag{4.6}$$

$$ST_r = P_r - V_r, \tag{4.7}$$

such that M is total number in r^{th} sub-band.

For each audio sample, all the above-mentioned features are extracted and combined into a single feature vector. This vector is fed to the deep learning model for training. For feature extraction, the Librosa library in Python is utilized. The window length (fftsize) was set to 2048, and the hop length was set to 512. Discrete Cosine Transform Type-2, along with the default sample rate, was applied to each audio clip.

4.3.2.2 Deep Learning Module

Once the relevant features were extracted, various deep-learning techniques were explored to determine the perceptual quality of the IIT-JMU-UGM Audio samples. Recurrent Neural Networks (RNNs) have emerged as a pioneering technique for deep learning with time-series-based data. Unlike regular deep neural networks and convolutional neural networks, RNNs are specifically designed to handle sequential or time-dependent data, such as audio and video. RNNs exhibit superior performance due to the presence of recurrent connections, which enable them to capture temporal dependencies in the data

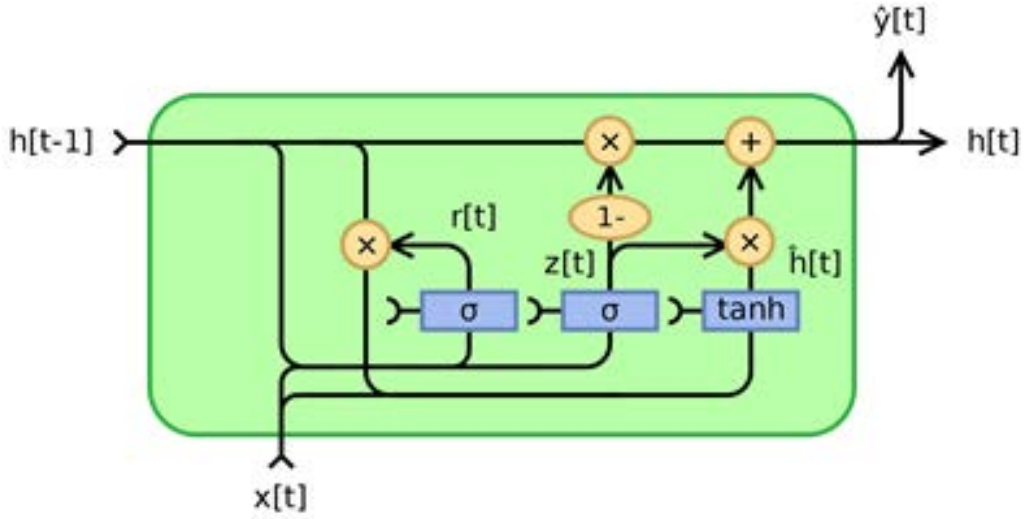


Figure 4.5: Architecture of the basic Gated Recurrent Unit [7].

[170]. However, a limitation of the basic RNN model is the vanishing gradient problem, which hampers its effectiveness in long-term dependencies. To address this issue, advanced types of RNNs such as "Long Short Term Memory (LSTM)" [171], "Bidirectional Long Short Term Memory (BiLSTM) " [172], and "Gated Recurrent Units " [7] have been developed.

LSTM is one of the popular variants of RNN and consists of cells with specialized gates: input, output, and forget gate. These gates regulate the information flow, allowing the network to adaptively remember or forget specific values over time. BiLSTM, on the other hand, employs two LSTM networks that process the input sequence in both forward and backward directions, capturing bidirectional dependencies. GRU shares a similar architecture with LSTM but has fewer gates and training parameters, resulting in faster computation and lower memory requirements. Despite having fewer components, GRU has demonstrated competitive performance in various tasks [173]. By selecting an appropriate variant of RNN, we aim to increase the efficiency of perceptual QA for the IIT-JMU-UGM Audio samples.

In various research studies, GRUs have been effectively applied in speech signal modeling, natural language processing tasks, and music modeling [174,175]. Figure 4.5, depicts the standard architecture of a GRU. The main components of a GRU include:

1. **Update Gate** (z_t): It determines the amount of the previous hidden state that should be retained and the amount of new information from the current input that

should be incorporated.

2. **Reset Gate** (r_t): It determines the amount of the previous hidden state that should be forgotten to incorporate new information from the current input.

3. **Candidate Activation State ($\hat{h}_t$):** This is the candidate activation that will be added to the previous hidden state, but before applying the update gate. It is computed using the reset gate, the previous hidden state, and the current input.

4. **Hidden State** (h_t)**:** This state is a combination of the previous hidden state and the new memory content, weighted by the update gate.

Mathematically, the equations for the GRU are as follows [7]:

$$z_t = \sigma(W_z x_t \ + \ U_z h_{t-1} + b_z) \tag{4.8}$$

$$r_t = \sigma(W_r x_t \ + \ U_r h_{t-1} + b_r) \tag{4.9}$$

$$\hat{h}_t = \phi(W_h x_t \ + \ U_h(r_t \odot h_{t-1}) \ + \ b_h) \tag{4.10}$$

$$h_t = (1 - z_t) \odot h_{t-1} \ + \ \ z_t \odot \hat{h}_t \tag{4.11}$$

where z_t, r_t correspond to the update and the reset gate vectors at time step t. Further, x_t, h_t represent the input and output vector respectively. The candidate activation vector is given by $\hat{h}_t$. The parameters for the feed-forward connections are given as W_z, W_r, and W_h. While, the parameters corresponding to the recurrent weights are given as U_z, U_r, and U_h. It consists of trainable bias vectors b_z, b_r, and b_h, that are added before applying the non-linearities. The Element-wise multiplication is represented by $\odot$. For both the update and the reset gates, sigmoidal activation function σ is used. While the ϕ represents the hyperbolic tangent function for the candidate state.

Furthermore, a Stacked/Deep RNN architecture is comprised of a number of RNN layers that are sequentially connected. The output of one layer acts as the input to the subsequent layer. Each layer combines the learned representations from the preceding layer and forwards them to the subsequent higher layer. Consequently, the model obtains enhanced representations of the provided data [176,177]. Motivated by these observations, this work uses a deep learning model using stacked GRU architecture for audio quality assessment.

Table 4.2: Model summary of the proposed model for audio quality assessment.

S.No	Layer	Hidden Units	Dropout Rate	Output Shape	Parameters
1.	**GRU-1**	64	–	(None, 128, 64)	20160
2.	**Dropout**	–	0.2	(None, 128, 64)	–
3.	**GRU-2**	32	–	(None, 128, 32)	9312
4.	**Dropout**	–	0.2	(None, 128, 32)	–
5.	**GRU-3**	8	–	(None, 8)	984
6.	**Dropout**	–	0.2	(None, 8)	–
7.	**Dense**	1	–	(None, 1)	9

The feature vector obtained in the first step is fed to the designed stacked GRU model. We made use of three GRU layers (GRU-1, GRU-2, and GRU-3) stacked together, each followed by a Dropout layer. To prevent overfitting, the Dropout layer is employed. During training, this layer randomly sets certain input units in the model to zero at each step. The number of hidden units in each of the GRU layers is set to 64, 32, and 8, respectively. A Dense Layer with one output linear function is added to obtain the final quality scores. The output value corresponds to the quality score from the range 1 to 5, (bad to excellent). The summary of the proposed model is given in Table 4.2.

While training, the objective was to ensure that the proposed model, along with its hyperparameters, achieved optimal performance without overfitting. To assess the model's training progress and evaluate its performance, two plots were generated: the plot of loss versus the number of epochs and the plot of Pearson Linear Correlation Coefficient (PLCC or r)versus the number of epochs as given in Fig. 4.6 and Fig.4.7 respectively. The loss versus epochs plot shows the progression of the model's loss function as the training proceeds. Initially, there is a significant decrease in the loss value as the model learns and adjusts its parameters. However, as the training progresses, the loss may reach a plateau, indicating that the model has learned all it can from the available data. This plateau serves as an indication of convergence, and the loss value at this point is saved as the final model. The PLCC versus epochs plot illustrates the correlation between the objective quality scores and the actual subjective scores as the model trains over epochs. As the training continues, the PLCC value increases, indicating an improved alignment

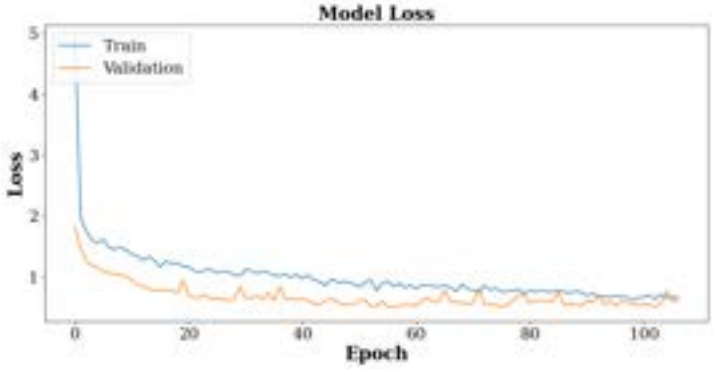


Figure 4.6: Plot between Loss and Epochs.

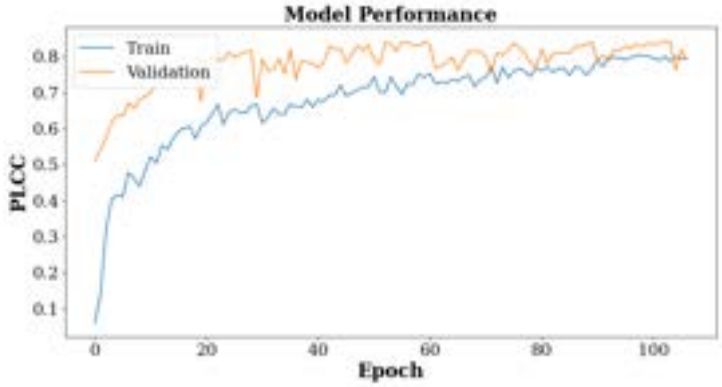


Figure 4.7: Plot between PLCC and Epochs.

between the predicted and ground-truth scores.

Further to prevent overfitting, early stopping was employed, which stops the training process if the loss does not decrease for a certain number of consecutive epochs. This prevents the model from memorizing the training data excessively and helps ensure generalization to unseen data. The plotted results demonstrate the gradual convergence of the model without overfitting, indicating that the proposed model and its hyperparameters are effective in achieving optimal performance. Also, a dropout rate of 0.2 was applied to reduce overfitting, and the model was trained for a total of 300 epochs.

To determine the optimal loss function for this model, many experiments were conducted (discussed in Results Section), employing different loss functions such as “Mean Square Error (MSE)”, “Mean Absolute Error (MAE)”, and “Mean Squared Logarithmic Error (MSLE)”. The final model utilized the MSE function as the loss function to decrease the difference between the predicted outputs and the training labels. In the process of optimizing the model, different batch sizes were experimented with, including 8, 16, 32, 64, and 128. After evaluating the performance, the optimal batch size was determined to

be 16. The Adam optimizer was employed with specific hyperparameters set as follows: the value of Beta-1 and Beta-2 equal to 0.9 and 0.999, the learning rate equals 0.01, and an Epsilon factor of 1e-8. These parameters were chosen to ensure efficient convergence and optimization during training. For the dataset split, 70% of the IIT-JMU-UGM Audio samples were allocated for training, and the remaining 15% each for testing and validation purposes. This split allowed for robust evaluation and validation of the model's performance on unknown data. Furthermore, other types of RNNs were also explored and evaluated. The details and results of these alternative RNN models will be discussed in the subsequent section.

The implementation of the proposed metric was carried out in Python 3, leveraging various libraries such as NumPy and Librosa. The deep learning framework utilized was Keras with a Tensorflow backend. The model training and testing were performed on a DELL G3 15 laptop with the 8th generation "Intel Core i7 processor" and "Nvidia GTX GPU". This hardware configuration facilitated efficient computation and acceleration during the training process. In this environment, the feature extraction and model training process took approximately 18 minutes to complete. For predicting the quality of an individual audio clip with a duration of 8 seconds, the average prediction time was approximately 0.520 seconds. This indicates that the model is capable of providing quick quality assessments for audio samples.

4.4 Experimental Results And Analysis

This section evaluates two main aspects: the analysis of the IIT-JMU-UGM Audio Dataset, and the performance evaluation of the proposed quality assessment model.

4.4.1 Analysis of the proposed IIT-JMU-UGM Audio Dataset

In order to show that the proposed IIT-JMU-UGM Audio Dataset is comprehensive and can overcome the issues associated with the existing datasets, we conducted various empirical studies, each of which is listed below.

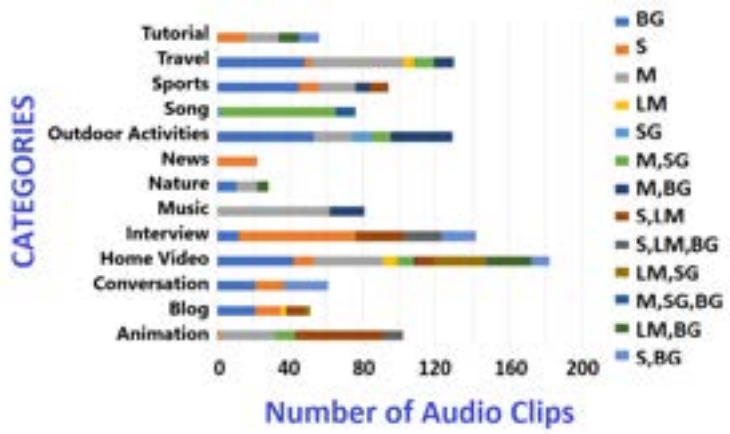


Figure 4.8: Histogram representing various categories of context and content in the IIT-JMU-UGM Audio dataset. Content is annotated as song (SG), music (M), light music (LM), speech (S), and background sounds (BG).

4.4.1.1 Database Diversity Analysis

To assess the degree of diversity present in the dataset, a histogram analysis was performed. Figure 4.8 illustrates the distribution of the dataset across thirteen distinct context categories, including tutorials, home videos, songs, music, sound effects, sports, outdoor activities, and more. These samples were captured using a variety of devices such as smartphones, handheld devices, and HD cameras, by both amateurs and professionals. The clips encompass a wide range of content, including music, songs, human voices, speech from different languages, background music, sound effects, machine sounds, animal sounds, and other ambient sounds, as indicated by the different colors in the plot. Additionally, approximately 10% of the clips implicitly exhibit various distortions resulting from factors such as varying capture device quality, environmental conditions, compression artifacts, and more. This characteristic makes the proposed dataset representative of real-world scenarios, incorporating a significant degree of generality and realism.

4.4.1.2 Subjective Testing Analysis

To validate the consistency of the subjects during the subjective testing, we made use of Fleiss' Kappa technique [178]. This technique is a statistical measure to calculate the degree of agreement between multiple subjects. The Kappa Score ($KScore$) can be obtained as;

$$KScore = \frac{\bar{M} - \bar{M}_a}{1 - \bar{M}_a}. \tag{4.12}$$

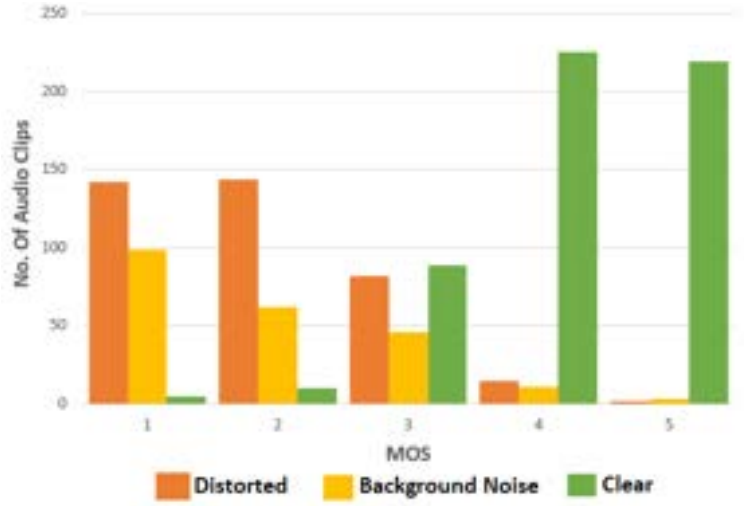


Figure 4.9: Histogram of MOS and the type of perceptual annoyance in the IIT-JMU-UGM Audio Dataset.

The value $1 - M_a$ determines the degree of agreement which is obtainable above chance. $\bar{M}$ - $\bar{M}_a$ gives the amount of agreement that is actually obtained above chance. The higher the value of $KScore$, the greater the degree of agreement between the subjects. We carried on this inter-rater reliability test for the subjective test on IIT JMU UGM Audio Database and found $KScore$ to be 0.405 which is interpreted as "Moderate agreement among the subjects". Thus, validating the reliability of our subjective testing.

In the subjective testing phase, the subjects/participants were not only required to rate the audio clips based on their quality but also to identify the type of perceptual annoyance they noticed in each clip. They could choose from three options: background noise, distorted sound (broken), or clear (no degradation). The histogram plot shown in Fig. 4.9, which depicts the relationship between the Mean Opinion Score and the type of perceptual annoyance, reveals an interesting pattern. It can be observed that audio clips with lower MOS scores had a higher number of perceptual distortions, while clips with higher MOS scores had fewer perceptual annoyances. This finding indicates that the participants were actively engaged in the testing process and paid attention to the perceptual quality and presence of distortions in the audio clips. Therefore, this further strengthens the reliability of the subjective testing procedure.

4.4.1.3 Class Balance

Figure 4.10 provides an overview of the MOS distribution across the entire dataset. It can be observed that the plot exhibits a diverse range of quality scores, indicating that

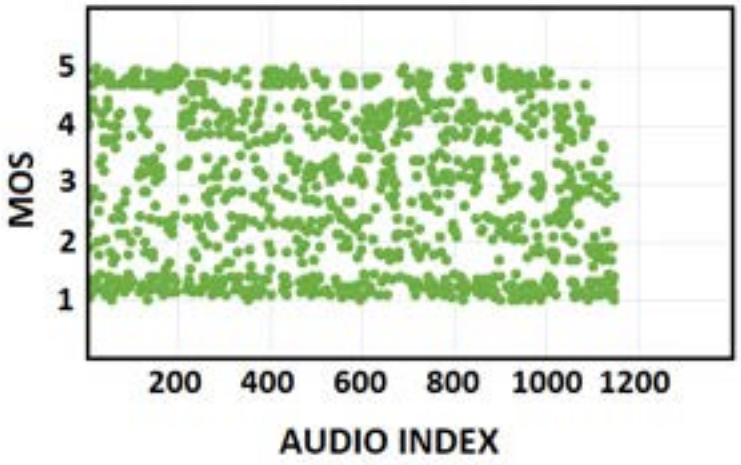


Figure 4.10: Overall distribution of MOS on the IIT-JMU-UGM Audio Dataset.

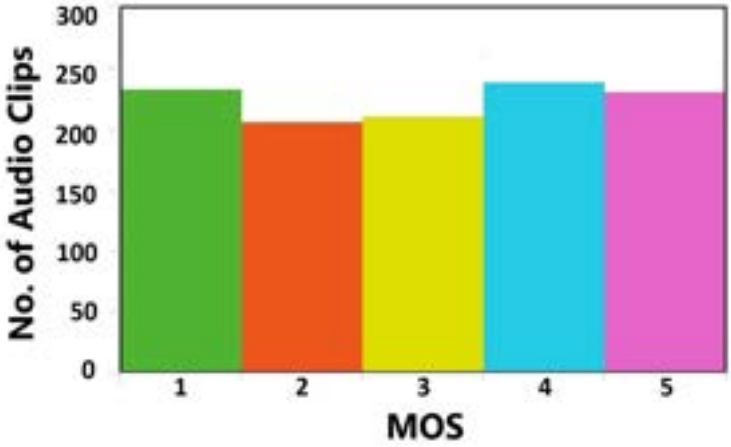


Figure 4.11: MOS distribution into five discrete quality classes on the IIT-JMU-UGM Audio Dataset.

the dataset comprises audio samples with varying levels of perceptual quality.

Furthermore, Fig. 4.11 presents a histogram of the mean opinion scores categorized into five discrete quality levels ranging from 1 to 5. The plot demonstrates that the dataset represents the different quality levels in a fairly balanced ratio. This balanced distribution of quality levels further supports the claim that the proposed dataset is comprehensive and generalized, encompassing a wide range of perceptual quality variations in the audio samples.

4.4.1.4 Effect of Bit rate

Figure 4.12 illustrates a scatter plot depicting the relationship between the bit rate and the MOS. A common perception is that higher bit rates correspond to better audio quality, while lower bit rates result in poorer audio quality. However, the plot reveals a different observation that challenges this notion. The scatter plot indicates that the relationship

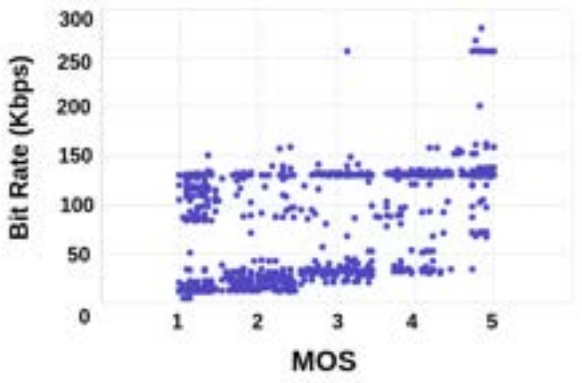


Figure 4.12: Scatter plot between bit-rate and MOS.

between audio bit rate and perceptual quality is not strictly linear or directly proportional. This finding suggests that other factors, such as background sound, context, noise, and possibly the audio encoding and compression techniques used, can significantly influence the perceived quality of UGM audio data.

This observation underscores the need for a comprehensive and generalized quality metric that takes into account multiple factors and their interactions to accurately assess the perceptual quality of UGM audio. Merely relying on bit rate as a sole indicator of quality may not capture the full complexity of the audio content and the listener's perception.

4.4.2 Result Analysis of the Proposed Quality Assessment Model

In this section, we analyze the performance of the proposed method in two ways. First, we compare the proposed model with existing SOTA techniques. Next, we perform an extensive ablation study of various hyperparameters used in the model. For this, we employ the standard co-relation evaluation methods, i.e. r, ρ, τ, and RMSE.

4.4.2.1 Performance Comparison

In order to assess the performance of our proposed method, we compared it with several SOTA QA metrics, including both intrusive and non-intrusive approaches. Among the intrusive metrics, we considered SSISDR [92], NISQA [93], ViSQOL [94], PESQ [88], and STOI [91]. Among the non-intrusive metrics, we evaluated SRMR [97], Wawenets [179], MOSNET [102], NIST-STNR [100], WADA [180], and SNRVAD [101]. Table 4.3 presents the performance of these objective quality metrics on the IIT-JMU-UGM Audio Dataset. The results show that the proposed deep learning method gives superior results compared to existing algorithms, r, ρ, and τ equal to 0.834, 0.810, of 0.683. It outperforms all the intrusive metrics, including ViSQOL, STOI, SISDR, PESQ, and NISQA. ViSQOL, despite being one of the highest-performing metrics, is limited in its applicability as it is designed for specific sample rates (48kHz). Additionally, the PESQ metric is computationally demanding. Moreover, the proposed method surpasses all the mentioned non-intrusive approaches, including Wawenets, SRMR, NIST-STNR, SNRVAD, MOSNET, and WADA. Overall, our proposed model outperforms the benchmark metrics. This can be attributed to the fact that these methods were primarily designed for speech quality assessment, while the IIT-JMU-UGM Audio Dataset is more generalized and consists of diverse sounds present in user-generated data.

4.4.2.2 Statistical Significance Test

In order to assess the statistical significance of the proposed metric, we conducted an F-Test between the proposed method and the existing metrics. The results of the F-Test are indicated in the last column of Table 4.4. It is noteworthy that the F-Score is '+1' for all cases, indicating that the proposed audio quality assessment technique is statistically superior to all the other SOTA techniques. This further validates the effectiveness of the proposed approach.

Table 4.3: Performance comparison of the proposed algorithm against various audio quality metrics.

S. No.	Metric	Type	r	ρ	τ	RMSE
1.	**Proposed Metric-GRU**	Non-Intrusive	**0.834**	**0.818**	**0.625**	**0.776**
2.	Wawenets [179]	Non-Intrusive	0.683	0.633	0.458	0.815
3.	SRMR [97]	Non-Intrusive	0.297	0.293	0.197	1.192
4.	NIST-STNR [100]	Non-Intrusive	0.273	0.255	0.174	1.276
5.	SNRVAD [101]	Non-Intrusive	0.228	0.208	0.140	1.274
6.	MOSNET [102]	Non-Intrusive	0.210	0.200	0.131	1.278
7.	WADA [180]	Non-Intrusive	0.119	0.144	0.098	1.336
1.	ViSQOL [94]	Intrusive	0.734	0.715	0.523	0.736
2.	STOI [91]	Intrusive	0.704	0.652	0.480	0.924
3.	SISDR [92]	Intrusive	0.697	0.652	0.475	1.020
4.	PESQ [88]	Intrusive	0.683	0.641	0.468	1.207
5.	NISQA [93]	Intrusive	0.410	0.332	0.232	1.106

Table 4.4: Results of the F-Test conducted between the proposed GRU metric and the various SOTA techniques

Metric	**SRMR**	**Wavenets**	**NIST-STNR**	**SNR-VAD**	**VISQOL**	**STOI**
Score	+1	+1	+1	+1	+1	+1

Table 4.5: Performance comparison of the proposed stacked GRU architecture using different number of GRU layers.

Number of Layers	**r**	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$	**RMSE**
1	0.605	0.603	0.426	1.052
2	0.736	0.703	0.518	0.971
3	**0.834**	**0.818**	**0.625**	**0.776**

4.4.2.3 Ablation Study

This section depicts an extensive ablation study to determine the effect of various parameters and conditions on the results of the proposed metric.

- **Analysis of Depth of the Network**

  The proposed model utilizes a stacked GRU architecture with three layers. The decision to use this specific number of layers is justified in Table 4.5. The ablation study conducted evaluates the model's performance using varying numbers of GRU layers. The results indicate that when using one or two layers of GRU, the model achieves r values of 0.605 and 0.736" respectively. However, with the inclusion of a third layer, there is a notable improvement in the r score, which reaches 0.834. As shown in the table, the deep GRU model consistently outperforms the shallow models on the performance parameters, indicating that the additional layers contribute to better results. Based on these findings, the use of a three-layer deep GRU architecture is justified as it yields improved results in terms of various evaluation metrics.

- **Analysis of Different Backbone Networks**

  In order to assess the relative performance of our proposed stacked GRU model, we compared it with other commonly used RNN models, namely Simple RNN, LSTM,

Table 4.6: Performance comparison of the proposed stacked GRU architecture with different RNN models.

S. No.	Model	r	ρ	τ	RMSE
1.	**GRU**	**0.834**	**0.818**	**0.625**	**0.776**
2.	**LSTM**	0.801	0.778	0.595	0.816
3.	**BiLSTM**	0.761	0.736	0.545	0.860
4.	**Simple RNN**	0.617	0.554	0.423	1.073

Table 4.7: Performance comparison of the proposed stacked GRU architecture using different optimization algorithms.

S. No.	Optimizer	r	ρ	τ	RMSE
1.	**ADAM**	**0.834**	**0.818**	**0.625**	**0.776**
2.	**RMS Prop**	0.821	0.798	0.610	0.779
3.	**AdaDelta**	0.818	0.795	0.611	0.799
4.	**SGD**	0.798	0.786	0.603	0.802

and BiLSTM. Each model was configured with specific parameters to evaluate their effectiveness. For the basic RNN model, we employed three Simple RNN layers with hidden units set to $64, 32$, and 8, and dropout values of $0.4, 0.3$, and 0.3 respectively. In the stacked LSTM model, we combined three LSTM layers with sizes of $64, 32,$, and 8, and set the dropout to 0.05, and recurrent dropout values 0.35 for all layers. The stacked BiLSTM model consisted of two layers with dropout values of 0.4 and 0.35, and recurrent dropout set to 0.35 for both layers. The final quality score was obtained from the last Dense layer with a size of 1 for all models. To prevent overfitting, k-fold validation, and standard measures were applied.

As shown in Table 4.6, the performance of the stacked GRU model outperformed the other RNN models. Additionally, it was observed that the proposed GRU model achieved convergence in approximately 18% less time compared to LSTM, and about 23% less time compared to BiLSTM. Therefore, not only did the GRU model exhibit superior performance in terms of r, but it also demonstrated faster computation time compared to the other RNN architectures.

Table 4.8: Performance comparison of the proposed stacked GRU architecture using different loss functions.

S. No.	Loss Function	r	ρ	τ	RMSE
1.	**MSE**	**0.834**	**0.818**	**0.625**	**0.776**
2.	**MSLE**	0.815	0.785	0.601	0.776
3.	**MAE**	0.796	0.768	0.588	0.858

Table 4.9: Performance of the 1-D CNN model.

Technique	r	ρ
1D CNN	0.556	0.541

- **Analysis of Optimizers**

  To assess the effect of different optimizers (Adam, RMS Prop, AdaDelta, and SGD) on the results of the proposed model, an ablation study is conducted. As depicted in Table 4.7, the Adam optimizer exhibits the highest performance, closely followed by RMSProp. The Adam optimizer combines the advantages of the gradient descent with the momentum algorithm and the Root Mean Square (RMS) Prop algorithm. It possesses favorable characteristics such as computational efficiency, easy convergence, simplicity of implementation, and low memory requirements.

- **Analysis of Loss Function**

  Furthermore, the performance of many loss functions, such as "Mean Squared Logarithmic Error (MSLE)", "Mean Absolute Error (MAE)", and "Mean Square Error (MSE)" was analyzed. The results in Table 4.8 point that MSE exhibits better performance compared to the other techniques. MSE calculates the square of the difference between the actual and predicted values. Its simplicity and efficiency make it a preferred choice for this task.

- **Analysis of CNN based Model**

  In the study, we conducted experiments using a combination of 1D CNN, Max-Pooling, Dropout, and finally Dense layer on raw audio samples. However, the obtained results, as presented in Table 4.9, were unsatisfactory. One possible explanation for this outcome is the limited kernel size of 1D CNNs, which primarily

focus on capturing local dependencies and fail to consider the influence of nonadjacent features. Consequently, the overall performance of the models is adversely affected. Thus, in contrast to the CNN model, the Recurrent Neural Networks excel at capturing temporal relationships in time-series data and demonstrate superior efficiency during both the training and inference phases.

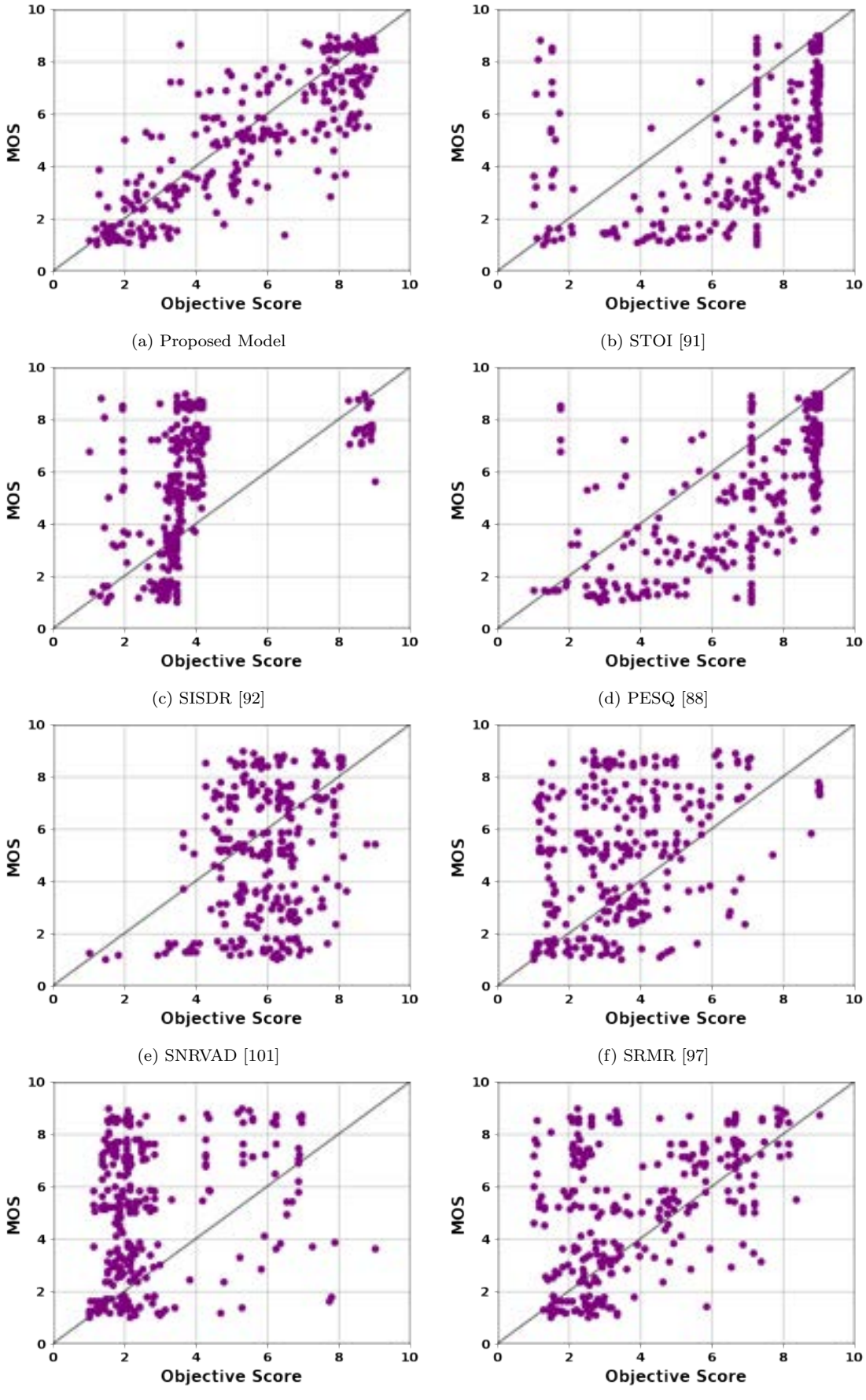
- **Analyis of Spectrogram Based Model**

  We also explored Spectrogram images to train a CNN-based deep learning model as done in [181]. However, due to the inherent diversity and variation in the user-generated audio clips, it was observed that there was a minimal correlation between the spectrograms. As a result, the obtained results were not satisfactory. The lack of consistent patterns or relationships in the spectrograms made it challenging for the CNN model to efficiently learn and obtain meaningful features from the audio data.

4.4.2.4 Scatterplot Analysis

Figure 4.13 shows the scatter plots depicting the objective scores obtained by various quality assessment metrics, including STOI, SISDR, PESQ, SNRVAD, SRMR, NIST-STMR, NISQA, and the corresponding subjective scores (MOS). For better visualization, a subset of three hundred samples was randomly selected and plotted. It is evident from the scatter plot that the proposed model exhibits a higher degree of linearity and better convergence compared to the other metrics. Such high linearity and efficient mapping between objective and subjective scores further validate the efficiency of the proposed model in accurately assessing audio quality.

In summary, the conducted experiments clearly demonstrate that the existing quality assessment metrics do not effectively correlate with subjective evaluations for a wide range of distortions in UGM audio samples. In contrast, the proposed QA metric outperforms these existing algorithms and exhibits a strong correlation with human auditory perception. By developing a deep learning-based model specifically designed for UGM audio data and utilizing the comprehensive IIT-JMU-UGM Audio Dataset, the proposed metric successfully captures the perceptual quality of diverse audio samples. The model's performance surpasses that of both intrusive and non-intrusive metrics, indicating its efficiency in evaluating UGM audio quality.



(a) Proposed Model

(b) STOI [91]

(c) SISDR [92]

(d) PESQ [88]

(e) SNRVAD [101]

(f) SRMR [97]

(g) NIST-STNR [100]

(h) NISQA [93]

Figure 4.13: Scatter plot between MOS and objective scores of different quality assessment metrics.

4.5 Conclusion

To facilitate the assessment of quality in user-generated multimedia audio, a novel dataset called IIT-JMU-UGM Audio Dataset was developed. This dataset encompasses diverse content, context, and degrees of distortions commonly found in real-time multimedia applications. Human subjective testing was conducted on this dataset to obtain ground truth information about quality. The investigation revealed that while low bit rate degradations significantly affect perceived audio quality, the relationship between bit rate and quality is not necessarily linear. Other factors such as context, content, and background sounds also have a profound impact on overall quality. To address these various factors, a robust non-intrusive quality metric was proposed, utilizing a stacked Gated Recurrent Unit architecture. The metric takes into account perceptually important audio features like MFCC, spectral centroid, chroma, and spectral contrast. When applied to the IIT-JMU-UGM Audio Dataset, the proposed metric outperformed existing SOTA intrusive and non-intrusive methods.

Chapter 5

Conclusion and Future work

5.1 Conclusion

This thesis aims to explore and propose quality assessment techniques for multimedia data. The thesis explores two important streams of multimedia, one is the DIBR view QA, and the other is the domain of UGM audio QA. The importance and necessity of quality assessment metrics are initially discussed, supported by a detailed literature review. Moreover, each of the chapters presents some drawbacks of the existing techniques and the motivation for carrying out the proposed work.

In the first chapter, an extensive literature review is presented, exploring the need, application, and scope of quality assessment metrics. Existing quality assessment methods for DIBR views are examined, and their drawbacks and limitations are highlighted. The study also delves into existing datasets and metrics for quality assessment of audio present in user-generated multimedia.

In the second chapter, a full-reference metric for the quality assessment of DIBR images is proposed. This work involves analyzing maps generated by the Non-Subsampled Contour Transform, which offers valuable quality-related features in DIBR views. A backbone CNN model is used to extract deep features from these maps. The image quality score is computed by comparing the features of the synthesized with its reference views. The results of this proposed metric outperform existing QA metrics. The main contribution of this work is the introduction of NSCT as a quality-aware feature extractor. It highlights both the edge details as well as the relevant texture information which is important in DIBR view quality assessment.

In the third chapter, a no-reference quality assessment metric is introduced. This work proposes a novel method for calculating ground truth scores for individual image blocks in an image. The predicted block-level quality values are then aggregated to determine the overall quality of the entire image. Experimental results demonstrate that the proposed algorithm outperforms existing objective metrics for DIBR synthesized views. The main contribution of this work is exploiting the fact that the human perceptual system is very sensitive to the high-intensity degradations which are specifically localized in DIBR images. Thus, by identifying and measuring these distortions, valuable insights into perceptual quality can be gained.

A summary of the results obtained by the proposed FR and NR assessment metrics is given in table 5.1 for the IETR DIBR dataset. Similarly, the results of the metrics for the IVY DIBR dataset are shown in Table 5.2. As mentioned earlier, both the proposed metrics perform much better than the existing techniques for quality assessment.

Table 5.1: Summary of results of the proposed metrics on the IETR dataset.

Technique	**r**	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$	**RMSE**
Proposed FR metric	0.8207	0.8187	0.6203	0.1417
Proposed NR metric	0.7211	0.7091	0.5114	0.1718

Table 5.2: Summary of results of the proposed metrics on the IVY dataset.

Technique	**r**	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$	**RMSE**
Proposed FR metric	0.7580	0.7375	0.5418	9.4090
Proposed NR metric	0.6693	0.6283	0.4402	10.5856

Furthermore, Figure 5.3 showcases several examples extracted from the IETR dataset. It includes the subjective scores as well as the normalized predicted scores obtained by the application of the proposed FR and NR metrics. These examples demonstrate the robust correlation between the proposed method and the subjective scores.

Chapter four focuses on the quality assessment of audio present in user-generated multimedia. It begins by addressing the limitations of existing audio databases and the need for a new database. With this motivation, a novel repository is developed consisting of

DMOS: 0.908
Proposed NSCT-FR: 0.907
Proposed NR: 0.960

DMOS: 0.754
Proposed NSCT-FR: 0.724
Proposed NR: 0.839

DMOS: 0.993
Proposed NSCT-FR: 0.919
Proposed NR: 0.878

Table 5.3: Examples from the IETR dataset showing a comparison of the subjective score (DMOS) and the predicted scores obtained by the proposed metric.

diverse UGM audio data. The clean reference signals are subjected to various distortions to mimic the degradations present in real-world scenarios. The proposed audio database incorporates samples with implicit distortions, and diverse contexts, content, distortion types, and intensities. To establish ground truth scores, an extensive subjective assessment test is conducted. Next, the work delves into the utilization of a deep learning model for QA. It explores various techniques for assessing audio quality, specifically the use of Recurrent Neural networks is capitalized which yields significantly improved results when compared to existing techniques for audio quality assessment.

5.2 Future Work

In the future, this work can be extended in various directions of quality assessment and its applications. Some of the domains in which the work may be extended are listed below.

1. **Joint DIBR quality assessment and enhancement model:**
   A possible domain for exploration is a joint quality assessment and enhancement which is an integrated framework that simultaneously evaluates the quality of data while enhancing it. By combining both evaluation and enhancement processes, this unified approach aims to optimize the overall image quality, ensuring superior results for a wide range of applications and datasets. The workflow of a comprehensive quality enhancement and assessment model can be outlined in Fig. 5.1. In this approach, the predicted quality score of the enhanced image serves as a loss function during the training of the corresponding enhancement model. Thus the process aims to replicate the enhancements based on the preferences of the human perceptual system.

2. **Advancements in AQA for UGM data:**
   Chapter 4 introduced the IIT-JMU-UGM Audio Dataset as a benchmark for UGM AQA. Although it served as a successful benchmark, the dataset has some limitations. Firstly, it mainly focused on only two explicit distortion types: low bit rate and background noise. Additionally, only about 10% of the samples included real-world degradations, thus limiting diversity for quality assessment. In this regard, in our next work presented in [8], a few improvements are introduced.

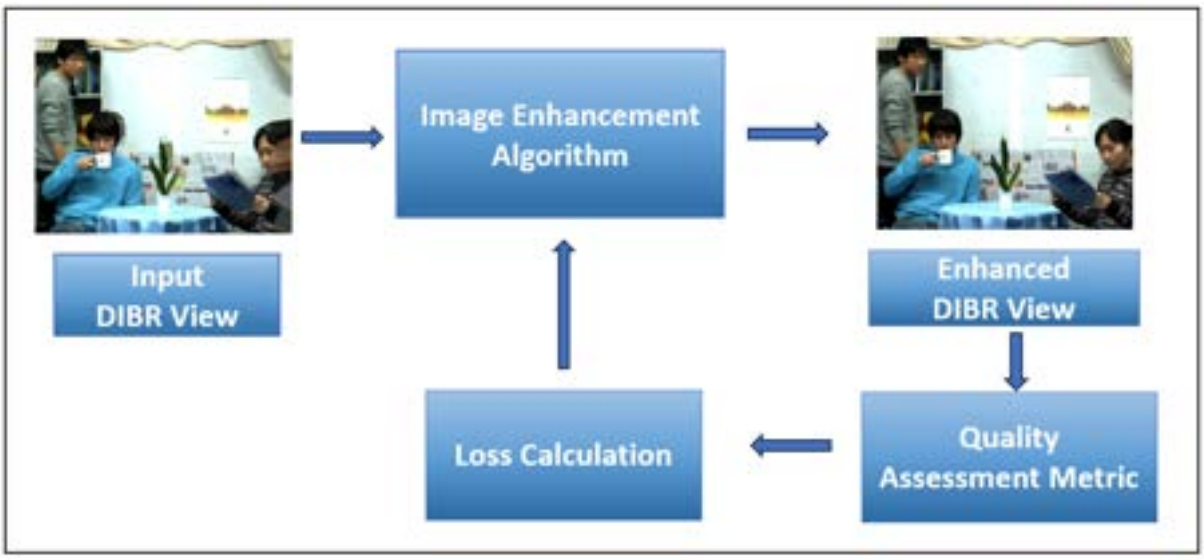


Figure 5.1: The proposed joint DIBR quality assessment and enhancement model.

First, an enhanced version of the earlier proposed dataset called IIT-JMU-UGM Audio Dataset-2 is developed. This repository includes a wider range of real-world scenarios by incorporating audio clips with diverse contexts, content, distortion types, and intensities. It also incorporates implicitly distorted audio alongside the original dataset. The final dataset consists of more than two thousand audio samples, each accompanied by their respective subjective scores.

Next, the proposed work also incorporates the concept of Transformer-based learning (Fig. 5.2, which provides an efficient and non-intrusive technique for evaluating the quality of UGM audio. This novel approach surpasses the performance of existing SOTA algorithms by achieving a performance improvement of over 4%.

Along the same lines, other powerful and advanced deep learning concepts can be explored in the future for developing enhanced audio quality assessment techniques for UGM data.

3. **Video quality assessment metric:**

   As video data is an extension of images, the future methods for assessing DIBR image quality can be extended and made more efficient for video quality assessment. This would entail evaluating both traditional (static) distortions and temporal distortions present in the video data. A more precise evaluation of such datasets can be achieved by combining local and global quality scores and incorporating temporal quality cues.

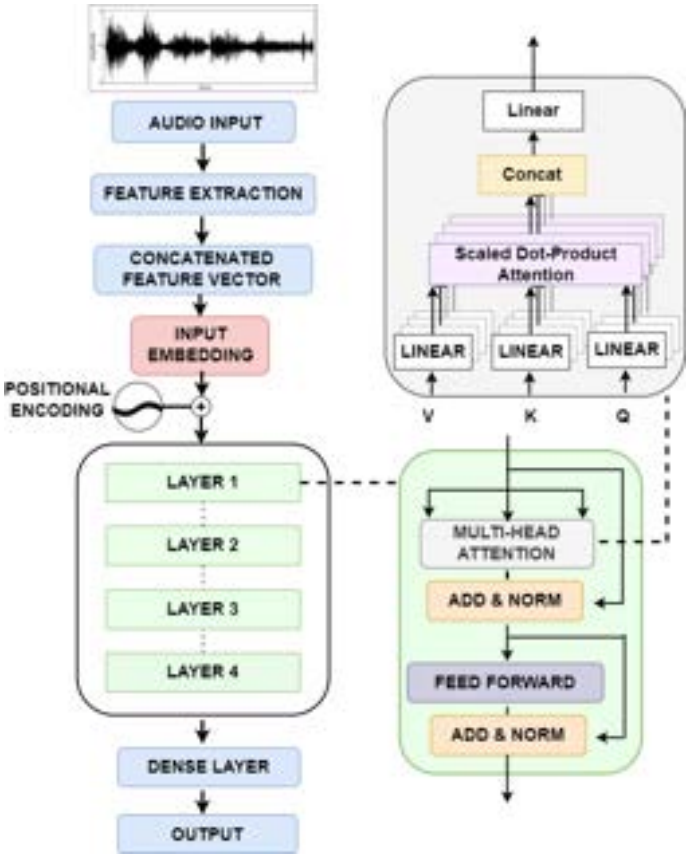


Figure 5.2: Architecture of the Transformer-based quality assessment metric proposed in [8].

4. **A comprehensive multimedia quality assessment metric:**
   An additional area worthy of exploration is the development of a perceptual quality metric that holistically evaluates the overall quality of multimedia clips, encompassing video, audio, images, 3-D views, graphics, and screen content. Such a metric would take into account real-world scenarios where end-users seek a satisfying quality of experience. The effect on content and context may also be explored in such studies. This comprehensive metric may find utility in various domains, including movie theatres, gaming, Over-the-top (OTT) media, and other immersive applications that involve the integration of multimodal media elements.

5. Furthermore, potential improvements in model architecture such as Transformers, and Diffusion models can be pursued, potentially leading to enhanced performance in DIBR view quality assessment.

Bibliography

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[2] Y. Hu and P. C. Loizou, "Subjective Comparison of Speech Enhancement Algorithms," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006, pp. I–I.

[3] A. Schmidt-Nielsen, E. Marsh, J. Tardelli, P. Gatewood, E. Kreamer, T. Tremain, C. Cieri, and J. Wright, "Speech in noisy environments (SPINE) training audio LDC2000S87," *Linguistic Data Consortium (LDC). University of Pennsylvania*, 2000.

[4] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1992.

[5] A. L. Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: theory, design, and applications," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3089–3101, 2006.

[6] Sadbhawna, V. Jakhetiya, D. Mumtaz, B. N. Subudhi, and S. C. Guntuku, "Stretching artifacts identification for quality assessment of 3d-synthesized views," *IEEE Transactions on Image Processing*, vol. 31, pp. 1737–1750, 2022.

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[8] D. Mumtaz, A. Jena, V. Jakhetiya, K. Nathwani, and S. C. Guntuku, "Transformer-based quality assessment model for generalized user-generated multimedia audio content," in *Proc. Interspeech 2022*, 2022, pp. 674–678.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[12] A. Tai, M. Binjie, X. Bin, H. Chunlei, X. Shiming, and P. Chunhong, "Patch loss: A generic multi-scale perceptual loss for single image super-resolution," *Pattern Recognition*, vol. 139, p. 109510, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323002108

[13] J. Kim and C. Lee, "Reliable perceptual loss computation for gan-based super-resolution with edge texture metric," *IEEE Access*, vol. 9, pp. 120 127–120 137, 2021.

[14] D. Amir and Y. Weiss, "Understanding and simplifying perceptual distances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 226–12 235.

[15] A. Q. de Oliveira, M. Walter, and C. R. Jung, "An artifact-type aware dibr method for view synthesis," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1705–1709, 2018.

[16] C. Fehn, "A 3d-tv approach using depth-image-based rendering (dibr)," in *Proc. of VIIP*, vol. 3, no. 3, 2003.

[17] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[18] S. Tian, L. Zhang, L. Morin, and O. Deforges, "A benchmark of dibr synthesized view quality assessment metrics on a new database for immersive media applications," *IEEE Transactions on Multimedia*, vol. 21, no. 5, p. 1235–1247, 2019.

[19] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-D synthesized view assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343, 2011.

[20] Y. Jung, H. G. Kim, and Y. M. Ro, "Critical binocular asymmetry measure for the perceptual quality assessment of synthesized stereo 3d images in view synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 7, pp. 1201–1214, 2016.

[21] I. Ahn and C. Kim, "A novel depth-based virtual view synthesis method for free viewpoint video," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 614–626, 2013.

[22] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[23] M. Solh and G. AlRegib, "Hierarchical hole-filling for depth-based view synthesis in ftv and 3d video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 495–504, 2012.

[24] G. Luo, Y. Zhu, Z. Li, and L. Zhang, "A hole filling approach based on background reconstruction for view synthesis in 3d video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1781–1789.

[25] D. Wang, Y. Zhao, Z. Wang, and H. Chen, "Hole-filling for dibr based on depth and gradient information," *International Journal of Advanced Robotic Systems*, vol. 12, no. 2, 2015.

[26] C. Zhu and S. Li, "Depth image based view synthesis: New insights and perspectives on hole generation and filling," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 82–93, 2016.

[27] O. Stankiewicz, K. Wegner, M. Tanimoto, and M. Domański, "Enhanced view synthesis reference software (vsrs) for free-viewpoint television," 2013.

[28] R. Song, H. Ko, and C.-C. J. Kuo, "Mcl-3d: A database for stereoscopic image quality assessment using 2d-image-plus-depth source," *Journal of Information Science and Engineering*, vol. 31, 03 2014.

[29] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Dibr synthesized image quality assessment based on morphological wavelets," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.

[30] S. Tian, L. Zhang, L. Morin, and O. Deforges, "A full-reference image quality assessment metric for 3-d synthesized views," *Proc. Image Quality Syst. Perform. Conf., IST Electron. Imag., Soc. Imaging Sci.Technol.*, vol. 12, p. 3661–3665, 2018.

[31] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Multi–scale synthesized view assessment based on morphological pyramids," *Journal of Electrical Engineering*, vol. 67, pp. 1–9, 01 2016.

[32] L. Li, Y. Zhou, K. Gu, W. Lin, and S. Wang, "Quality assessment of dibr-synthesized images by measuring local geometric distortions and global sharpness," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 914–926, 2017.

[33] L. Li, Y. Zhou, J. Wu, F. Li, and G. Shi, "Quality index for view synthesis by measuring instance degradation and global appearance," *IEEE Transactions on Multimedia*, vol. 23, pp. 320–332, 2021.

[34] S. Tian, L. Zhang, L. Morin, and O. Déforges, "Sc-iqa: Shift compensation based image quality assessment for dibr-synthesized views," in *2018 IEEE Visual Communications and Image Processing (VCIP)*, 2018, pp. 1–4.

[35] S. Mahmoudpour and P. Schelkens, "Synthesized view quality assessment using feature matching and superpixel difference," *IEEE Signal Processing Letters*, vol. 27, pp. 1650–1654, 2020.

[36] S. Chaudhary, A. Mazumder, D. Mumtaz, V. Jakhetiya, and B. N. Subudhi, "Perceptual quality assessment of dibr synthesized views using saliency based deep features," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2628–2632.

[37] S. Ling, J. Li, P. L. Callet, and J. Wang, "Perceptual representations of structural information in images: Application to quality assessment of synthesized view in ftv scenario," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1735–1739.

[38] Sadbhawna, V. Jakhetiya, S. Chaudhary, B. N. Subudhi, W. Lin, and S. C. Guntuku, "Perceptually unimportant information reduction and cosine similarity-based quality assessment of 3d-synthesized images," *IEEE Transactions on Image Processing*, vol. 31, pp. 2027–2039, 2022.

[39] S. Ling and P. L. Callet, "Image quality assessment for dibr synthesized views using elastic metric." Proceedings of the 25th ACM International Conference on Multimedia, 2017, p. 1157–1163.

[40] Y. Zhang, H. Zhang, M. Yu, S. Kwong, and Y. Ho, "Sparse representation-based video quality assessment for synthesized 3d videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 509–524, 2020.

[41] Sadbhawna, V. Jakhetiya, B. Subudhi, S. Jaiswal, L. Li, and W. Lin, "Context region identification based quality assessment of 3d synthesized views," *IEEE Transactions on Multimedia*, pp. 1–11, 2022.

[42] P. Zhenyu, P. Qiuping, S. Feng, G. Wei, and L. Weisi, "Lggd+: Image retargeting quality assessment by measuring local and global geometric distortions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3422–3437, 2022.

[43] S. Mahmoudpour and P. Schelkens, "Unifying structural and semantic similarities for quality assessment of dibr-synthesized views," *IEEE Access*, vol. 10, pp. 59 026–59 036, 2022.

[44] K. Ding, Y. Liu, X. Zou, S. Wang, and K. Ma, "Locally adaptive structure and texture similarity for image quality assessment," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. ACMM '21. Association for Computing Machinery, 2021, p. 2483–2491.

[45] W. Xuejin, S. Feng, J. Qiuping, M. Xiangchao, and H. Yo-Sung, "Measuring coarse-to-fine texture and geometric distortions for quality assessment of dibr-synthesized images," *IEEE Transactions on Multimedia*, vol. 23, pp. 1173–1186, 2021.

[46] R. Soundararajan and A. C. Bovik, "Rred indices: Reduced reference entropic differencing for image quality assessment," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517–526, 2012.

[47] S. Mahmoudpour and P. Schelkens, "Reduced-reference image quality assessment based on internal generative mechanism utilizing shearlets and rényi entropy analysis," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–5.

[48] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "Speed-qa: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, 2017.

[49] J. Wu, Y. Liu, G. Shi, and W. Lin, "Saliency change based reduced reference image quality assessment," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4.

[50] K. Gu, V. Jakhetiya, J. Qiao, X. Li, W. Lin, and D. Thalmann, "Model-based referenceless quality metric of 3d synthesized images using local image description," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 394–405, 2018.

[51] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

[52] V. Jakhetiya, K. Gu, S. P. Jaiswal, T. Singhal, and Z. Xia, "Kernel-ridge regression-based quality measure and enhancement of three-dimensional-synthesized images," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 1, pp. 423–433, 2020.

[53] S. Tian, L. Zhang, L. Morin, and O. Déforges, "Niqsv+: A no-reference synthesized view quality assessment metric," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1652–1664, 2018.

[54] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, and P. Le Callet, "Multiscale natural scene statistical analysis for no-reference quality evaluation of dibr-synthesized views," *IEEE Transactions on Broadcasting*, vol. 66, no. 1, pp. 127–139, 2020.

[55] D. D. K. D. D. Sandić-Stanković and P. L. Callet, "Fast blind quality assessment of dibr-synthesized video based on high-high wavelet subband," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5524–5536, 2019.

[56] G. Wang, Z. Wang, K. Gu, and Z. Xia, "Blind quality assessment for 3d-synthesized images by measuring geometric distortions and image complexity," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4040–4044.

[57] G. Wang, Z. Wang, K. Gu, L. Li, Z. Xia, and L. Wu, "Blind quality metric of dibr-synthesized images in the discrete wavelet transform domain," *IEEE Transactions on Image Processing*, vol. 29, pp. 1802–1814, 2020.

[58] L. Li, Y. Huang, J. Wu, K. Gu, and Y. Fang, "Predicting the quality of view synthesis with color-depth image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2509–2521, 2021.

[59] J. Yan, Y. Fang, R. Du, Y. Zeng, and Y. Zuo, "No reference quality assessment for 3d synthesized views by local structure variation and global naturalness change," *IEEE Transactions on Image Processing*, vol. 29, pp. 7443–7453, 2020.

[60] S. Ling, J. Li, Z. Che, J. Wang, W. Zhou, and P. L. Callet, "Re-visiting discriminator for blind free-viewpoint image quality assessment," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[61] G. Yue, C. Hou, K. Gu, T. Zhou, and G. Zhai, "Combining local and global measures for dibr-synthesized image quality evaluation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 2075–2088, 2019.

[62] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, and X. Gao, "No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4566–4579, 2019.

[63] X. Wang, K. Wang, B. Yang, F. W. Li, and X. Liang, "Deep blind synthesized image quality assessment with contextual multi-level feature pooling," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 435–439.

[64] S. Lao, Y. Gong, S. Shi, S. Yang, T. Wu, J. Wang, W. Xia, and Y. Yang, "Attentions help cnns see better: Attention-based hybrid image quality assessment network," *arXiv preprint arXiv:2204.10485*, 2022.

[65] M. Cheon, S. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *CVPR Workshop*, 2021, pp. 433–442.

[66] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[67] Z. Lin, Z. Lei, M. Xuanqin, and M. David, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[68] Z. Dongsheng, Z. Huan, C. Jiangzhong, Z. Xu, Y. Ximei, and L. Kuen, "Evaluating quality of dibr-synthesized views based on texture and perceptual hashing similarity," in *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*, 2022, pp. 1–6.

[69] Sadbhawna, V. Jakhetiya, D. Mumtaz, and S. P. Jaiswal, "Distortion specific contrast based no-reference quality assessment of dibr-synthesized views," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–5.

[70] G. Wang, Z. Wang, K. Gu, L. Li, Z. Xia, and L. Wu, "Blind quality metric of DIBR-synthesized images in the discrete wavelet transform domain," *IEEE Transactions on Image Processing*, vol. 29, pp. 1802–1814, 2020.

[71] I. Rec, "P. supplement 23: ITU-T coded-speech database," *International Telecommunication Union*, 1998.

[72] B. M. Fazenda, P. Kendrick, T. J. Cox, F. Li, and I. Jackson, "Perception and automated assessment of audio quality in user generated content: An improved model," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[73] E. Rothauser, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[74] C. Creusere, K. Kallakuri, and R. Vanam, "An Objective Metric of Human Subjective Audio Quality Optimized for a Wide Range of Audio Fidelities," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 129 – 136, 02 2008.

[75] J. F. Santos and T. H. Falk, "Towards the development of a non-intrusive objective quality measure for DNN-enhanced speech," in *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.

[76] Z. Li, J.-C. Wang, J. Cai, Z. Duan, H.-M. Wang, and Y. Wang, "Non-Reference Audio Quality Assessment for Online Live Music Recordings," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. Association for Computing Machinery, 2013, p. 63–72.

[77] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 631–635.

[78] Team, CoreSV, "CoreSV Listening Test," 2014. [Online]. Available: http://listening-test.coresv.net/results.htm.

[79] G. Waters, "Sound quality assessment material—recordings for subjective tests: User's handbook for the EBU-SQAM compact disk," *European Broadcasting Union (EBU), Tech. Rep*, pp. 1–13, 1988.

[80] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE Challenge — Corpus description and performance evaluation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[81] Itu. [Online]. Available: https://extranet.itu.int/brdocsearch

[82] H. Martinez and M. Farias, "A no-reference audio-visual video quality metric," *Journal of Electronic Imaging*, vol. 23, p. 061108, 11 2014.

[83] Bizzard. [Online]. Available: http://www.festvox.org/blizzard/index.html

[84] M. K. Nandwana, J. Van Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The voices from a distance challenge 2019 evaluation plan," *arXiv preprint arXiv:1902.10828*, 2019.

[85] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[86] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*.

[87] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[88] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) -a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[89] J. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II-Perceptual Model," *AES: Journal of the Audio Engineering Society*, vol. 61, pp. 385–402, 06 2013.

[90] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[91] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[92] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[93] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131.

[94] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[95] P. Manocha, A. Finkelstein, R. Zhang, N. J.Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," in *Interspeech*, Oct. 2020.

[96] J. Serrà, J. Pons, and S. Pascual, "Sesqa: Semi-supervised learning for speech quality assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 381–385.

[97] T. H. Falk, C. Zheng, and W. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[98] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-

domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[99] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, 2020.

[100] NIST. [Online]. Available: https://labrosa.ee.columbia.edu/~dpwe/tmp/nist/doc/stnr.txt

[101] SNRVAD. [Online]. Available: https://labrosa.ee.columbia.edu/projects/snreval/

[102] C. Lo, W. Fu, W. C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 09, pp. 1541–1545.

[103] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks," in *Interspeech 2016*, pp. 342–346.

[104] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 316–320.

[105] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[106] W.-C. Tseng, C.-Y. Huang, W.-T. Kao, Y. Y. Lin, and H. yi Lee, "Utilizing self-supervised representations for mos prediction," in *Proc. Interspeech*, 08 2021, pp. 2781–2785.

[107] M. Yu, C. Zhang, Y. Xu, S.-X. Zhang, and D. Yu, "MetricNet: Towards Improved Modeling For Non-Intrusive Speech Quality Assessment," in *Proc. Interspeech 2021*, 2021, pp. 2142–2146.

[108] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020.

[109] P. Manocha, A. Kumar, B. Xu, A. Menon, I. D. Gebru, V. K. Ithapu, and P. Calamia, "SAQAM: Spatial Audio Quality Assessment Metric," in *Proc. Interspeech 2022*, 2022, pp. 649–653.

[110] Q. Huang and T. Hain, "Exploration of Audio Quality Assessment and Anomaly Localisation Using Attention Models," in *Proc. Interspeech 2020*, 2020, pp. 4611–4615.

[111] X. Dong and D. S. Williamson, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in *Proc. Interspeech*, 2020, pp. 4631–4635.

[112] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Warp-q: Quality prediction for generative neural speech codecs," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 401–405.

[113] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2019.

[114] N. Nessler, M. Cernak, P. Prandoni, and P. Mainar, "Non-Intrusive Speech Quality Assessment with Transfer Learning and Subject-Specific Scaling," in *Proc. Interspeech 2021*, 2021, pp. 2406–2410.

[115] A. Ragano, E. Benetos, and A. Hines, "More for less: Non-intrusive speech quality assessment with limited annotations," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2021, pp. 103–108.

[116] Sadbhawna, V. Jakhetiya, S. Chaudhary, B. N. Subudhi, W. Lin, and S. C. Guntuku, "Perceptually unimportant information reduction and cosine similarity-based

quality assessment of 3d-synthesized images," *IEEE Transactions on Image Processing*, vol. 31, pp. 2027–2039, 2022.

[117] S. Chaudhary, A. Mazumder, D. Mumtaz, V. Jakhetiya, and B. Subudhi, "Perceptual quality assessment of dibr synthesized views using saliency based deep features," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2628–2632.

[118] S. Montabone and A. Soto, "Human detection using a mobile platform and novel features derived from a visual saliency mechanism," *Image and Vision Computing*, vol. 28, no. 3, pp. 391–402, 2010.

[119] J. Yan, Y. Fang, R. Du, Y. Zeng, and Y. Zuo, "No reference quality assessment for 3d synthesized views by local structure variation and global naturalness change," *IEEE Transactions on Image Processing*, vol. 29, pp. 7443–7453, 2020.

[120] J. Zhou, A. L. Cunha, and M. N. Do, "Nonsubsampled contourlet transform: construction and application in enhancement," in *IEEE International Conference on Image Processing*, vol. 1, 2005, pp. I–469.

[121] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint, arXiv:1409.1556*, 2014.

[122] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

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

[124] C. Ji, Z. Peng, W. Zou, F. Chen, G. Jiang, and M. Yu, "No-reference quality assessment for 3d synthesized images based on visual-entropy-guided multi-layer features analysis." *Entropy*, vol. 23, no. 6, p. 770, 2021.

[125] X. Sui, M. Ding, J. Yan, Y. Fang, Y. Zuo, and Z. Tan, "Objective quality assessment of synthesized images by local variation measurement," *Signal Processing: Image Communication*, vol. 92, p. 116096, 2021.

[126] L. Li, X. Chen, Y. Zhou, J. Wu, and G. Shi, "Depth image quality assessment for view synthesis based on weighted edge similarity," in *CVPR Workshops*, 2019, pp. 17–25.

[127] X. Wang, K. Wang, B. Yang, F. W. B. Li, and X. Liang, "Deep blind synthesized image quality assessment with contextual multi-level feature pooling," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 435–439.

[128] M. S. Farid, M. Lucenteforte, and M. Grangetto, "Perceptual quality assessment of 3d synthesized images," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 505–510.

[129] K. Gu, G. Zhai, X. Yang, and W. Zhang, "An efficient color image quality metric with local-tuned-global model," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 506–510.

[130] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2011.

[131] P. H. Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8288, pp. 53–, 02 2012.

[132] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[133] F. Battisti, E. Bosc, M. Carli, P. L. Callet, and S. Perugia, "Objective image quality assessment of 3d synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78–88, 2015.

[134] S. Mahmoudpour and P. Schelkens, "Synthesized view quality assessment using feature matching and superpixel difference," *IEEE Signal Processing Letters*, vol. 27, pp. 1650–1654, 2020.

[135] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, p. 209–212, 2013.

[136] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[137] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 6, p. 2957–2971, 2017.

[138] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3664–3673.

[139] S. Christian, V. Vincent, I. Sergey, S. Jonathon, and W. Zbigniew, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.

[140] H. Andrew, Z. Menglong, C. Bo, K. Dmitry, W. Weijun, W. Tobias, A. Marco, and A. Hartwig, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv: 1704.04861*, 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[141] F. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[142] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, 2012.

[143] S. Christian, I. Sergey, V. Vincent, and A. Alex, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[144] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[145] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[146] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[147] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV: 16th European Conference*. Springer, 2020, pp. 491–507.

[148] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[149] P. C. Madhusudana and R. Soundararajan, "Subjective and objective quality assessment of stitched images for virtual reality," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5620–5635, 2019.

[150] Y. Li, L. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, 2016, pp. 685–689.

[151] V. Jakhetiya, K. Gu, T. Singhal, S. C. Guntuku, Z. Xia, and W. Lin, "A highly efficient blind image quality assessment metric of 3-d synthesized images using outlier detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4120–4128, 2019.

[152] S. Tian, L. Zhang, L. Morin, and O. Deforges, "Niqsv: A no reference image quality assessment metric for 3d synthesized views," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 1248–1252.

[153] M. S. Farid, M. Lucenteforte, and M. Grangetto, "Objective quality metric for 3d virtual views," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3720–3724.

[154] The consumer digital video library. [Online]. Available: https://cdvl.org/

[155] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Transactions on Multimedia*, vol. 8, 11 2006.

[156] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[157] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[158] I. Martín-Morató, M. Cobos, and F. J. Ferri, "Adaptive mid-term representations for robust audio event classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2381–2392, 2018.

[159] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[160] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of Voice Quality Features in Detecting Depression," in *Proc. Interspeech 2018*, 2018, pp. 1676–1680.

[161] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2020.

[162] Y. Shan, J. Wang, X. Xie, L. Meng, and J. Kuang, "Non-intrusive speech quality assessment using deep belief network and backpropagation neural network," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 11 2018, pp. 71–75.

[163] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO IST Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[164] F. Zalkow and M. Müller, "Ctc-based learning of chroma features for score–audio music retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2957–2971, 2021.

[165] M. Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015, vol. 5.

[166] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[167] T. Takahashi, T. Hori, C. M. Wilk, and S. Sagayama, "Semi-supervised nmf in the chroma domain applied to music harmony estimation," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1636–1641.

[168] K. O'Hanlon and M. B. Sandler, "Comparing cqt and reassignment based chroma features for template-based automatic chord recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 860–864.

[169] D.-N. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai, "Music type classification by spectral contrast feature," *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 113–116 vol.1, 2002.

[170] M. Xia, H. Shao, X. Ma, and C. W. de Silva, "A stacked gru-rnn-based approach for predicting renewable energy and electricity load for smart grid operation," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 7050–7059, 2021.

[171] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[172] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[173] Y. Gao and D. Glowacka, "Deep gate recurrent neural network," in *Proceedings of The 8th Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 63. PMLR, 16–18 Nov 2016, pp. 350–365.

[174] A. A. Wazrah and S. Alhumoud, "Sentiment analysis using stacked gated recurrent unit for arabic tweets," *IEEE Access*, vol. 9, pp. 137 176–137 187, 2021.

[175] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[176] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[177] T. Chaudhuri, M. Wu, Y. Zhang, P. Liu, and X. Li, "An attention-based deep sequential gru model for sensor drift compensation," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7908–7917, 2020.

[178] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[179] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 331–335.

[180] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, 2008, pp. 2598–2601.

[181] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, pp. 535–544, 2020.

List of publications

Journals

1. **D. Mumtaz**, V. Jakhetiya, K. Nathwani, B. N. Subudhi and S.C. Guntuku, "Non-Intrusive Perceptual Audio Quality Assessment for User-Generated Content Using Deep Learning," IEEE Transactions on Industrial Informatics, doi:10.1109/TII.2021.3139010. **(Impact factor: 12.3)**

2. **D. Mumtaz**, Sadbhawna, V. Jakhetiya, B. N. Subudhi, W. Lin, "Non-Subsampled Contourlet Transform and Ground-truth Score Generation based Quality Assessment for DIBR-Synthesized Views" Submitted after Major Revision, IEEE Transactions on Multimedia **(Impact factor: 7.3)**

3. Sadbhawna, V. Jakhetiya, **D. Mumtaz**,, B. N. Subudhi and S. C. Guntuku, "Stretching Artifacts Identification for Quality Assessment of 3D-Synthesized Views," in IEEE Transactions on Image Processing, vol. 31, pp. 1737-1750, 2022. **(Impact factor: 10.6)**

Conferences

1. **D. Mumtaz**,, A. Jena, V. Jakhetiya, K. Nathwani, and S. C. Guntuku, "Transformer-Based Quality Assessment Model For Generalized User-Generated Multimedia Audio Content", in Proc. Interspeech 2022, 2022, pp. 674–678 **(A Ranking Conference)**

2. Sadbhawna, V. Jakhetiya, B.N. Subudhi, H. Shakya, **D. Mumtaz**,, "Do we need a new large-scale quality assessment database for Generative Inpainting based 3D View Synthesis?" Accepted in AAAI 2022 Student Abstract.

3. S. Chaudhary, A. Mazumder, **D. Mumtaz**, V. Jakhetiya, B.N. Subudhi, “Perceptual Quality Assessment of DIBR Synthesized Views Using Saliency Based Deep Features”, 2021 IEEE International Conference on Image Processing (ICIP), 2628-2632

4. Sadbhawna, V. Jakhetiya, **D. Mumtaz**, and S. P. Jaiswal, ”Distortion Specific Contrast Based No-Reference Quality Assessment of DIBR-Synthesized Views,” 22nd International Workshop on Multimedia Signal Processing (MMSP), 2020, pp. 1-5.