# THÈSE

Pour l'obtention du grade de

## DOCTEUR DE L'UNIVERSITÉ DE POITIERS

FACULTÉ DES SCIENCES FONDAMENTALES ET APPLIQUÉES

DIPLÔME NATIONAL - ARRÊTÉ DU 25 MAI 2016

Ecole Doctorale : MATHÉMATIQUE, INFORMATIQUE, MATÉRIAUX, MÉCANIQUE, ÉNERGÉTIQUE

Secteur de Recherche : **Traitement du Signal et des Images**

PRÉSENTÉE PAR :

**ABDERREZZAQ SENDJASNI**

# ÉVALUATION OBJECTIVE ET SUBJECTIVE DE LA QUALITÉ DES IMAGES À 360 DEGRÉS

Directeur de Thèse :
**MOHAMED-CHAKER LARABI**

Soutenue le 05 Janvier 2023
Devant la Commission d'Examen

**JURY**

| | | |
|---|---|---|
| Aljosa SMOLIC | Professeur, Luzern Univ. of Applied Sciences and Arts, Suisse | Rapporteur |
| Sophie TRIANTAPHILLIDOU | Professeur, Univ. of Westminister, Royaume-Uni | Rapporteure |
| Laura TONI | Professeur, UCL, Royaume-Uni | Examinatrice |
| Majdi KHOUDEIR | Professeur, Univ. de Poitiers, France | Examinateur |
| Steven Yves LE MOAN | Professeur, NTNU, Norvège | Examinateur |
| Mohamed-Chaker LARABI | Maitre de Conférences HDR, Univ. de Poitiers, France | Examinateur |
| Faouzi ALAYA CHEIKH | Professeur, NTNU, Norvège | Examinateur |

# Abstract

360-degree images, *a.k.a.* omnidirectional images, are in the center of immersive media. With the increase in demands of the latter, mainly thanks to the offered interactive and immersive experience, it is paramount to provide good quality of experience (QoE). This QoE is significantly impacted by the quality of the content. Like any type of visual signal, 360-degree images go through a sequence of processes including encoding, transmission, decoding, and rendering. Each of these processes has the potential to introduce distortions to the content. To improve the QoE, image quality assessment (IQA) is one of the strategies to be followed. This thesis addresses the quality evaluation of 360-degree images from the objective and subjective perspectives. By focusing on the influence of Head Mounted Displays (HMDs) on the perceived quality of 360-degree images, a psycho-visual study is designed and carried out using four different devices. For this purpose, a 360-degree image datasets is created and a panel of observers is involved. The impact of HMDs on the quality ratings is identified and highlighted as an important factor to consider when conducting subjective experiments for 360-degree images. From the objective perspective, we first comprehensively benchmarked several convolutional neural network (CNN) models under various configurations. Then, the processing chain of CNN-based 360-IQA is improved at different scales, from input sampling and representation to aggregating quality scores. Based on the observations of the above studies as well as the benchmark, two 360-IQA models based on CNNs are proposed to accurately predict the quality of 360-degree images. The obtained observations and conclusions from the various contributions shall bring insights for assessing the quality of 360-degree images.

***Keywords*: 360-degree images, Image quality assessment, Visual perception, Convolutional Neural Networks, objective and subjective quality.**

# Résumé

Les images à 360 degrés, aussi appelées images omnidirectionnelles, sont au cœur des contenus immersifs. Avec l'augmentation de leur utilisation, notamment grâce à l'expérience interactive et immersive qu'ils offrent, il est primordial de garantir une bonne qualité d'expérience (QoE). Cette dernière est considérablement impactée par la qualité du contenu lui-même. En l'occurrence, les images à 360 degrés, comme tout type de signal visuel, passent par une séquence de processus comprenant l'encodage, la transmission, le décodage et le rendu. Chacun de ces processus est susceptible d'introduire des distorsions dans le contenu. Pour améliorer la qualité d'expérience, toutes ces dégradations potentielles doivent être soigneusement prises en compte et réduites à un niveau imperceptible. Pour atteindre cet objectif, l'évaluation de la qualité de l'image est l'une des stratégies devant être utilisée. Cette thèse aborde

l'évaluation de la qualité des images à 360 degrés des points de vue objectifs et subjectif. Ainsi, en s'intéressant à l'effet des visiocasques sur la qualité perçue des images 360 degrés, une étude psycho-visuelle est conçue et réalisée en utilisant quatre dispositifs différents. À cette fin, une base de données a été créé et un panel d'observateurs a été impliqué. L'impact des visiocasques sur la qualité a été identifié et mis en évidence comme un facteur important à prendre en compte lors de la réalisation d'expériences subjectives pour des images à 360 degrés. D'un point de vue objectif, nous avons d'abord procédé à une étude comparative extensive de plusieurs modèles de réseaux de neurones convolutifs (CNN) sous diverses configurations. Ensuite, nous avons amélioré la chaîne de traitement de l'évaluation de la qualité basée sur les CNN à différentes échelles, de l'échantillonnage et de la représentation des entrées à l'agrégation des scores de qualité. En se basant sur les résultats de ces études, et de l'analyse comparative, deux modèles de qualité basés sur les CNN sont proposés pour prédire avec précision la qualité des images à 360 degrés. Les observations et les conclusions obtenues à partir des différentes contributions de cette thèse apporteront un éclairage sur l'évaluation de la qualité des images à 360 degrés.

***Mot-clé*: Images 360 degrés, évaluation de la qualité d'image, perception visuelle, réseaux de neurones convolutifs, qualité subjective et objective.**

# Abstrakt

360-graders bilder, også kjent som rundstrålende bilder, er i sentrum av oppslukende medier. Med økningen i forventninger til sistnevnte, hovedsakelig takket være den aktiverte interaktive og oppslukende opplevelse, er det avgjørende å gi god kvaliteten på opplevelsen (QoE). Denne QoE er betydelig påvirket av kvaliteten på innholdet. Som alle typer visuelle signaler går 360-graders bilder gjennom en sekvens av prosesser, inkludert koding, overføring, dekoding og gjengivelse. Hver av disse prosessene har potensial til å introdusere forvrengninger til innholdet. For å forbedre QoE er vurdering av bildekvalitet (IQA) en av strategiene å følge. Denne oppgaven tar for seg kvalitetsevaluering av 360-graders bilder fra objektive og subjektive perspektiver. Ved å fokusere på påvirkningen av Head Mounted Displays (HMD-er) på den oppfattede kvaliteten til 360-graders bilder, er en psyko-visuell studie designet og utført ved hjelp av fire forskjellige enheter. For dette formålet opprettes et 360-graders bildedatasett og et panel av observatører er involvert. Virkningen av HMD-er på kvalitetsvurderingene identifiseres og fremheves som en viktig faktor når du utfører subjektive eksperimenter for 360-graders bilder. Fra det objektive perspektivet benchmarket vi først flere konvolusjonelle nevrale nettverk (CNN) under forskjellige konfigurasjoner. Deretter forbedres prosesseringskjeden til CNN-baserte 360-IQA i forskjellige skalaer, fra input-sampling og representasjon til aggregering av kvalitetspoeng. Basert på observasjonene av de ovenfornevnte studiene så vel som benchmark, foreslås to 360-IQA-modeller basert på CNN-er for å nøyaktig forutsi kvaliteten på 360-graders bilder. De

innhentede observasjonene og konklusjonene fra de ulike bidragene skal gi innsikt for å vurdere kvaliteten på 360-graders bilder.

***Nøkkelord*: 360-graders bilder, vurdering av bildekvalitet, visuell persepsjon, konvolusjonelle nevrale nettverk, objektiv og subjektiv kvalitet.**

# Acknowledgements

*Words cannot express my gratitude to my supervisor Mohamed-Chaker Larabi for his dedicated support. Chaker continuously provided encouragement and was always willing and enthusiastic to assist in any way he could throughout the research project. I could not have undertaken this journey without him, who generously provided knowledge and expertise. His guidance and overall insights have made my Ph.D. program an inspiring experience for me.*

*I would like to extend my sincere thanks to the reviewing committee, including Prof. Sophie TRIANTAPHILLIDOU and Prof. Aljosa SMOLIC for their precious time and thorough review of this dissertation. They raised many precious points in the manuscript, and I hope that I have managed to address them here. Furthermore, I want to show my gratitude to Dr. Laura TONI, Prof. Majdi KHOUDEIR, and Dr. Steven Yves LE MOAN for their acceptance to examine my work.*

*Lastly, I would be remiss in not mentioning my family, especially my parents, brothers, and sisters. Their belief in me has kept my spirits and motivation high during this process. I would also like to thank everyone, in whom I found emotional and moral support.*

# Contents


**Contents** . . . . . **viii**
**Figures** . . . . . **xi**
**Tables** . . . . . **xv**
**General Introduction** . . . . . **xxii**
**I Background and State-of-the-art** . . . . . **1**
I.1 Background and overview . . . . . 1
I.1.1 Immersive media . . . . . 1
I.1.2 360-degree images . . . . . 3
I.1.3 360-degree image quality assessment . . . . . 10
I.1.4 Evaluation of 360-IQA models . . . . . 13
I.2 Related work . . . . . 16
I.2.1 Subjective 360-IQA . . . . . 16
I.2.2 Objective 360-IQA . . . . . 18
I.3 Conclusion . . . . . 24
**II On the Influence of Head-Mounted Displays On Quality Ratings** . . . . . **27**
II.1 Introduction . . . . . 27
II.2 360-IQA databases . . . . . 28
II.3 Study Design . . . . . 30
II.3.1 The proposed 360-IQAD database . . . . . 30
II.3.2 Subjective assessment protocol . . . . . 31
II.4 Results and discussion . . . . . 34
II.4.1 Effects of HMD on subjective ratings . . . . . 34
II.4.2 Simulator Sickness Assessment . . . . . 38
II.5 Conclusion . . . . . 41
**III Convolutional Neural Networks for 360-IQA: A Benchmark** . . . . . **43**
III.1 Introduction . . . . . 43
III.2 The proposed benchmark: design and architecture . . . . . 44
III.2.1 Pre-trained CNN models . . . . . 44
III.2.2 Content-based splitting strategy . . . . . 47
III.2.3 Projection-based training . . . . . 49
III.2.4 Radial-based training . . . . . 50
III.2.5 Patch-based training . . . . . 52
III.2.6 Training on 2D IQA databases . . . . . 52
III.3 Results and discussion . . . . . 54

III.3.1 Experimental setup . . . 54
III.3.2 Data biases evaluation . . . 55
III.3.3 Projection-based evaluation . . . 56
III.3.4 Radial-based evaluation . . . 59
III.3.5 Patch-based evaluation . . . 60
III.3.6 Training on 2D IQA databases evaluation . . . 63
III.3.7 Computational complexity . . . 66
III.3.8 Overall performance evaluation . . . 67
III.4 Summary . . . 69
III.5 Conclusion . . . 70
**IV Pre- and Post-processing for CNN-based 360-IQA . . . 72**
IV.1 Introduction . . . 72
IV.2 Visual scanpath for patch-based 360-IQA . . . 73
IV.2.1 Visual scanpath for data-augmentation . . . 75
IV.2.2 Visual scanpath for patch qualities aggregation . . . 77
IV.3 Adaptive patch sampling . . . 81
IV.4 Input representation (normalization) . . . 83
IV.4.1 Basic Scaling . . . 83
IV.4.2 Local Normalization based methods . . . 84
IV.4.3 Difference based methods . . . 84
IV.4.4 Histogram based methods . . . 85
IV.4.5 Whitening based methods . . . 85
IV.5 Experimental setup . . . 86
IV.6 Results and Discussion . . . 86
IV.6.1 Data-augmentation . . . 86
IV.6.2 Aggregating patch qualities . . . 89
IV.6.3 Adaptive patch sampling . . . 94
IV.6.4 Patch normalization . . . 95
IV.7 Takeaways . . . 100
IV.8 Conclusion . . . 100
**V Perceptually-Weighted CNN For 360-IQA Using Visual Scan-Path And JND 103**
V.1 Introduction . . . 103
V.2 Proposed model . . . 104
V.2.1 Pre-processing . . . 104
V.2.2 Network Architecture . . . 106
V.3 Results and discussion . . . 108
V.3.1 Experimental setup: . . . 108
V.3.2 Performance comparison . . . 108
V.3.3 Ablation study . . . 109
V.4 Conclusion . . . 111
**VI Attention-aware Patch-based CNN for Blind 360-degree Image Quality Assessment . . . 114**
VI.1 Introduction . . . 114

VI.2 The proposed model . . . . . 115
VI.2.1 Inputs generation . . . . . 115
VI.2.2 Patches normalization . . . . . 115
VI.2.3 Model architecture . . . . . 116
VI.2.4 Loss function . . . . . 120
VI.2.5 Quality scores aggregation . . . . . 121
VI.3 Experiments and results . . . . . 122
VI.3.1 Experimental setup . . . . . 122
VI.3.2 Performance comparison with SOTA models . . . . . 124
VI.3.3 Cross-datasets evaluation . . . . . 128
VI.3.4 Ablation study . . . . . 129
VI.4 Conclusion . . . . . 137
**General Conclusion** . . . . . **140**
**Bibliography** . . . . . **145**

# Figures


I.1 Generic scheme of a 360-degree image generation chain. . . . . 3
I.2 Illustration of the non uniform and oversampling causing substantial variation in pixel density [25]. . . . . 8
I.3 Examples of several types of projection for 360-degree images, from which one can see that the content is deformed and discontinuous boundaries are introduced. (a) ERP, (b) EAP, (c) CMP, (d) ACP, (e) COHP1, (f) vertical SSP, (g) COHP2, (h) CISP, and (i) RSP [34]. . . . . 9
I.4 Categorization of 360-IQA methods. . . . . 10
I.5 Subjective experiment environment using a HMDs for 360-degree images. (a) compatible computer, (b) HMD tracking stations called lighthouse, (c) human observer, and (d) HMD [36]. . . . . 11
I.6 Full- vs. no-reference 360-IQA. . . . . 12
I.7 ($a$) ERP image, ($b$) radial content of the blue rectangle with a 90° field of view, and ($c$) blue rectangle extracted on the ERP image. (d) stretched content due to ERP. . . . . 20
I.8 2D plane to spherical surface mapping of 360-degree images [70]. . . . 21

II.1 (Left) spatial information (SI) versus colourfulness index (CFI) plot of CVIQ, OIQA, and MVAQD, and (right) histogram of their MOS values re-scaled to [0, 1]. . . . . 31
II.2 Pristine images in the proposed database. 1-4 rows are images taken from JVET and SUN360. Row 5 are created as synthesized images. . . 32
II.3 Spatial information (SI) versus colourfulness index (CFI) plot of the selected pristine images used for the construction of the database. . . . 32
II.4 Illustration of the adopted subjective assessment protocol. . . . . 33
II.5 HMDs used in the subjective study. . . . . 34
II.6 Histograms of the rating scores obtained by the used HMDs. . . . . 35
II.7 Confidence interval of the MOS generated by the used HMDs. . . . . 36
II.8 Probability plot of MOS against the normal distribution quantiles. . . . 37
II.9 Pairwise multiple significance plot between HMDs. "NS" stands for no significance. . . . . 38
II.10 Box plot of MOS per level of distortion, where J-*, GB-* and GN-* stand for JPEG, Gaussian Blur and Gaussian Noise with 4 different levels. . . 39

II.11 Simulator-sickness scores for the considered HMDs in terms of total scores (TS), oculomotor (O), and disorientation (D). . . . . 40

III.1 Architecture of the CNN models: Top layers replaced by a regression block composed of a global average pooling (GAP) layer, a fully connected layer (FC), a dropout layer and a final FC layer to output the predicted score $S$. $V_{\mathbf{F}} \in \mathbb{R}^{D\times1\times1}$ represents the extracted features vector. 45
III.2 Process of transfer learning from a source to a target domain. . . . . 46
III.3 Illustration of the source-to-target domain transfer. The labels of the classification database (ImageNet [121]) are classes, whereas the IQA database (TID2013 [130]) are the MOSs (continuous values). . . . . 47
III.4 Samples from the used databases: (top) CVIQ and (bottom) OIQA. . . 48
III.5 Spatial information (SI) / colorfulness information (CFI) plot of pristine images in CVIQ and OIQA databases. . . . . 49
III.6 ERP to CMP re-projection resulting in six faces: left, front, right, back, top and bottom. . . . . 50
III.7 Viewport selection for the spherical content configuration. Blue areas represent the selected viewports. In total, 24 regions surrounding the equatorial line (From 18° to 162°) are extracted from the spherical content. . . . . 51
III.8 Architecture of the multichannel CNN. $Ri$ with $i \in \{1,2,..,n\}$ stand for the extracted regions. Architecture adopted for $C_{CMP}$ $(n = 6)$ and $C_{Radial}$ $(n \in \{8,16,24\})$. . . . . 52
III.9 Architecture of the $C_{Patches}$. $P_{I.i}$ with $i \in \{1,2,3,...,n\}$, $n = 24$ for the radial sampling and $n = 6$ for the CMP sampling. $S_{P_{I.i}}$ represents the predicted quality score of patch $i$ from the 360-degree image $I$. . . . . 53
III.10 Contrast $(val_loss - loss)/(val_loss + loss)$ between training and validation losses for all models trained on 2D-IQA databases ($0 \rightarrow$ equal loss between training and validation losses). 'All' stands for combined datasets. . . . . 64
III.11 Computational complexity in terms of required prediction time per image.The average over ten samples is provided. . . . . 67

IV.1 Steps in a typical deep-learning based 360-IQA framework. . . . . 72
IV.2 Overview of the processes addressed in this chapter. . . . . 73
IV.3 CNN models for IQA. (top) multichannel vs. (bottom) patch-based CNN. 74
IV.4 360-degree images viewed using head-mounted displays. Blue area in the ERP represents the window extracted from the sphere. . . . . 75
IV.5 Illustration of standard data-augmentation [158]. . . . . 76
IV.6 Patch selection using fixations from the scanpaths and extraction on the sphere. . . . . 77
IV.7 Mapping of predicted local qualities $S_{P_i}$ (per patch) to global quality $S_I$ (per 360-degree image). . . . . 78

IV.8 Architecture of the proposed model. Features are only extracted from individual patches $P_i$ by ResNet-50. $V_F \in \mathbb{R}^{D\times1\times1}$ represents the extracted feature vector. . . . . 80
IV.9 Latitude and content's importance based patches sampling from the sphere. . . . . 81
IV.10 Overview of the adaptive sampling process on the sphere and patches' labelling. . . . . 82
IV.11 Computational time for VOs individually on CVIQ. . . . . 89
IV.12 Computational time for VOs individually on OIQA. . . . . 90
IV.13 Contrast (max-min)/(max+min) between training and validation losses for the five folds (0 → equal loss between training and validation and 1→ important gap between both losses) on CVIQ. T and V represent the reached loss values for training and validation, respectively. . . . . 91
IV.14 Contrast (max-min)/(max+min) between training and validation losses for the five folds (0 → equal loss between training and validation and 1→ important gap between both losses) on OIQA. T and V represent the reached loss values for training and validation, respectively. . . . . 92
IV.15 Performance of Minkowski mean in terms of PLCC/SRCC on OIQA (left) and CVIQ (right). . . . . 93
IV.16 Performance of Percentile pooling in terms of PLCC/SRCC on OIQA (left) and CVIQ (right). . . . . 93
IV.17 Overall Statistical Significance on CSIQ, LIVE, and TID. (Left) better vs. worse analysis and (Right) statistical significance. A white/black square: row method is statistically better/worse than the column one; gray square: statistically indistinguishable. . . . . 97
IV.18 Statistical significance on CSIQ per individual degradation. . . . . 98
IV.19 Statistical significance on LIVE per individual degradation. . . . . 99

V.1 Different scan-paths considered as virtual observers (VOs). Each scan-path is composed of eight ordered fixations. The radius of each fixation reflect the fixation duration. The color blue represent the first viewed viewport and the red one corresponds to the last viewport. . . . . 105
V.2 (Top) Examples of extracted viewports and (Bottom) their corresponding JND probability maps. . . . . 105
V.3 Architecture of the proposed model: The green rectangle depicts the overall network structure, the magenta rectangle depicts the local quality predictor structure, and the blue rectangle depicts the JND features extractor network. . . . . 106
V.4 Data splitting for training the proposed model. (Red) testing sets (blue) training sets. . . . . 108
V.5 Scatter plots of predicted quality scores versus MOS of the final model SP360IQA-F-JND (Best performance on the left and worst performance on the right among the $VOs$). . . . . 110

V.6 PLCC and SRCC of individual VOs with regard to the version of the model. . . . . 111

VI.1 Local normalization applied on patches with different distortions. (Top) extracted patches, (bottom) normalized versions. . . . . 116
VI.2 Architecture of the proposed model. $F$ and $F'$ stand for feature maps, and $S_{P'_i}$ represents the predicted quality score associated with patch $P'$. 117
VI.3 Architecture of the Conv Block with $\otimes$ element-wise multiplication and $\oplus$ element-wise addition. The $3 \times 3$ and $1 \times 1$ correspond to the kernel sizes for the convolution layers, and $2 \times 2$ represents the stride for the pooling layer (GeM). . . . . 118
VI.4 Architecture of the used regression block. "GAP" corresponds to global average pooling, and "FC" stands for a fully connected layer. $V_F$ is the generated feature vector. . . . . 120
VI.5 Illustration of saliency map to highlight the important regions. (a) a 360-degree image, (b) its saliency, and (c) their superposition. . . . . 121
VI.6 Evolution of PLCC and RMSE during training the proposed model on OIQA and CVIQ. . . . . 129
VI.7 LCN vs. RGB patches. (top) RGB and (bottom) LCN. . . . . 130
VI.8 Statistical significance analysis on OIQA among the sampling methods with respect to the size of $\alpha_0$ using the Krasula *et al.* method. "A-128/A-256" stand for the adaptive sampling with $\alpha_0 = 128/256$ respectively. "U-128/U-256" stand for the uniform sampling. For the significance plots, a white/black square: row model is statistically better/worse than the column one; gray square: statistically indistinguishable. . . . . 132
VI.9 Statistical significance analysis on OIQA for the SPA module/skip-connections using the Krasula *et al.* method. 1: No SPA/No skip-connections, 2: SPA/No skip-connections, 3: SPA/Short-connection, 4: SPA/Long-connection, 5: SPA/Short+Long connections. For the significance plot, a white/-black square: row model is statistically better/worse than the column one; gray square: statistically indistinguishable. . . . . 134
VI.10 Statistical significance analysis on OIQA among the Huber, MSE, and MAE loss functions using the Krasula *et al.* method. For the significance plots, a white/black square: row model is statistically better/-worse than the column one; gray square: statistically indistinguishable. 135
VI.11 Contrast (val_loss - loss)/(val_loss + loss) between training and validation losses for the five folds, F-1 to F-5, ($0 \rightarrow$ equal loss between training and validation. . . . . 136
VI.12 Evaluating the performances of the OR-based local qualities' aggregation by varying the value of $\lambda$. . . . . 137

# Tables


I.1 QoE influencing factors for 360-degree viewing experience. . . . . 4
I.2 Summary of 360-degree content projections . . . . . 7
I.3 Summary of subjective studies for 360-degree content. . . . . 19
I.4 Summary of traditional and deep learning-based no-reference 360-IQA models. . . . . 22

II.1 Summary of state-of-the-art 360-degree image databases. . . . . 29
II.2 Characteristics of the considered HMDs. . . . . 34
II.3 SOS parameter $a$ of all HMDs' rating scores. . . . . 35
II.4 Virtual reality sickness questionnaire [20]. . . . . 39

III.1 Number of parameters (in million) in each selected model without their top layers, and the dimension of the output vector for feature representation (fv). . . . . 46
III.2 Characteristics of the used 2D IQA databases. . . . . 54
III.3 Performance evaluation of the splitting strategies on CVIQ/OIQA databases. The best performing models are highlighted in **bold** for rows and underlined for columns. (a) and (b) stands for the SI/CFI-based schemes . . . . . 55
III.4 Performance evaluation of cross-database validation under the $C_{ERP}$. Best performing models in **bold**. . . . . 57
III.5 Performance evaluation in terms of PLCC and SRCC of pre-trained models using $C_{CMP}$. Best performances are highlighted in **bold** for each database. . . . . 58
III.6 Performance evaluation of cross database validation under the $C_{CMP}$. Best performing models in **bold**. . . . . 59
III.7 Performance evaluation of pre-trained models with the $C_{Radial}$ on CVIQ/OIQA databases in terms of PLCC/SRCC. Best performing model is highlighted in **bold** for CVIQ and underlined for OIQA. . . . . 60
III.8 Performance evaluation of pre-trained models with the $C_{Patches}$ on CVIQ/OIQA database in terms of PLCC, SRCC. Best performances are highlighted in **bold** for columns and underlined for rows. . . . . 61

III.9 Performance evaluation of pre-trained models with the $C_{Patches}$ on CVIQ/OIQA database in terms of PLCC, SRCC. Best performances are highlighted in **bold** for rows and underlined for columns. . . . . 62
III.10 Performance evaluation of $C_{2D}$ in terms of PLCC, SRCC. The best performing models are highlighted in **bold** for columns and underlined for rows on each dataset. 'All' stands for combined datasets. . . . . 63
III.11 The number of FLOPs with regard to the input shapes. . . . . 67

IV.1 Summary of basic statistic based aggregation methods. . . . . 79
IV.2 Summary of training and evaluation setup of the aforementioned studies. 86
IV.3 Performance evaluation of the model. The Best performance is highlighted in **bold**. The mean of 5 folds is provided . . . . . 87
IV.4 Computational complexity in terms of training time for data-augmentation on CVIQ and OIQA databases. The mean of 5 folds is provided. . . . . 88
IV.5 Performance evaluation of the pooling strategies in terms of PLCC, SRCC, and RMSE. The best performance is highlighted in **bold** and second-best underlined . . . . . 93
IV.6 Performance comparison with state-of-the-art mutlichannel-based models . . . . . 94
IV.7 Performance evaluation on OIQA, CVIQ, and MVAQD. The median and standard deviation (SD) over five-folds are provided. Best performance is highlighted in **bold**. . . . . 95
IV.8 Labels for the used normalization methods. . . . . 95
IV.9 Performance comparison of the selected normalization on CSIQ, LIVE, and TID2013. The median over five folds is taken. The best performance is highlighted in **bold** and second-best underlined . . . . . 96

V.1 Performance comparison with state-of-the-art quality models in terms of PLCC and SRCC. Best performance is highlighted in bold. . . . . 109
V.2 Standard deviation, maximum and minimum performance in terms of PLCC, SRCC, and RMSE of virtual observers. Best PLCC values are highlighted in bold and SRCC underlined. . . . . 110

VI.1 Performance comparison with SOTA models on the OIQA database. The Best performance is highlighted in **bold** and the second-best is underlined. $\text{Ours}_{\text{Avg}}$, $\text{Our}_{\text{OR}}$, and $\text{Our}_{\text{OR+SAL}}$ stand for the proposed model with average, OR-based, and OR + saliency based aggregation of local qualities, respectively. . . . . 125
VI.2 Performance comparison with SOTA models on the CVIQ database. The Best performance is highlighted in **bold** and the second-best is underlined. $\text{Ours}_{\text{Avg}}$, $\text{Our}_{\text{OR}}$, and $\text{Our}_{\text{OR+SAL}}$ stand for the proposed model with average, OR-based, and OR + saliency based aggregation of local qualities, respectively. . . . . 127

VI.3 Cross-database performances comparison of the proposed model with SOTA with respect to their complexity. The best performance is highlighted in **bold**. . . . . 128
VI.4 Ablating the sampling methods with regard to the patches' size, *i.e.* $\alpha_0$. 131
VI.5 Ablating the use of the SPA module and skip-connections. The Best performance is highlighted in **bold** and the second-best is underlined. 133

# List of abbreviations

**IQA** Image Quality Assessment

**QoE** Quality of Experience

**HVS** Human Visual System

**OQA** Objective Quality Assessment

**SQA** Subjective Quality Assessment

**NR** No-Reference

**FR** Full-Reference

**MOS** Mean Opinion Scores

**DMOS** Differential Mean Opinion Scores

**ERP** Equirectangular Projection

**CMP** Cube-Map Projection

**EAP** Equal-Area Projection

**OHP** Octahedron Projection

**ISP** Icosahedron Projection

**ACP** Adjusted Cube-map Projection

**SSP** Segmented Sphere Projection

**RSP** Rotated Sphere Projection

**VR** Virtual Reality

**AR** Augmented Reality

**MR** Mixed Reality

**HMD** Head Mounted Display

**SDE** Screen Door Effect

**ITU** International Telecommunication Union

**VQEG** video quality expert group

**IPD** Interpupillary Distance

**FoV** Fieled of View

**NSS** Natural Scene Statistics

**PLCC** Pearson Linear Correlation Coefficients

**SRCC** Spearman Rank-order Correlation Coefficients

**RMSE** Root Mean Square Error

**MAE** Mean Absolute Error

**MSE** Mean Squared Error

**AUC** Area Under the Curve

**ROC** Receiver Operating Characteristic

**PSNR** Peak Signal-to-Noise Ratio

**SSIM** Structural SIMilarity

**ACR** Absolute Category Rating

**SS** Single-Stimulus

**SAMVIQ** Subjective Assessment of Multimedia Video Quality

**HM** Head Movement

**HTC** High Tech Computer

**FHD** Full High Definition

**PPD** Pixels Per Degree

**HDR** High dynamic Range

**WGN** White Gaussian Noise

**GB** Gaussian Blur

**JPEG** Joint Photographic Experts Group

**LCN** Local Contrast Normalization

**TL** Transfer-Learning

**CNN** Convolutional Neural Networks

**GAP** Global Average Pooling

**GeM** Generalized Mean Pooling

**BN** Batch Normalization

**FC** Fully Connected

**ReLU** Rectified Linear Units

**SPA** Apatial Attention

**LQs** Local Qualities

**SD** Standard Deviation

**FLOPs** Floating-Point Operations

**JND** Just-Noticeable Difference

# General Introduction

## Context of the thesis

During the last decade, immersive applications have known an impressive growth due to the offered immersive and interactive visual experience. At the center of the used media, one can find 360-degree images, *a.k.a.* omnidirectional images. The latter offer captivating visual experience of real world scenes and virtual environments as in the case of virtual reality (VR). Foreseeing the massive opportunities in this field, many well-known companies such as Facebook and YouTube started offering 360-degree content and tools. As a result, more than 70 million of 360-degree images have been uploaded to Facebook [1] just in a year and YouTube brought 360-degree videos to live-streaming [2]. As demands grow and consumers' expectation rises, it is of paramount importance to guarantee the quality of experience (QoE) of users. However, the specific characteristics of 360-degree images make it particularly challenging to deal with such a content. Before being displayed to end users, 360-degree images, like any type of visual signal, go through a sequence of steps including encoding, transmission, decoding, and rendering. Each of these processes has the potential to introduce distortions to the content. Moreover, additional processes specific for 360-degree images are required, such as stitching and sphere-to-plane projection introduce geometric distortions. Besides, the used devices, *i.e.* head-mounted displays (HMDs), is prone to the screen door effect (SDE). All the highlighted issues alter the visual experience. To improve the users' QoE, 360-degree characteristics must be carefully taken into account as well as the various processing steps involved in the pipeline. Image quality assessment (IQA) is one strategy among many that might be used to accomplish this goal.

IQA is considered as one of the most difficult image processing tasks [3, 4]. Obviously, the human eye is the ultimate receiver of visual signals, requiring an IQA model that agrees with the way the human visual system (HVS) processes and perceives visual signals. A considerable effort has been made to accurately predict the perceptual quality of images according to the understanding of the HVS for traditional content, such as 2D images, where the viewing conditions are quite simple. Still, many challenges arise when dealing with immersive content in general, and 360-degree in particular. The nature of such a content and the used devices, *i.e.* VR headsets, require a deeper understanding, especially by means of psycho-visual experiments. In

addition, the HVS perception for VR environment is still in its infancy and not fully studied. Many factors influence 360-degree QoE assessment related to the (i) users: cyber-sickness, immersion and presence, and (ii) device: limited field of view (FoV) and resolution [5]. Understanding all these factors and their influence on visual perception is paramount for developing accurate and generalized IQA methods.

The goal of this thesis is to identify the factors that affect 360-IQA and investigate their impact in order to propose robust and accurate 360-IQA models. The present work comprises three complementary parts that contribute toward this aim. The first one focuses on the exploration of the influence of HMDs on subjective quality ratings of 360-degree images. The second part tackles the investigation of the use of convolutional neural networks (CNNs) for 360-IQA at various levels of the processing chain, including pre-processing, model architecture, and quality aggregations. Finally, the third part aims at building 360-IQA models based on CNN by taking advantage of conclusions drawn from the previous parts. The findings throughout the thesis shall pave the way toward more reliable, accurate, and robust 360-IQA tools.

## Contributions

In this thesis, quality assessment of 360-degree images in deeply investigated from several perspectives. The following contributions are presented in this dissertation:

- Psycho-visual evaluation of 360-IQA with a focus on the influence of HMDs on the subjective ratings.
- The use of transfer-learning from well-known and widely adopted CNNs under various configurations for 360-IQA.
- With a focus on pre- and post-processing steps for CNN-based 360-IQA models to optimize the processing chain at different levels, the following are present:
    - A specifically designed data-augmentation for training patch-based CNN models.
    - An adaptive, lightweight, and consistent patch sampling strategy by incorporating the exploration behavior of users and the content importance.
    - A comprehensive performance evaluation of input data representation prior to training CNNs.
- Two 360-IQA models :
    - A multichannel CNN with visual scanpath and just-noticeable difference (JND). The information about visual trajectories and JND are used to account for the HVS properties and make the network closer to human judgment.
    - An attention aware patch-based CNN model, incorporating spatial attention to help the model focus on spatially meaningful features. In addition, skip-connections within the spatial attention module are integrated to align the preserved features via spatial attention.

## Organization of the manuscript

This dissertation is organized in six chapters and presents the assessment of 360-degree image quality from several points of view.

- The first chapter begins with an introduction to immersive media, with a focus on 360-degree images. The major elements of 360-degree images are then discussed, as well as the most significant challenges facing the processing of this emerging type of content, specifically 360-IQA. Finally, a brief literature review related subjective and objective quality assessment of 360-degree content is presented. A series of 360-IQA models are summarized so as to provide the reader with an overview about 360-IQA models.
- The second chapter reports and discusses the influence of HMDs on quality ratings of 360-degree images. It describes the psych-visual study designed for this purpose, and the constructed database. Then, it introduces the conducted experiments and the applied statistical analysis.
- The third chapter presents an extensive investigation of the use of CNNs for 360-IQA. It describes the design of the benchmark with its different configurations. Then, an extensive discussion is provided with some recommendations.
- The fourth chapter presents several studies on the pre- and post-processing for CNN-based 360-IQA, including data-augmentation and labelling, adaptive inputs sampling, and data representation.
- The fifth chapter introduces a 360-IQA model based on multichannel CNNs, designed to incorporate several HVS properties, including visual scanpaths and JND probability maps.
- The sixth chapter is dedicated to the description of a patch-based CNN 360-IQA model with spatial attention and hierarchically features reuse. Based on the observations from chapter III and IV, an adaptive 360-IQA framework is designed, starting from input selection and data representation to the architecture of the model and aggregation of patch quality.

This dissertation finishes with general conclusions about the conducted work and provides openings and perspectives for future work.

# Chapter I

# Background and State-of-the-art

## I.1 Background and overview

### I.1.1 Immersive media

According to the Cambridge dictionary, immersion can be described as the fact of becoming completely involved in something. Therefore, Biocca *et al.* [6] identified immersion as a system property and defined it as:

> *The term "immersion" refers to the degree to which immersive media environments submerges the perceptual system of the user in computer-generated stimuli. The more the system blocks out stimuli from the physical world, the more the system is considered to be immersive.*

According to Gisbergen *et al.* [7], *immersion* is created through six dimensions, including presence, perspective, proximity, point of view, participation, and place. This viewpoint can be traced back to initial research on telepresence equates immersion to the system's ability to provide user's senses with surrogate stimuli replacing or complementing real-life signal input [8].

Thanks to the aforementioned qualities and characteristics of immersion, immersive media have drawn considerable interdisciplinary interest over the past decades, which successfully delivered various frameworks for immersive media. The latter involves multi-modal human-computer interaction, where the user immersion inside a virtual space feels as a part of the physical world. Therefore, immersive media is known to invoke a user's sense of being there with respect to the degree of immersiveness. From an experiential perspective, this combines the physical and psychological concepts of immersion [8, 9]. Immersive systems make use of technology such as displays in order to provide the users with immersive experiences, including virtual reality (VR), augmented reality (AR), high dynamic range (HDR), etc.

Following Schuemie *et al.* [10], immersive media can be characterized by :

**Immersivity:** can be measured by the combination of sensory cues with content cues essential for user emplacement and engagement.

**Interactivity:** refers to the interaction of the users with the virtual environment through a digital interface.

**Explorability:** can be described as the possibility for users to explore and move freely.

**Believability:** relates to the fidelity and validity of sensory features within the generated environments.

**Plausibility:** concerns the coherence and consistency of the content features for the user to form mental concepts.

In the Qualinet white paper on definitions of immersive media experience (IMEx) [8], immersive media is summarized as:

> *A high-fidelity simulation provided and communicated to the user through multiple sensory and semiotic modalities. Users are emplaced in a technology-driven environment with the possibility to actively partake and participate in the information and experiences dispensed by the generated world.*

Immersive media are known to stimulate physical senses to the point where we experience psychological immersion [11, 12]. It has the ability to make users involved in the simulated virtual environment, giving the impression that it is real and that they are present in it. The most known immersive media technologies that provide impassiveness and engagements are:

**Virtual Reality (VR):** occludes physical space to provide interactive and non-interactive experiences of a fully computer-simulated “virtual” world or a photographically “captured” real world [8, 13]. A virtual reality system should have three characteristics: response to user actions, real-time 3D graphics and a sense of immersion [14].

**Augmented Reality (AR):** it is defined as a real-time direct or indirect view of a physical real-world environment that has been enhanced/augmented by adding virtual computer-generated information to it [15]. Therefore, it can be considered as a digital content overlays a real-world environment. It enhances reality rather than replacing it. Milgram *et al.* [16] defined AR as the "middle ground" between virtual environment (synthetic) and telepresence (real).

**Mixed Reality (MR):** combines real and virtual content that allows real-time interaction and aims at blending real and virtual environments [17]. Milgram *et al.* described MR as a “stronger” version of AR. It is potentially bound to specific hardware or devices.

VR is one of the most popular immersive media technologies. As of today, there are an estimated of 171 million VR users worldwide. At the heart of VR applications, 360-degree visual images. The latter are also known as panoramic, omnidirectional, and

spherical images, as they cover a 360-degree range. The viewer may enjoy an immersive experience by viewing 360-degree content of real-world or computer-generated ones using head-mounted displays (HMDs).

Several challenges with regard to the processing of 360-degree images arise as it gaining more popularity. In particular, the assessment of quality of experience (QoE). The latter is strongly related to:

- The viewer presence (i.e. the degree of immersion).
- The viewer's behavior (Head rotation represented by the Roll-Pitch-Yaw movement).
- The cyber-physical system (Device).
- Simulator sickness.
- The quality of the media itself (Content).
- Physiological state of the viewers.

The elements that influence QoE for VR services, according to the international telecommunication union ITU-T G.1035 [5] recommendation, might be connected to the users and / or the used system. With a focus on 360-degree QoE, Table I.1 summarizes the most significant factors. It is of paramount importance to understand the extent to which these factors impact QoE. The improvement of the QoE relies on the fact that each component contributing to the final QoE is studied and then modeled. This helps to design and develop accurate and consistent quality assessment models.

## I.1.2 360-degree images

In this part, we describe the typical 360-degree content communication pipeline. Fig. I.1 depicts the key processes involved in the generation of this type of content, from acquisition to display and visualization.

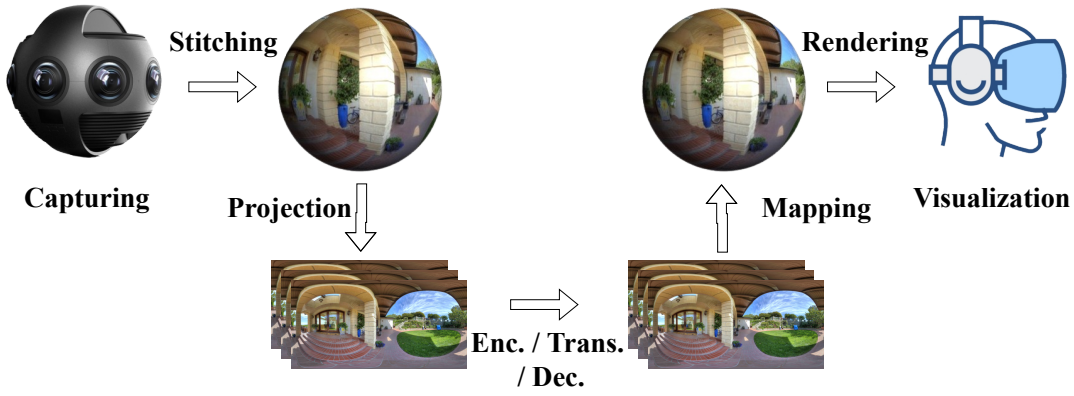



**Figure I.1:** Generic scheme of a 360-degree image generation chain.

**Table I.1:** QoE influencing factors for 360-degree viewing experience.

| Category | Factors | Brief description |
|---|---|---|
| Human | Vision | Visual abnormalities such as astigmatism and chromatic aberration may occur in the human eye. Such vision problems may negatively affect the user experience. When vision problems can be corrected by lenses, having the user wear their normal glasses may be a solution, although this may be uncomfortable. |
| | Simulator sickness | Cybersickness or visually induced motion sickness triggered by visual stimuli is undesirable phenomenon. It is caused by the sensory conflict arising between the visual and vestibular system [18]. While viewing 360-degree content in an HMD, a user may experience symptoms of simulator sickness such as fatigue, sweating, vertigo or nausea [19, 20]. Simulator sickness is an important factor that affects QoE. Different factors exist such as resolution, FoV, the orientation of users, HMD, player, etc, by which simulator sickness can be affected [21, 22]. |
| | Immersion | The tendency to experience immersion and the level of presence in VR environment, including 360-degree ones, vary individually according to each users. |
| System | Content | The quality of 360-degree content is crucial for the user's experience. As it has additional requirements when compared with traditional multimedia content, such as high resolutions and stitching, it is important that it is generated at a good quality. |
| | Device | Unlike traditional viewing devices, HMD wearing comfort may greatly impact the final QoE. Besides, the FoV and resolution of the HMD are very important features impacting the viewing experience. With a wider FoV, a user is more likely to feel present in the scene. A wide FoV can increase immersion, however it can also cause simulator sickness. This is mainly because the large visual input brought from large FoV may cause conflicts with the vestibular and proprioceptive systems [23]. Regrading the resolution of the HMD, an appropriate screen resolution, relative to the resolution of human eye, would provide the best and most comfortable experience. |

- **Capturing:** it is typically performed using multiple cameras, that are time synchronized and uniformly placed on a rig. The end result is a set of images which encompasses the entire field of view which is 360 degrees.
- **Stitching:** after the acquisition, the 2D images captured by all cameras are combined to create a 360-degree image resulting in a spherical representation. This process is called *stitching*. The latter must be performed in such a way to avoid visual distortions, particularly the scene motions and exposure differences. In a real world scenario, the intensity varies spatially, and so does the contrast across the spatial dimensions of visual scenes. For panoramic stitching, the ideal set of images will have a reasonable amount of overlap to overcome lens distortion and have enough detectable features. The set of images will have consistent exposure between frames to minimize the probability of seams occurring.
- **Projection:** in order to process, store or transmit the spherical 360-degree image, it is mapped into a planar representation. This process is called sphere-to-2D plane projection. The most widely used planar projection is the equirectangular projection (ERP) and Cube-map projection (CMP) [24]. The main issue with these projections is the variation in pixel density due to non-uniform and oversampling as illustrated in Fig. I.2. Other projections have been investigated in order to solve the shortcomings of ERP and CMP. Each projection has different characteristics. Fig. I.3 presents examples of several projection format.
- **Encoding:** after obtaining a planar representation of the 360-degree image, 2D standard codec can be used in order to decrease its spatial redundancy. Before encoding, an additional step called tilling can be applied, which divides the 360-degree image into several tiles that are independently encoded. This is useful to control which quality each tile will receive, e.g. tiles not perceptually relevant can be encoded with lower quality.
- **Transmission:** the bitstream generated by the encoding step is then stored or sent to the client over a fixed or wireless communication channel.
- **Decoding:** this step performs the inverse operation of the encoder in order to reconstruct the 360-degree image.
- **Inverse Mapping:** in order to render the 360-degree image, a spherical representation is usually used. Therefore, the transmitted planar image has to be mapped back into a sphere, by applying the corresponding inverse mapping transformation.
- **Rendering:** the images that are displayed to the user are a part of the entire viewing sphere. Depending on the user viewing direction, a selected part of the sphere is projected on a 2D plane, resulting in the so-called **viewport**. Viewers can select the viewport to focus on the content using head movement (HM) within a sphere, while eye movement (EM) determines which region can be captured at high resolution within the viewport. This viewing mechanism is quite specific to this type of application.
- **Display:** the output of the rendering step is a 2D image that can be projected on a display. The displays for 360-degree images can be categorized in two

types: the first corresponds to a navigable image on a flat 2D display (e.g., a smartphone screen), where the viewing direction can be controlled by moving the display itself or swiping using hands. The second type corresponds to a virtual reality headsets, *a.k.a* head-mounted displays (HMDs). The latter is used to track the users' head movements so as to compute the corresponding viewport, on the one hand. On the other hand, HMDs enhance the immersiveness of the users compared to using 2D flat screens.

Despite the utility of HMDs in offering an immersive experiences, still several issues regarding the display must be considered:

**Pupillary distance or interpupillary distance (IPD):** is the distance measured in millimeters between the centers of the pupils of the eyes. Everyone is different and the value changes based on whether you're looking at something close up or far away. The optical center of the lenses must be positioned correctly in relation to the center of the subject's pupils or undesired results can ensue; such as eye fatigue, headaches and even nausea.

**Screen Door Effect (SDE):** is a visual artifact caused by the display of the HMD, which consist of the space between each pixel that been magnified. That space results in the black visual grid that occur when watching through the HMD's lenses. This problem is less noticeable on higher-resolution displays, which have higher pixels per square inch (PPI). This means the pixels are packed more tightly together and there is less space between them. As the space between pixels shrinks, the screen door effect becomes less noticeable.

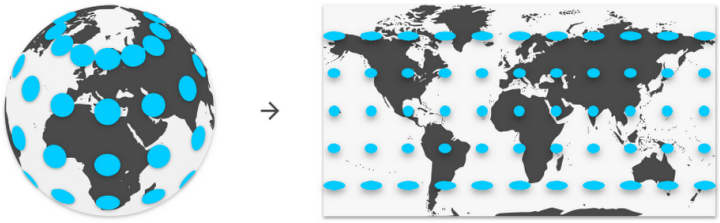
**Field of View (FoV):** is one of the most important aspects of virtual reality. The wider the field of view, the more present the user is likely to feel in the experience. Monocular FOV describes the field of view for a single eye. The binocular FoV is the combination of the two monocular FoV. When combined, a viewable area of 200°-220° can be provided. Where the two monocular fields of view overlap, there is the stereoscopic binocular field of view, about 114°, where we are able to perceive things in 3D.

Despite the fact that ERP content is widely adopted as a standard *de facto*, it does not represent the viewed content by the users through HMDs. In addition, due to the non-uniform sampling, severe geometric distortion appears at the polar regions. Several works tackled the projection problem in order to reduce the over-sampling on the one hand, and improve the consistency with the actual viewed content on the other. Table I.2, summarizes some of the most important works on 360-degree content projections. Additionally, Fig. I.3 visually illustrates several projection format.

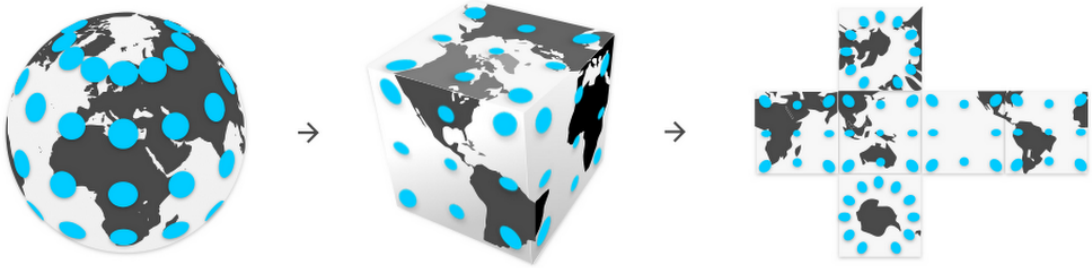
360-degree image projections can be categorized as map-based, patch-based, tile-based, and rotation-based projection. The map-based projection are straightforward and aim to simply project the spherical content to 2D plane. Patch-based projections were introduced to solve the oversampling problem of the map-based ones. The use

**Table I.2:** Summary of 360-degree content projections

| Projection type | Name | Description |
|---|---|---|
| Map-based | Equirectangular projection (ERP) Fig. I.3 (a) | The ERP's horizontal and vertical coordinates correlate to the sphere's longitude and latitude. As longitude ranges from 0 to $2\pi$ and latitude ranges from $-\pi/2$ to $\pi/2$, horizontal and vertical coordinates in ERP correspond to longitude and latitude on the sphere. |
| | Equal-area projection (EAP) [25]éFig. I.3 (b) | The EAP is proposed as a solution to the oversampling problem in ERP by adaptively decreasing the sampling rate in vertical coordinate. It can prevent oversampling and ensure that the area is equal to the original sphere while incurring a more severe shape distortion. |
| | Cube-map projection (CMP) Fig. I.3 (c) | The CMP implies a bounded cubic box surrounds the sphere; the pixels on the sphere are first projected to the cube, which is then unfolded into six surfaces and reorganized for compact representation. |
| Patch-based | Dodecahedron-projection [26] | With this projection, the sphere is projected to a rhombus dodecahedron, then divided and reassembled to generate a three by four rectangle. The rearrangement of the generated portions is made in such a way to minimize discontinuity. |
| | Octahedron-projection (OHP) [27] Fig. I.3 (e) and (g) | The OHP projects the spherical content onto the faces of the octahedron. The rearrangement of the faces may differ according to the task. For instance, JVET recommends compacting the faces using different layout resulting in COHP. |
| | Icosahedron projection (ISP) [28] Fig. I.3 (h) | The ISP is based on the OHP where the faces are rearranged into a compact format too. |
| | Adjusted Cube-map projection (ACP) [29] Fig. I.3 (d) | The ACP is an improved version of the CMP where the oversampling caused by nonuniform projection at different angles. A nonlinear modification is proposed with a sample rate adjustment mechanism based on location. |
| Tile-based | Li *et al.* [30] | The poles in the sphere are projected to circles instead of tiles, to eliminate distortions |
| | Yu *et al.* [31] | The proposed projection divides the ERP into several tiles. The idea is to prevent oversampling by decreasing the sampling rate in the horizontal direction. |
| | Segmented sphere projection (SSP) [32] Fig. I.3 (f) | The SSP is an extension of the projection proposed by Li *et al.* [30]. The number of tiles is decreased to three so as to generate less discontinuous boundaries. |
| Rotation-based | Rotated Sphere Projection (RSP) [33] Fig. I.3 (i) | This projection unfolds the sphere at two separate rotation angles and stitches it together like a baseball surface [34]. |



(a) Equirectangular projection (ERP)



(b) Cube-map projection (CMP)

**Figure I.2:** Illustration of the non uniform and oversampling causing substantial variation in pixel density [25].

of polyhedron with numerous faces to approach the ideal sampling rate is the most widely adopted strategy. It appears that as the number of faces increases, the oversampling rate decreases. However, this is accomplished at the cost of introducing content's discontinuity. The same goes for the tile-based projections. The constant spacing of latitudes and longitudes leads to a constant vertical sampling density. Since each latitude is stretched horizontally to fit the desired rectangle, it leads to varying horizontal sampling density [31]. The tile-based projection aims at changing the sampling density while retaining the rectangular shape. Here, ERP representation is sliced into multiple tiles. Regarding the rotation-based projection, rotating the original sphere surface before projecting it into a 2D plane enhances the coding efficiency as demonstrated by Zakharchenko *et al.* [35].

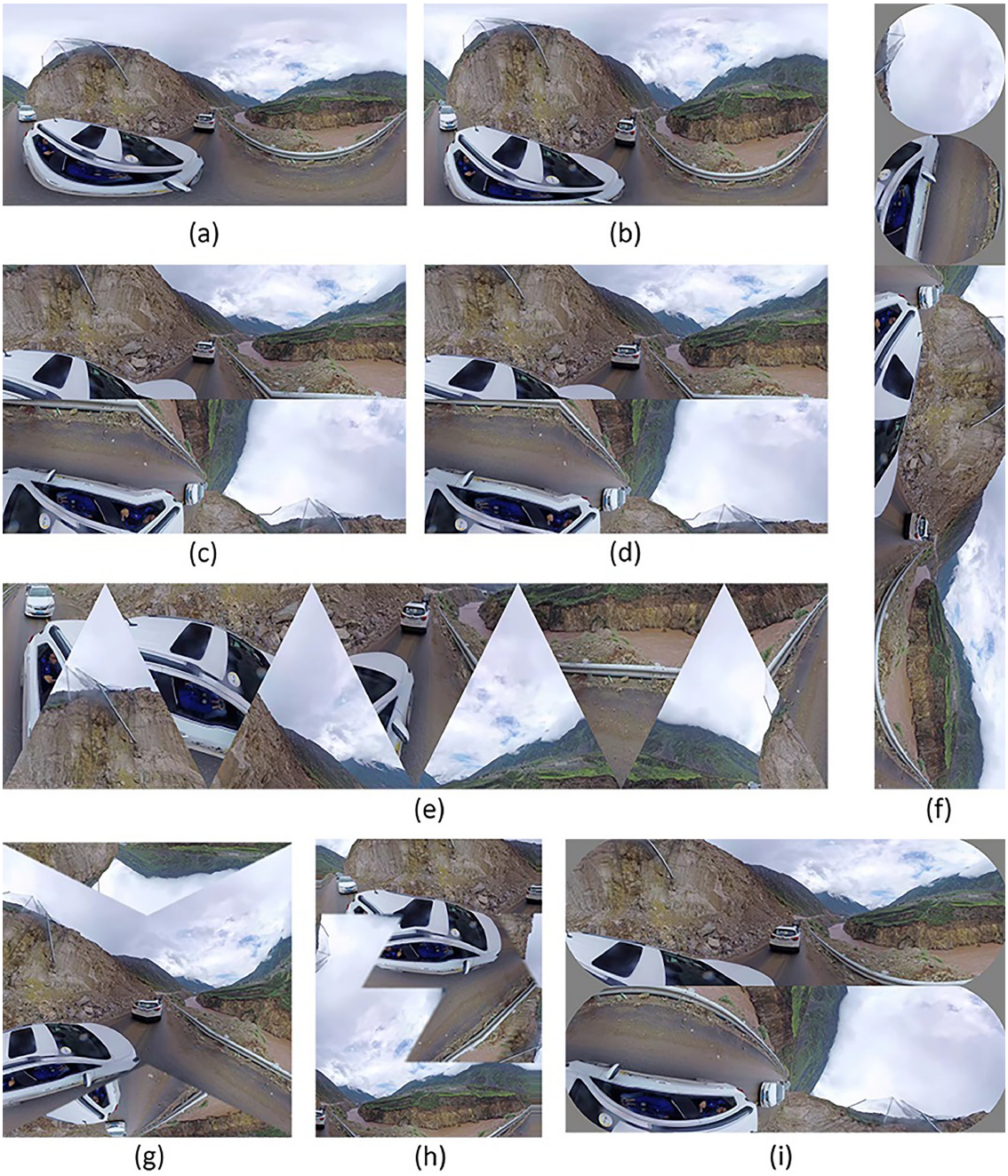



**Figure I.3:** Examples of several types of projection for 360-degree images, from which one can see that the content is deformed and discontinuous boundaries are introduced. (a) ERP, (b) EAP, (c) CMP, (d) ACP, (e) COHP1, (f) vertical SSP, (g) COHP2, (h) CISP, and (i) RSP [34].

### I.1.3 360-degree image quality assessment

Image quality assessment (IQA) is crucial for most image processing applications. IQA can be used to monitor image quality for any imaging-based system. For instance, visual content providers use IQA to examine the quality of the digital content transmitted to the customers and garantee a better QoE. IQA can be employed to evaluate and benchmark image compression systems and algorithms. In particular, IQA can help to evaluate which compression algorithm provides the best perceptual quality while requiring less bitrate.

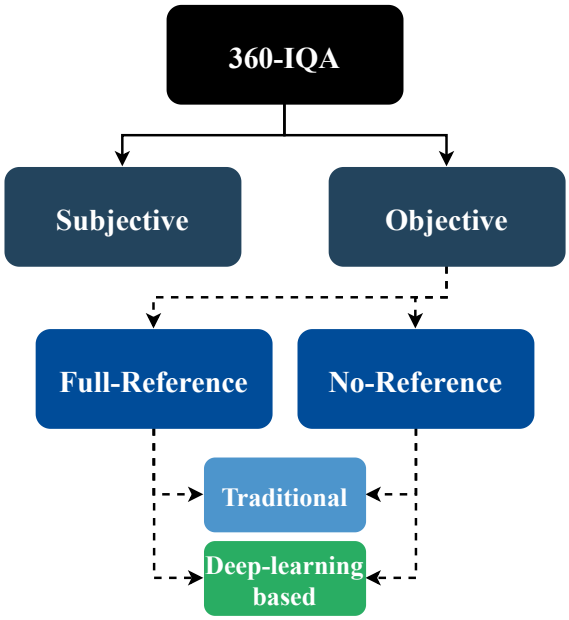



**Figure I.4:** Categorization of 360-IQA methods.

Despite the fact that 360-degree images gained an impressive popularity in a short time, 360-IQA still in it infancy and not fully explored. Compared to 2D-IQA where a significant progress has been achieved in terms of performance, accuracy, and robustness, 360-IQA is yet to be broadly uncovered. 360-IQA comprises two classes as shown in Fig. I.4. First, subjective quality assessment (SQA) involving human observation. With SQA, the exploration behavior and visual perception of individuals in immersive environments and for a set of images can be studied by following recommendations and specific protocols. The collection of human opinion serves to build IQA databases. Second, the objective quality assessment (OQA) focuses on developing computational approaches that mimic how 360-degree images are viewed and perceived. According to 360-IQA state-of-the-art, OQA for 360-degree images is divided into two categories, full-reference (FR) and no-reference (NR). Within each class traditional models based on natural scene statistics (NSS) or pixel-to-pixel differences are proposed in addition to deep-learning based ones.

SQA remains the most reliable way to evaluate image quality while being tedious

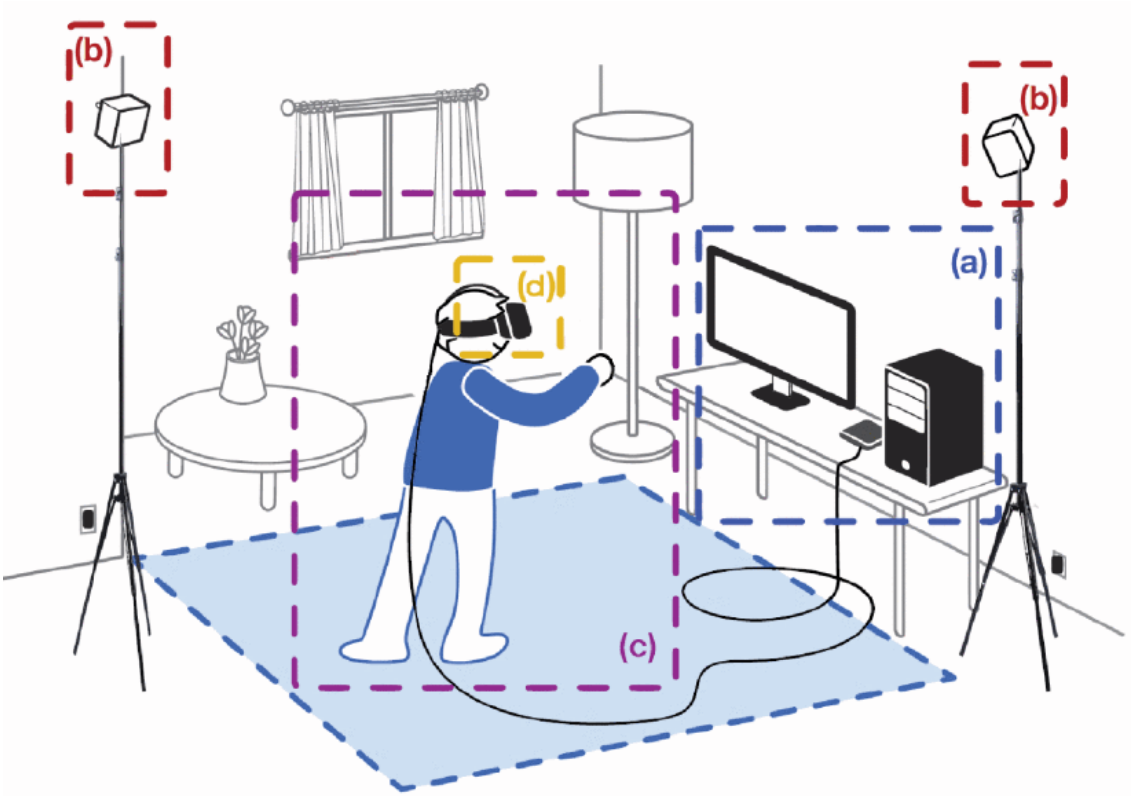



**Figure I.5:** Subjective experiment environment using a HMDs for 360-degree images. (a) compatible computer, (b) HMD tracking stations called lighthouse, (c) human observer, and (d) HMD [36].

and money consuming. Therefore, objective quality evaluation ensures a trade-off by providing a computational approach for predicting image quality. However, the reliability, predictability, and robustness of OQA models are dependent and conditioned on the reliability of SQA. As the latter plays the ground truth on which OQA is evaluated and ascertained, it must be carefully designed.

SQA are based on psycho-physics research, a field that studies the relationship between physical stimulus and human perception. An SQA method consists on collecting subjective opinions from human observers, where the latter judge and rate image quality according to their perception and preference. The collected data is mapped to a subjective score used as ground truth. SQA methods can be classified into:

**Single-stimulus:** observers only see the distorted images and are unaware of the pristine ones.

**Double-stimulus:** both pristine and distorted images are available to observers.

SQA for 360-degree images uses single-stimulus methods, as the HMD has a single screen. A typical experimental setup is shown in Fig. I.5. Usually, the observer sits on

a swivel chair to reduce side effects, such as cyber-sickness, on the one hand. On the other hand, the observer has the ability to turn around so as to explore the full scene. However, it is still possible for the observers to stand during the experiment. The HMD is connected to a compatible computer in order to render the received signal. According to the type of HMD, tracking stations may be required. These stations power the presence and immersion by helping the HMD and controllers track their exact locations. Basically, the stations facilitate the VR experience. They are the proxy in between the computer, the peripherals, and the users. This is accomplished by constantly flooding the room with a non-visible light [37].

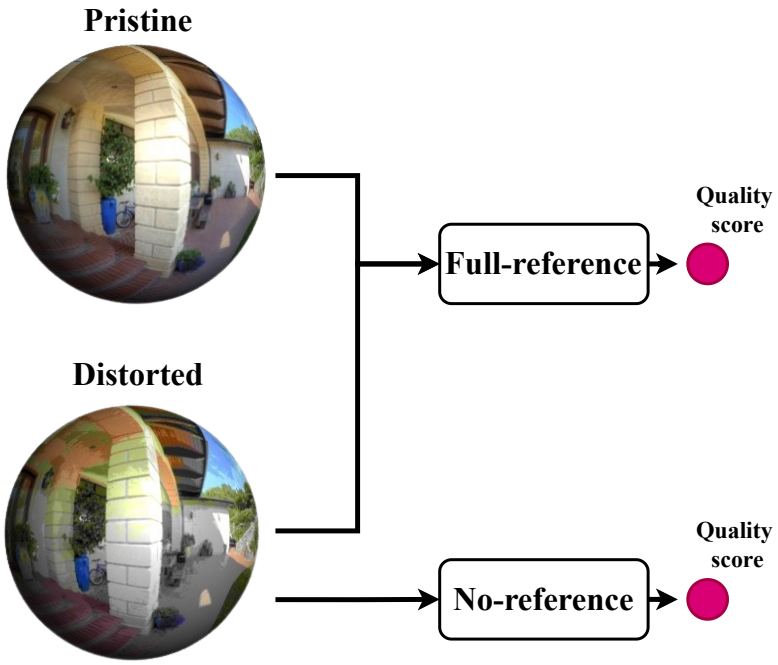



**Figure I.6:** Full- vs. no-reference 360-IQA.

Regarding OQA categories, the difference among them lies in the availability of pristine images. In case of 360-IQA, only FR and NR models have been investigated. The FR models make use of both pristine images and their distorted versions as illustrated in Fig. I.6 (a). By having access to the pristine images, FR models can measure the (i) difference in the visual signal refereed to as *fidelity*, and (ii) the amount of information loss. Image quality and fidelity are often used interchangeably. However, they do not refer to the same aspect. Daly *et al.* [38] describe image fidelity as a subset of overall image quality that specifically addresses the visual equivalence of two images. It is also referred to as the ability to discriminate between two images in terms of how closely the image represents the real source distribution [10, 39]. It is known that image fidelity and quality measures are not always positively correlated [39]. For instance, an enhanced image considered as a distorted version may be preferred over the pristine one.

Image quality is described as the weighted combination of visually significant attributes of an image [40]. Therefore, the assessment of image quality focuses on perceptual assessments to determine if an image is pleasant for human viewers. In NR cases, the model have only access to the distorted version, and therefore can not compute the fidelity as there are no sources to compare with. The fact of missing the pristine images makes determining the level of quality challenging. In particular, images with multiple distortions and captured in the wild. A quality of an image could be affected by several attributes [4, 41–43], such as:

**Blur:** can affect the amount of detail in an image. It is mainly caused by imaging systems used for capturing, in particulars the lens and sensors.

**Noise:** is a random variation in image density that appears as pixel level changes.

**Dynamic range:** is the range of light levels that a camera can record. It is connected to noise, where excessive noise suggests a poor dynamic range.

**Contrast:** is the log-log slope of the tone reproduction curve. High contrast frequently results in a loss of tones and, as a result, loss of details.

**Artifacts:** are visual representational abnormalities affecting the details within images. They may be induced by the capturing system, stitching, compression, and transmission losses.

### I.1.4 Evaluation of 360-IQA models

According to the the ITU and the video quality expert group (VQEG), the performances and reliability of objective models can be evaluated by their correlation with subjective opinion scores, *i.e.* MOS or DMOS. This is achieved with respect to three aspects: prediction accuracy, prediction monotonicity and prediction errors. Therefore,VQEG recommends the use of the Pearson linear correlation coefficients (PLCC), Spearman rank-order correlation coefficients (SRCC), the root mean square error (RMSE), and the mean absolute error (MAE) as evaluation metrics. PLCC evaluates the accuracy of the correlation, whereas the SRCC evaluates its monotonicity. A value close to 1 relates to a better accuracy and monotonicity for PLCC and SRCC, respectively. As for the RMSE and MAE, they evaluate the prediction errors where lower values refers to a lower average error in the prediction.

Prior to calculating the PLCC, RMSE, RMSE, and MAE for different objective quality assessment models, a five-parameter logistic nonlinear fitting function is recommended by the ITU-R recommendations [44]. It is used to map predicted quality scores to a common scale. However, it is to be recalled that the use of the logistic function cannot be done if the native correlation is below ±0.7. Otherwise the correlation value cannot be considered as reliable because regression quality is very low. The fitting function can be computed as follows:

$$f(\hat{y}) = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + \exp^{\beta_2(\hat{y} - \beta_3)}} \right) + \beta_4 \hat{y} + \beta_5 \tag{I.1}$$

where $\hat{y}$ denotes the objective score and $f(\hat{y})$ represents the corresponding mapped score. $\beta_i (i = 1, 2, 3, 4, 5)$ correspond to the logistic function parameters to be fitted.

The aforementioned performance metrics can be calculated as follows:

**PLCC:**

$$PLCC = \frac{\sum_{i=1}^{N} (y_i - \mu_{y_i})(f(\hat{y}_i) - \mu_{f(\hat{y}_i)})}{\sqrt{\sum_{i=1}^{N} (y_i - \mu_{y_i}) * \sum_{i=1}^{N} (f(\hat{y}_i) - \mu_{f(\hat{y}_i)})}}, \tag{I.2}$$

where $y_i$ and $f(\hat{y}_i)$ denote the $i-th$ subjective and mapped objective quality values. $\mu_{y_i}$ and $\mu_{f(\hat{y}_i)}$ represent the corresponding mean values of $y_i$ and $f(\hat{y}_i)$, respectively.

**SRCC:**

$$SRCC = 1 - \frac{6 \sum_{i=1}^{N} {d_i}^2}{N(N^2 - 1)}, \tag{I.3}$$

where $N$ is the number of image samples, $d_i$ indicates the rank difference between the subjective and objective evaluations for the $i-th$ image.

**RMSE:**

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - f(\hat{y}_i))^2}{N}}. \tag{I.4}$$

**MAE:**

$$MAE = \frac{\sum_{i=1}^{N} |y_i - f(\hat{y}_i)|^2}{N}. \tag{I.5}$$

In addition to PLCC, SRCC, RMSE, and MAE, some evaluation metrics have been proposed to extensively evaluate the performances of an objective IQA model. For instance, the metric proposed by Krasula *et al.* [45] provides more insight on the behavior and performances of an IQA model. It examines the statistical significance among several models by comparing the area under the curve (AUC) values from receiver operating characteristic (ROC) analysis [46, 47], mostly known for binary classification tasks. Hence, the ability of an IQA model to classify images according to their perceived quality can be highlighted. Basically, it compares the capacity to distinguish

different/similar and better/worse among image pairs and provide a percentage of correct classification denoted as $C_0$.

The statistical significance of the Krasula *et al.* method is calculated according to method proposed by Hanley *et al.* [48]. Hence, given two objective models $a$ and $b$, a critical ratio $c_{ab}$ between the AUC for $a$ and $b$ is measured. This is accomplished as follows:

$$c_{ab} = \frac{AUC_a - AUC_b}{\sqrt{SE_a^2 + SE_b^2 - 2rSE_aSE_a}}, \tag{I.6}$$

where $SE_a$ and $SE_b$ are the standard errors for $AUC_a$ and $AUC_b$ computed according to Hanley *et al.* [49], respectively, and $r$ is an estimated correlation between the two areas. The standard error $SE$ is obtained as:

$$SE = \sqrt{\frac{AUC(1-AUC) + (n_{g1}-1)(Q_1 - AUC^2) + (n_{g2}-1)(Q_2 - AUC^2)}{n_{g1}n_{g2}}} \tag{I.7}$$

where $n_{g1}$ and $n_{g2}$ are the numbers of elements in each group in the ROC analysis, and $Q_1$ and $Q_2$ are:

$$Q_1 = \frac{AUC}{(2-AUC)} \tag{I.8}$$

$$Q_2 = \frac{2AUC^2}{(1+AUC)}. \tag{I.9}$$

In details, the Krasula *et al.* method provides:

- For the different vs. similar analysis:
  - AUC values showing how well can a model distinguish between significantly different and similar stimuli.
  - Threshold for the model's scores difference providing 95% probability that the images are significantly different.
- From the better vs. worse analysis:
  - Percentage of correct recognition of the qualitatively better stimulus from the pairs.
  - AUC values showing how well can the model recognize qualitatively better stimulus from the pair.

Such a metric necessitates additional data to the MOS, such as the standard deviations. Unfortunately, almost all available databases only provide the MOS. It becomes paramount to share more subjective and statistical data within IQA databases. This will help to understand objective models behavior towards specific content, conditions, and even degradation.

## I.2 Related work

In this section, we review previous studies on subjective and objective quality assessment for 360-degree content.

### I.2.1 Subjective 360-IQA

SQA is the most reliable method for assessing QoE, particularly for applications intended to human users. However, it necessitates a significant number of human observers as well as a substantial amount of money and resources. In case of immersive applications in general and 360-degree images in particular, a specific setup is required with advanced technologies such as HMDs. SQA is typically used to create databases for training and testing OQA methods. The subjective ratings could be used as a foundation for developing new algorithms and paradigms. For instance, the user's behavior can be modeled by means of visual attention modeling. The latter can be exploited in various applications, including image generation, object classification, target detection and tracking, etc.

Since SQA must be carefully performed in order to produce and collect accurate results, the ITU and VQEG developed guidelines on how to perform such experiments. For instance the ITU-Rec BT.500 [44] among other recommendations lists the protocol to setup the experiment environment, test conditions, nature of content, category of participants, data collection and analysis, etc. Following such guidelines ascertain the reliability of the experiments, collected data, and most importantly the drawn conclusions. Unfortunately, during the conduction of this thesis, there were no guidelines nor recommendations on how to conduct reliable subjective experiments for immersive applications in general, and 360-degree images in particular.

Because it is still in its early stages, existing SQA of 360-degree content studies employ 2D-related methods and standards [44]. Especially, the test method, collection of opinion scores, and the data processing. Regarding the test methods, several factors are considered, including (i) session setting, (ii) material arrangement, and (iii) rating scales. As 360-degree images are an emerging type of media on the one hand, and uses specific devices, *i.e.* HMDs, on the other, a training session is required before the subjective experiment. This helps the participants to familiarize with the 360-degree content, virtual environment, and the HMDs. Many existing works applied directly test methods designed for 2D-IQA such as the absolute category rating (ACR) and ACR with hidden reference (ACR-HR) methods, recommended in the ITU-T P.910 [50]. This is motivated by the fact that these methods are single-stimulus (SS) [44] based methods, which agree with 360-degree images viewing through HMDs. The latter can only display one stimulus at a time. Therefore, the participants rate each stimulus independently and without comparison. There were a few attempts to fine-tune existing methods for 360-degree visual content. In particular, Singla *et al.* [51] proposed a modified version of the ACR named M-ACR, where each stimulus is viewed twice before the rating. Bo *et al.* [52] proposed the subjective assessment of multimedia

panoramic video quality (SAMPVIQ), a modified version of the subjective assessment of multimedia video quality (SAMVIQ) [53]. Technically, in SAMPVIQ, the test stimulus is separated into different groups, where the reference stimulus is identified and played to the observer first, followed by the distorted version. The observers are free to view the stimuli in any order they want. When all the stimuli in a group have been rated by the observers, the experiment goes on to the next group.

With regard to the rating scales and the data processing, the majority of works [21, 54–58] used either discrete or continuous rating scales. As for processing the opinion scores, the recommendation listed in [44] are followed. Hence, the mean opinion scores (MOS) or the differential MOS (DMOS) is generated. Yet, the work done by Xu *et al.* [59] resulted in the overall DMOS (O-DMOS) where the head-movement (HM) of observers are taken into account. Here, the HM accounts for the importance of different regions in the 360-degree image.

The use of HMDs is unavoidable for quality ratings of immersive applications. Most reported studies overlooked the distortion introduced by the HMDs, in particular SDE. Besides, Therefore, ascertaining the reliability of the subjective experiments when using various commercial HMDs is of paramount importance. In the literature, one can find some studies targeting several factors that may influence the perceived quality, including the resolution of the content, the HMDs, test methods, viewing modes, etc. In these studies, different devices from different manufacturers are used. In particular, Singla *et al* [21] considered studying the impact of the HTC Vive and the Oculus Rift using different content in terms of resolution (4k, FHD). They also recorded head movements of the viewers to determine their behaviors. Here the focus was rather on the resolution and not the HMD itself. Similarly, the effect of different resolutions on perceived quality is studied by Hofmeyer *et al.* [55] and Zou *et al.* [58], where the HTC Vive and HTC Vive Pro quality ratings are compared. In addition, Zou *et al.* included the impact of pixel density on the perceived quality. Hence, a high-resolution monitor is used by adjusting the distance between the viewer and the screen to obtain different densities. Here, they demonstrated that with a higher density, quality improves until a saturation at values greater than 60 pixels per degree (PPD), which corresponds to the retina resolvable resolution [60]. Rossi *et al.* [61] compared user navigation trajectories with several platforms, including the Oculus Rift, a laptop, and a tablet. The effect of device and content on navigation is highlighted by evaluating user behavior when watching various types of material, ranging from movie to action and documentary content. When watching content with a primary focus of attention, the HMD was found to lead to similar navigation among users compared to the other devices. A closely related work to Rossi *et al.*, is the comparative study by Zhang *et al.* [57]. Here, the authors analyzed the influence of viewing methods on subjective evaluation, including free-viewing, fixed trajectory viewing, and content-dependent viewing modes. The subjective evaluation was found to be significantly affected by the viewing mode. Perez *et al.* [56] focused on assessing the cyber-sickness caused by high motion 360-degree content. To do so, they proposed to isolate the camera motion from other factors defining anchors to control the gaze of the viewers.

Table I.3 summarizes several studies conducted on SQA for 360-degree visual content. All of them targeted 360-degree videos, as it is more challenging. However, the conclusions drawn can be applied to 360-degree images as well. Most of these studied contributed to the ITU-T recommendation P.919 [62, 63]. The latter, performed a cross-lab test within the immersive media Group (IMG) of the VQEG in order to validate and recommend test methodologies to evaluate the audiovisual quality of 360-degree visual content.

## I.2.2 Objective 360-IQA

In comparison with traditional IQA methods for 360-degree images, deep-learning based ones gained a considerable attention. Such choices are mainly motivated by the impressive performances brought by deep-learning techniques. However, as deep-learning based models require large scale databases in order to achieve good results, 360-IQA databases are not of sufficient size to allow such an achievement. Existing models adopted transfer learning (TL) and fine-tuning techniques to cop with these limitations. In the following, we review existing 360-IQA models with respect to their category illustrated by Fig. I.4.

### I.2.2.1 Full-reference 360-IQA

As 360-degree content gained more popularity, a few 360-IQA models have been proposed by extending traditional 2D models such as PSNR, SSIM or MSE. In particular, Yu *et al.* [70] introduced the Spherical PSNR (S-PSNR) which computes the PSNR on a spherical surface instead of the 2D representation. The weighted spherical PSNR (WS-PSNR) [71] uses the scaling factor $w(i,j)$ of the projection from a 2D plane to the sphere as a weighting factor for PSNR estimation. For each pixel $(i,j)$ in $I$ and $\hat{I}$ the $WS-PSNR$ is obtained as follows:

$$WS-PSNR = 10\log_{10}(\frac{MAX^2}{WMSE}), \tag{I.10}$$

where,

$$WMSE = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(I(i,j)-\hat{I}(i,j))^2.w(i,j)}{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}w(i,j)}, \tag{I.11}$$

and,

$$w(i,j) = \cos(\frac{\pi}{N}(j+\frac{1}{2}-\frac{N}{2})). \tag{I.12}$$

Similarly, Chen *et al.* extended the structural similarity index (SSIM) [72] to spherical SSIM [73] by exploiting the relationship of structural similarity between the 2D plane and the spherical domain. Hence, computing the luminance, contrast and structural similarities at each pixel in the spherical domain. The S-SSIM uses the same

**Table I.3:** Summary of subjective studies for 360-degree content.

| Study | Year | Description | Res. | HMD | Observers | Test method |
|---|---|---|---|---|---|---|
| Tran *et al.* [64] | 2017 | Assess the QoE with regard to the perceived quality, presence, screen, and cybersickness. | HD / FHD / 2K / 4K | Samsung Gear VR / Google Cardboard | 36 | - |
| Singla *et al.* [21] | 2017 | Study of the influence of content's resolution, observers' behavior and gender in the final quality rating and the simulator sickness. | FHD / 4K | HTC Vive / Ocolus Rift | 28 (M:13, F:15) | ACR (5 scales) |
| Bo *et al.* [52] | 2017 | Investigate the performances of different SQA methods and proposed a modified version of SAMVIQ [65] for 360-degree content. | 4K | HTC VIVE | 23 (M:13, F:10) | SSCQS / SAMVIQ / SAMPVIQ |
| Singla *et al.* [66] | 2017 | Study of quality evaluation under different coding bitrates. | FHD / 4K | Ocolus Rift | 30 (M:15, F:15) | M-ACR |
| Perez *et al.* [56] | 2018 | Assess the cybersickness caused by high motion 360-degree content by isolating the camera motion effect from other factors. | - | Google Daydream | 15 (M:5, F:10) | - |
| Xu *et al.* [67] | 2018 | Investigate the viewing directions of observers and an adaptive method for SQA. | 3K / 8K | HTC Vive | 40 (M:29, F:11) | SSCQS |
| Fremereya *et al.* [68] | 2019 | Study of the impact of different resolutions with two HMDs on the perceived quality. | 4K / 6K / 8K | HTC Vive / Vive Pro | 27 (M:13, F:14) / 28 (M:16, F:12) / 27 (M:14, F:13) | ACR (5 scales) |
| Hofmeyer *et al.* [55] | 2019 | Explore the impact of frame-rate, motion interpolation techniques, and VR players on the quality rating. | 3K | HTC Vive Pro | 12 (M:11, F:1) | ACR |
| Singla *et al.* [51] | 2019 | Comparative study on ACR/M-ACR/DSIS under different resolutions and bitrates. | 4K / 6K / 8K | HTC Vive Pro | 29 (M:18, F:11) / 30 (M:16, F:14) / 28 (M:19, F:9) | ACR / M-ACR / DSIS |
| Zhang *et al.* [57] | 2019 | Study the influence of the viewing method including free-viewing, fixed trajectory, and content-dependent viewing modes. | 3k | HTC Vive | 30 (M:17, F:13) | ACR |
| Zou *et al.* [58] | 2020 | Investigate the impact of screen resolution. | 4K / 8K | HTC Vive / ViVe Pro / Pico G2 | 30 (M17: F:13) | ACR (5 scales) |
| Rossi *et al.* [61] | 2020 | Investigate the users behavior in VR by analyzing navigation trajectories under different viewing platforms (HMD, laptop, and tablet). | 2k | Oculus Rift | 94 (M:65, F:29) | - |
| Chu *et al.* [69] | 2021 | Study the rating duration with respect to the ACR and M-ACR methods. | 2K / 3K / 4K / 8K | HTC Vive pro | 26 | ACR / M-ACR (5 scales) |

weights as WS-PSNR [71] computed as. Zakharchenko *et al.* proposed to compute PSNR on the craster parabolic projection (CPP), CPP-PSNR [74], after re-mapping pixels of both the pristine and distorted images from the spherical domain to CPP. Differently, the works in [75–77] allocate weights in the computation of PSNR using saliency. Following the path of extending 2D models, Croci *et al.* [78] proposed to apply traditional 2D-IQA metric such as PSNR, SSIM, MS-SSIM[79], and VMAF [80] on Voronoi patches. The spherical Voronoi diagram is utilized to sample patches on the spherical content. The selected models are then applied to sampling patches, and the overall quality score is calculated using a simple average of patch scores. The same framework is extended in [81] by incorporating visual attention prior to sample Voronoi patches.

As it can be seen from the abovementioned models, FR 360-IQA models focused on extending 2D models. The majority of these models are based on signal fidelity measurement, which does not consider characteristics of omnidirectional / VR perception and by that, cannot sufficiently reflect the perceived visual quality. Another deficiency of these models is that the fidelity difference is performed locally, pixel-by-pixel differences, and do not consider global artifacts. Additionally, the computation are performed on the projected content, which present various geometric distortions.

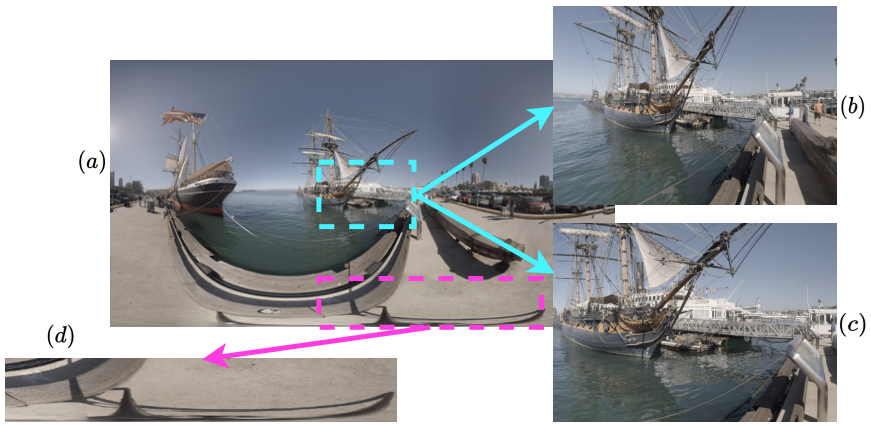



**Figure I.7:** ($a$) ERP image, ($b$) radial content of the blue rectangle with a 90° field of view, and ($c$) blue rectangle extracted on the ERP image. (d) stretched content due to ERP.

Following the strategy of using 2D models for 360-IQA, the framework proposed by Sui *et al.* [82] maps 360-degree images to moving camera videos by extracting sequences of viewports along visual scan-paths and then apply 2D-IQA models. The scan-paths are used to represent possible visual trajectories and then used to generate the set of viewports. The generated videos from the sampled viewports only contain global motion, representing a capturing process by a moving camera, where the patterns of movement are determined based on the users viewing behavior. Several 2D-IQA models including the PSNR, SSIM [73], VIF [83], NLPD [84], and DISTS [85] are then applied on the resulted videos with a temporal pooling across the set of

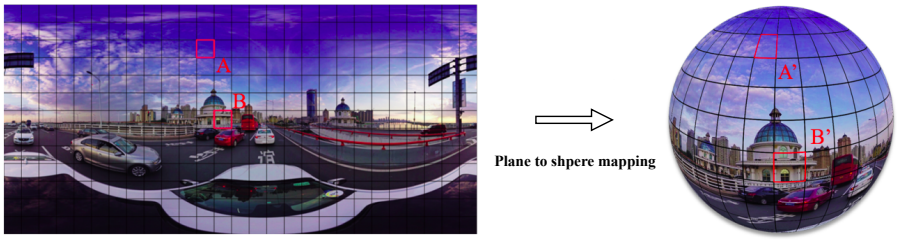



**Figure I.8:** 2D plane to spherical surface mapping of 360-degree images [70].

viewports.

To the best of our knowledge, only a single work on FR 360-IQA models based on deep learning has been proposed, the DeepVR-IQA model proposed by Kim *et al.* [86]. Here, the authors extracted $256px \times 256px$ patches from the Equirectangular (ERP) image, and then fed them to parallel pre-trained ResNet-50 [87]. The spherical positions of extracted patches are used along with their visual content. Thirty-two channels in total are used, each one is composed of a ResNet-50 and a self-defined multilayer perception (MLP). The resulted model is therefore highly complex. Furthermore, predicting quality based on ERP content is inadequate since it is geometrically distorted and does not represent the real perceived content as illustrated in Fig. I.7. First, the polar regions are stretched due to sphere-to-plane projection. Second, the viewport content extracted on the ERP differs from that on the sphere (radial), even if both are on the same location and near the equator (the least geometrically impacted region of the image). This is also supported by the changes of some regions according to their location during the sampling from 2D plane to spherical surface as illustrated in Fig. I.8.

#### I.2.2.2 No-reference 360-IQA

NR methods are the most challenging ones compared to FR models as they do not have access to pristine images. Despite this, they are widely adopted since they reflect real-world scenarios where pristine images are unlikely to be available, such as streaming applications. With a focus on NR 360-IQA, we present a literature review of current models in this section. These are mainly based on (i) natural scene statistics (NSS) or (ii) deep learning. Deep-learning based models adopt in general the multichannel paradigm. As IQA databases usually come with a ground truth label per image, a multichannel CNN is optimized to the overall quality. Whereas a traditional patch-based CNN will require a labeling step to label each patch individually.

MC360IQA proposed by Sun *et al.* [90] introduced the multichannel paradigm by using six pre-trained ResNet-34 [87] with a hyper architecture, where earliest activations are hierarchically added to the last one. The six hyper ResNet-34 are used in parallel and their outputs are concatenated and fused to a single quality score. By doing so, the lack of ground truth labels, *i.e.* MOS, per individual viewport is some-

**Table I.4:** Summary of traditional and deep learning-based no-reference 360-IQA models.

| Model | Year | Brief description | Input's type | Model's type | |
|---|---|---|---|---|---|
| | | | | Traditional | Deep-learning |
| Omni-IQA [88] | 2019 | Simple CNN and equator-based sampling of patches in addition to a weighted average of the patches quality. | Patches (ERP) | | ✓ |
| ASY-PIQA [89] | 2019 | Filtered high-frequency and low-frequency features of projected 360-degree images. | ERP | ✓ | |
| MC360IQA [90] | 2019 | Multichannel CNN using ResNet-34 with cubmap faces, each goes as input to a CNN channel. | cubmap faces | | ✓ |
| SSP-BOIQA [32] | 2020 | Features extracted from heatmap-based weighted bipolar and equatorial regions and regressed using a random forest. | Equatorial and bipolar regions | ✓ | |
| VGCN [91] | 2020 | Graph-CNN on encoded visual features by a multichannel ResNet-18 to model the dependency among selected viewports in addition to a global quality estimator on ERP images. | Viewports (ERP) and ERP | | ✓ |
| Zhou et al. [92] | 2021 | Multichannel Inception-V3 under shared weights with quality regression and distortion classification. | Cubmap faces | | ✓ |
| Ding et al. [93] | 2021 | Exploiting statistical characteristics with means of adjacent pixels correlation. | ERP | ✓ | |
| AHGCN [94] | 2021 | Modeling location and content based hyperedges using a hypergraph CNN on visual features encoded by a multichannel ResNet-18. | Viewports | | ✓ |
| SG360BIQA [95] | 2021 | Encoded visual and saliency features from cubmap faces using two CNNs. | Cubmap faces | | ✓ |
| SAP-net [96] | 2021 | Spatial-attention augmented ResNet-34 on wavelet-based enhancement to estimate the reference images. | Patches (ERP) | | ✓ |
| Liu et al. [97] | 2021 | Utilizing NSS and structural features. | ERP | ✓ | |
| MFILGN [98] | 2021 | Exploiting local-global naturalness and multi-frequency analysis. | Viewports (ERP) and ERP | ✓ | |

what solved as the six channels are trained to deliver a single predicted score per 360-degree image. However, it is achieved with an increasing complexity due to the multi CNNs. Foreseeing it performances for 360-degree quality assessment, the multichannel paradigm has been adopted by various works. For instance, Zhou *et al.* [92] proposed a very similar model using the same input type, *i.e.* the cubemap projection (CMP), with shared weights among the different CNN channels. By sharing the weights over parallel channels, they are updated similarly according to all inputs. However, by using a pre-trained model (Inception-V3 [99]), the authors made the choice to not update the different weights. In the adopted strategy, the weight sharing advantages are not exploited. Xu *et al.* [100] proposed the VGCN model, in which the dependencies among possible viewports are exploited using a graph CNN. The proposed model features twenty ResNet-18 [87] as channels for the sampled viewports, in addition to a global branch for ERPs that uses the deep bilinear CNN (DB-CNN) [101]. The latter is composed of two different CNNs, including VGG-16 [102]. This makes the model significantly complex. Flowing the same work, Fu *et al.* [94] proposed a similar architecture, where the interaction among viewports is modeled using hyper-graph construction. Their model named AHGCN demonstrated a good performance.

Differently to multichannel models, Truong *et al.* [88] proposed the Omni-IQA model with a patch-based training scheme. The ERP content is used as input where patches are sampled according to a latitude-based sampling strategy. Therefore, patches of $64 \times 64$ are sampled from the ERP images and used with a simple CNN. During the validation, an equator-bias average pooling of patches' scores is applied to estimate the overall quality. The same issue regarding the direct use of ERP holds for this work as well. Similarly, Yang *et al.* [96] proposed a patch-based model named SAP-net using the ResNet-34 [87] model as a backbone. The input patches are enhanced using a wavelet-based enhancement CNN, which are then used as references to compute the error maps. Their selection is made randomly on ERP images, not taking into account the geometric distortion nor the relevance of the patch content. Furthermore, a simple arithmetic mean pooling is performed on the obtained scores to compute the final score of the 360-degree images, overlooking the non-uniformity distribution of the quality. Miaomiao *et al.* [95] integrated saliency prediction within the CNN model combining SP-NET [103] for saliency features extraction, and ResNet-50 [87] for visual features extraction. The model is trained using CMP faces, then fine-tuned directly on ERP images. Both CMP and ERP represent distortions related to the projection process, especially ERP as discussed previously, making the adopted training strategy less consistent with the explored content.

The requirement for dedicated quality models for 360-IQA with traditional approaches, led to the design of few models based on NSS and structural characteristics [32, 97, 98]. For instance, Zhou *et al.* [98] proposed the MFLGN model by exploiting the naturalness features and multi-frequency analysis of ERP images in addition to sampled viewports. By combining the features from the ERP (global) and viewports (local) and regressing them via a support vector regression (SVR) [104]

module, their model is able to provide good results. The main idea behind the adoption of NSS-based 360-IQA models is to avoid the use of deep-learning, however the generalization to other content and distortions of such an approach remains less compared to deep-learning based models. As a results, others targeted specific distortions related to 360-IQA, such as the stitching problem [105, 106]. However, this restricts their capacity to generalize to other prevalent distortions and applications as well.

## I.3 Conclusion

This chapter provides an overview of immersive media, with a focus on 360-degree images. The fundamental characteristics of 360-degree images are highlighted, in addition to the most important issues facing the processing of this emerging type of content. Especially, perceptual quality assessment, where several aspect are implicated, including the nature of content, used devices and technologies, exploration behavior, and the psych-physical state of the user. Furthermore, a literature review of existing studies with regard to SQA and OQA is provided by considering their advantages and drawbacks.

Thereafter, this chapter highlights the usefulness of SQA studies and their importance in designing and developing accurate, predictive, and consistent IQA tools. SQA remains arduous and exceedingly difficult. This is more emphasised with immersive media in general, and 360-degree in particular. As visual perception in virtual environment still relatively uninvestigated, especially when HMDs are involved, SQA experiments may be subject to *unreliableness*. It becomes of paramount important to study all factors impacting the collection of subjective data, which are the foundation of OQA models development. The ascertainment of SQA experiments and the resulted data for 360-degree images, in particular the used HMDs, is crucial for paving the way toward consistent quality assessment methods.

Following that, OQA is investigated for 360-degree images and the major drawbacks are highlighted. The lack of large perceptually annotated databases is a major issue. It is acknowledged that reliable and representative databases would allow IQA models in general, and deep-learning based ones in particular, to achieve significant performances. With the impressive growth of deep-learning approaches, it has become critical to build a large scale database as representative of the field, *i.e.* 360-IQA.

Considering that building such a database is arduous, improving the processing chain of 360-IQa at all scales could be a solution to cope with such lack. In fact, designing adaptive strategies, starting from inputs sampling and data representations to model's architecture and training strategies, is effective. With a proper training strategy, deep-learning models may require less data to achieve robustness. In addition, making use of transfer-learning, fine-tuning, and domain adaptation techniques, works fine in such cases. However, their usefulness for a challenging type of content such as 360-degree necessitate extensive studies, as one can not rely on previous observations from other image processing tasks.

Despite the effectiveness of deep-learning on extracting, encoding, and representation of visual features, their capability may be enhanced with IQA-specific features. The latter can be used at different scales to assist the learning of the model. Among such features, one can find visual attention, exploration behavior, JND, etc.

# Chapter II

# On the Influence of Head-Mounted Displays On Quality Ratings

## II.1 Introduction

The most effective and reliable method to assess the quality of 360-degree images intended for human users, is to ask human subjects for their opinion, hence SQA. Because of the human participation in the process, SQA is impracticable for most applications. However, it gives useful data for evaluating the performance of objective quality models. Subjective studies offer a mechanism to evaluate the effectiveness of IQA tools. Beside, it serves as the foundation of model design toward the ultimate objective of mimicking, reproducing, and matching the human visual perception [107], and the exploration behavior in virtual environments.

In existing studies featuring subjective quality assessment of 360-degree images, the impact of HMDs is usually overlooked. To the best of our knowledge, there has been no extensive study on the impact of HMDs on the perceived quality of 360-degree images in general, and the quality ratings in particular. Besides, IQA requires the use of reliable and representative databases to accurately develop and evaluate OQA models. Currently, the existing ones are lacking diversity in terms of content and cannot be considered as representative of the field. In this chapter, we evaluate the impact of various HMDs on perceived quality, from the technological and the rendered content points of view. First, we explore whether the use of different HMDs results in different quality ratings for the same content and conditions. For this, we build a database and define a controlled paradigm to conduct the subjective experiment. Then, we study the comfort of the viewers by means of the simulator sickness questionnaires. A statistical study of the obtained results is performed so as to compare HMDs and draw conclusions related to QoE. It is worth noting that this research was conducted during the Covid-19, which made recruiting a large number of participants quite challenging.

## II.2 360-IQA databases

The availability of reliable and representative databases is a critical factor in developing image quality models. It allows obtaining accurate and well-generalized IQA models. Particularly, deep-learning based models, where the performance is only as accurate as the variety, reliability, and representativeness of the available training data. Unfortunately, there is a significant lack of 360-degree perceptual annotated image quality databases. It is mostly due to the complexity and difficulty of the construction task. Indeed, building an IQA database requires subjective experiments to gather human opinions represented as MOS, in addition to an appropriate environment and test conditions. For instance, for 360-IQA subjective experiments, observers view the test images using HMDs. These devices are far from offering a perfect representation of 360-degree content. They may introduce some impairments like the SDE with an impact on the quality ratings. Neglecting such phenomena may result in an unreliable evaluation. Another common issue with subjective scores in general is the non-linear nature of the obtained scores requiring a non-linear regression using a five parameter logistic function, as recommended in the ITU-R recommendations [44], prior to the performance evaluation. However, it is to be recalled that such a regression cannot be performed if the native correlation is below 0.7. Otherwise, the correlation value cannot be considered as reliable because the quality of regression would be very low.

To date, a few databases for 360-IQA are proposed in the literature including Huang et al. [36], CVIQ [90], OIQA [108], MVAQD [32], IQA-ODI [96], and Sui *et al.* [82]. Unfortunately, only CVIQ, OIQA, and MVAQD are large enough, publicly available, and widely adopted for 360-IQA models training. A previous comparative study in [109] showed a very low correlation between IQA metrics and MOS provided by Huang. *et al.* compared to CVIQ. This opens questions regarding the reliability of existing databases. Questions are still under investigation, especially regarding the use of HMDs for subjective experiments. It is a very delicate context as there are no recommendations nor guidelines on how to perform such experiments for such applications.

Table. II.1 summarizes the characteristics of existing databases in the literature in terms of number of reference/distorted images, number of subjects participating to the subjective experiments, quality distortion types, and the used HMD. In the following, we provide details on each database.

**Huang et al. [36]:** It contains 25 pristine 360-degree images used to create 12 versions for each one. Four distinct spatial resolutions and three JPEG quality factors (QF), where $QF \in \{25, 60, 100\}$ are used to create 300 distorted images at resolutions of 4k, 2K, 1080p, and 720p. The MOS was obtained using ACR with 98 subjects participating to the test (53 males and 45 females). The quality scale ranges from 0 (Bad) to 100 (Excellent). Each subject rated only three different image contents at the four spatial resolution and three quality factors.

**CVIQ [90]:** This database is composed of 16 360-degree images and 528 compressed

**Table II.1:** Summary of state-of-the-art 360-degree image databases.

| Database | Huang et al. [36] | CVIQ [90] | OIQA [108] | MVAQD [32] | IQA-ODI [96] | Sui *et al.* [82] |
|---|---|---|---|---|---|---|
| Ref images | 25 | 16 | 16 | 15 | 120 | 36 |
| Distorted images | 300 | 528 | 320 | 300 | 960 | 72 |
| Distortion type (Distortion level) | JPEG (3) / Downsampling (4) | JPEG (11) / AVC (11) / HEVC (11) | JPEG (5) / JPEG2000 (5) / BLUR (5) / WGN (5) | BLUR (4) / HEVC (4) / JPEG (4) / JP2K (4) / WGN (4) | JPEG (4) / Projection (4) | HEVC (3) / Stiching (3) |
| Number of subjects | 98 (M: 53, F: 45) | 20 (M: 14, F: 6) | 20 (M: 15, F: 5) | 26 (M:16 , F: 10) | 200 (M: 138, F: 62) | 22 |
| HMD | HTC Vive | HTC Vive | HTC Vive | HTC Vive Pro | HTC Vive | HTC Vive |

versions. The compression artifacts are obtained using eleven levels of : 1) JPEG compression with QF ranging from 50 to 0 with a step of 5, 2) H.264/AVC and H.265/HEVC with quantization parameters (QP) from 30 to 50 with a step of 2. The authors used the ACR method with a rating scale of 10-levels from the lowest to the highest quality to gather the MOS. The ACR method was adopted with the participation of 20 subjects (14 males and 6 females).

**OIQA [108]:** It includes 320 distorted 360-degree images created from 16 reference ones using four distortion types with five levels each. The used distortions include JPEG, JPEG 2000 compression (JP2K), Gaussian blur (BLUR) and white Gaussian noise (WGN). JPEG and JP2K are applied directly on ERPs, while GB and WGN are applied individually on small blocks, and then stitched back to ERP. Subjective scores are given in the range from 1 (bad) to 10 (excellent). 20 subjects were involved in the test (15 males and 5 females).

**MVAQD [32]:** This database is composed of 300 distorted 360-degree images generated from 15 pristine ones. The 300 images are generated using five distortion types, including BLUR with $\sigma \in \{0.5, 1, 2, 5\}$, HEVC with $QP \in \{27, 32, 37, 42\}$, JPEG with $QF \in \{70, 50, 30, 10\}$, JP2K with $Bpp \in \{0.6, 0.3, 0.1, 0.05\}$, and WGN with $\sigma^2 \in \{0.001, 0.005, 0.01, 0.02\}$. Each distortion is applied using four levels. The quality ratings are collected from 26 subjects (16 males and 10 females) using a five-grade quality scale ranging from 1 (bad) to 5 (excellent).

**IQA-ODI [96]:** This database is to date the largest proposed one. It includes 1080 distorted 360-degree images, of which 120 are pristine. The distorted images are generated by applying the JPEG compression with $QF \in \{60, 35, 15, 5\}$ on ERP images, and JPEG with $QF = 15$ on CMP, CPP, ISP, and OHP projections. The subjective scores are obtained from 200 subjects (138 males and 62 females). All the subjects

are divided into 10 groups to only rate 108 images. The scores are provided as the differential MOS in the range of [0, 100].

**Sui *et al.* [82]:** This database was constructed to model the user's exploration behavior. It contains 36 pristine 360-degree images used to generate 72 distorted ones. The used distortions are HEVC/H.265 with $QP \in \{38, 44, 50\}$ and the stitching distortion with a parameter $p \in \{0.5, 0.75, 1\}$, where higher $p$ values imply severe distortions. 22 subjects participated to the subjective experiment to rate the quality using a five-grade quality scale ranging from 1 (bad) to 5 (excellent).

In comparison to Huang et al., MVAQD, and IQA-ODI, only CVIQ and OIQA databases have received attention in the literature. In addition, the study in [109] showed a poor correlation between MOS provided with Huang et al. [36] and objective quality metrics. This obviously shows the ineffectiveness of certain databases in contrast to others, which may raise questions about their representativeness as well as their reliability. Besides, not all existing databases are made publicly available.

During the completion of this thesis, only CVIQ, OIQA, and MVAQD were publicly available. Even though, their content may be subject to less diversity in terms of content and complexity of scenes. To verify this, we computed the spatial information (SI) and colorfulness information (CFI) of the pristine images on the each database. SI represents an indicator of edge energy, giving an idea about the spatial complexity of an image, whereas CFI is a perceptual indicator of the variety and intensity of colours. The SI and CFI are obtained according to ITU-T P.910 [110] recommendations and the metric described in [111], respectively. The CFI versus SI plot is provided in Fig. II.1 (Left). One can observe a lack of diversity on CVIQ as illustrated, whereas the SI and CFI indicators are more spread out for OIQA and MVAQD, depicting a higher diversity.

We further compare the histogram of the MOS values on CVIQ, OIQA, and MVAQD. From Fig. II.1 (Right), one can observe that OIQA and MVAQD depict a distribution that is not far from a uniform one. The histogram demonstrates that the MOS values cover the quality range, and the number of observations is properly balanced, on the one hand. CVIQ, on the other hand, presents a different distribution of subjective ratings, with MOS values that are not distributed evenly over the quality range, which tends to demonstrate problems of miss-conception.

## II.3 Study Design

### II.3.1 The proposed 360-IQAD database

Twenty pristine 360-degree images are selected to create our database, from which 240 distorted versions are created. First, we chose 360-degree images from the joint video exploration team (JVET) test sequences [112] and the SUN360 database [113]. In addition, to account for the synthesized content related to VR, four scenes have

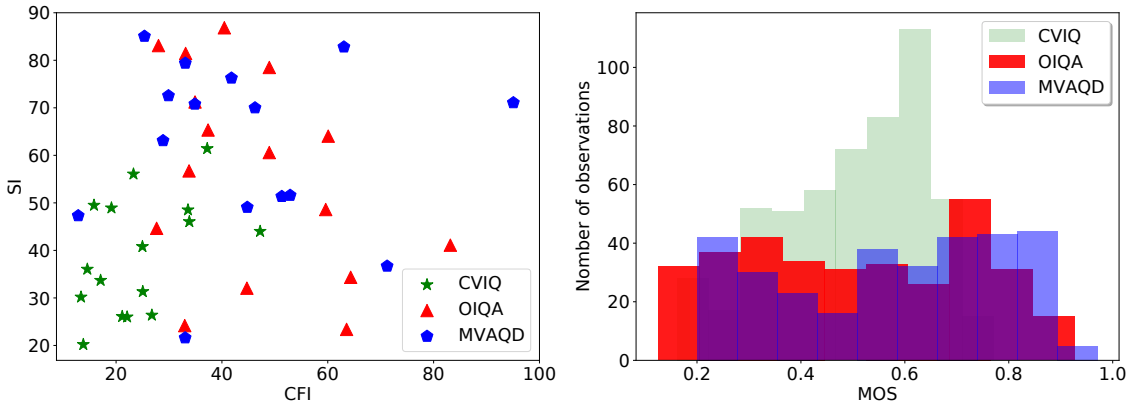



**Figure II.1:** (Left) spatial information (SI) versus colourfulness index (CFI) plot of CVIQ, OIQA, and MVAQD, and (right) histogram of their MOS values re-scaled to [0, 1].

been added to the database. The used images are given in Fig. II.2. As it can be seen, the images represent a variety of content types, including indoor and outdoor natural scenes, human scenes, landscapes, and nature as well as synthesized ones. Often, databases are constructed without paying attention to the diversity of the content. In our case, we account for two important characteristics, *i.e.* spatial complexity and colourfulness, described in the previous section. The CFI versus SI plot of the pristine images shown in Fig. II.3 aims at demonstrating the spatial and colour diversity of the selected images. One can notice that the used images span over the range of CFI and SI values.

Once selected, images are distorted using the JPEG compression with $QF \in \{80, 60, 40, 20\}$, BLUR with $\sigma \in \{0.5, 1, 1.5, 2\}$, and WGN with $\sigma^2 \in \{5, 50, 100, 150\}$. For each distortion type, four levels are applied to cover the perceived quality range from annoying to imperceptible. The levels are purposefully chosen in such a way that the perceived difference between them is obvious for observers. From each pristine 360-degree image, 12 distorted ones are generated.

### II.3.2 Subjective assessment protocol

In order to construct a reliable database, the selection of subjective protocol is of paramount importance. Unfortunately, at the time of the study there were no guidelines for conducting experiments for immersive applications. In our case, we built the test by relying on the ITU recommendation ITU-BT.500 [114]. Hence, the adopted protocol is depicted in Fig. II.4. It is scrupulously followed by each observer. First, the observer is screened for visual acuity and color blindness in order to collect reliable scores. Then, he is asked to complete a simulator sickness questionnaire (SSQ) before beginning the test. For this aim, we used the virtual reality sickness questionnaire (VRSQ) proposed in [20]. The VRSQ consists of nine questions in which the observer is asked to rate the severeness of nine symptoms on a four scale (None: 0, Slight:

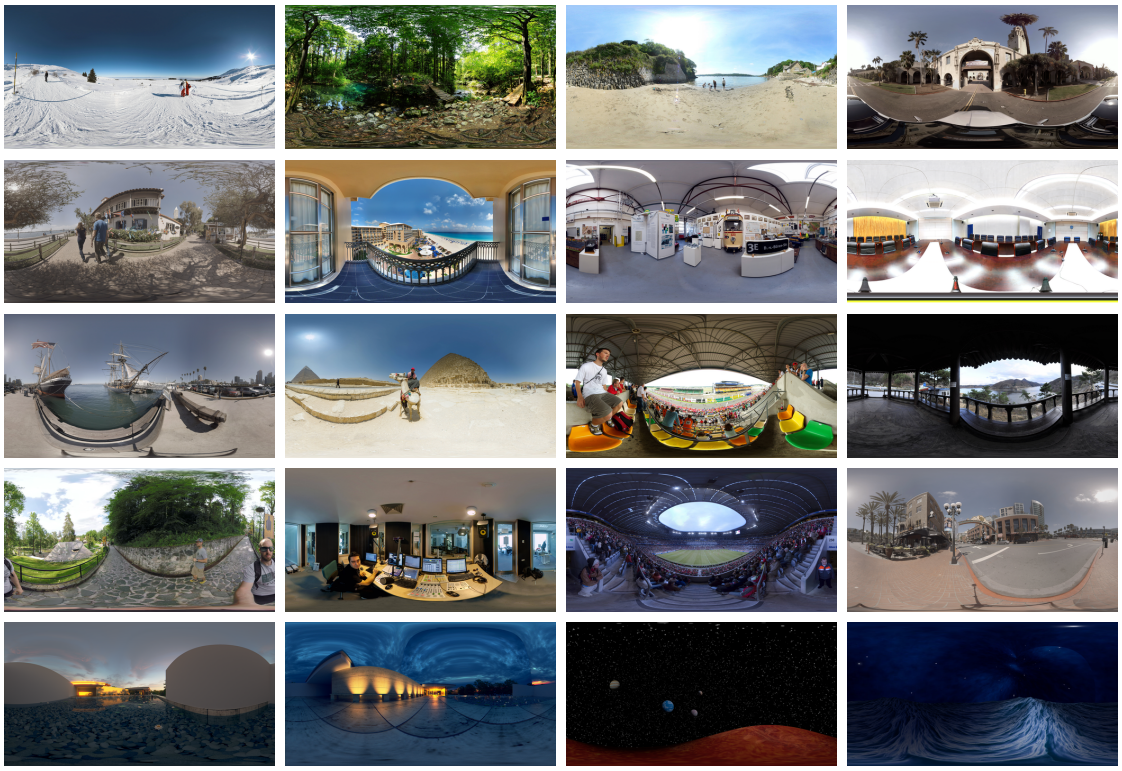


**Figure II.2:** Pristine images in the proposed database. 1-4 rows are images taken from JVET and SUN360. Row 5 are created as synthesized images.

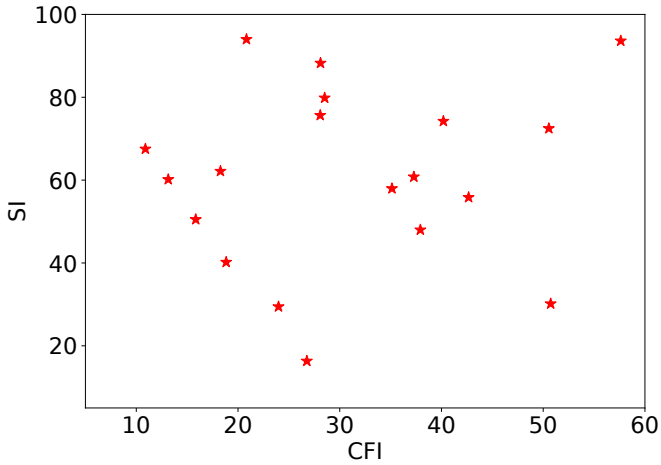



**Figure II.3:** Spatial information (SI) versus colourfulness index (CFI) plot of the selected pristine images used for the construction of the database.

1, Moderate: 2, Severe: 3). Individual symptoms are classified into three categories: oculomotor agitation (O), disorientation (D) and total score (TS). The use of SSQ prior to beginning the test serves as the observer's initial state and is used to compare the progression of the symptoms described in the VRSQ. It is known that conducting a subjective test in the morning or the afternoon may lead to a different assessment because of the psycho-visual state of the observer. The observer is then trained on a few samples of 360-degree images with perceptual qualities corresponding to those used in the test. The training is designed to familiarize the observer with the task at hand, and get familiar with VR environment, since not all observers are VR or HMD users. Samples used in this session are for training purposes only and are discarded from the experiment results.

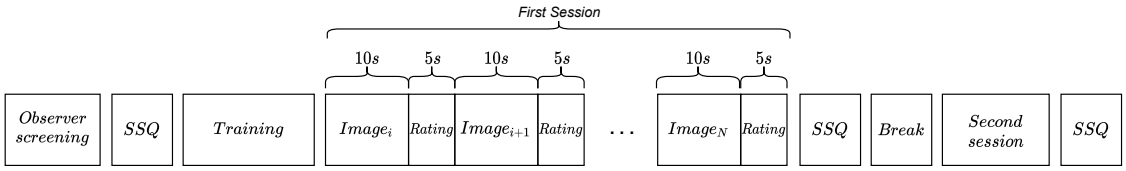



**Figure II.4:** Illustration of the adopted subjective assessment protocol.

After the training session, the first session starts by asking the observer to rate the quality of the impaired 360-degree images using a five-point quality scale ranging from 5 (excellent) to 1 (bad), following the ACR method. This quality scale should be sufficient to cover the quality levels used in the constructed database and where the maximum quality corresponds to the pristine images. In the first session, the observer rates a hundred and thirty images, corresponding to a duration of 32.5 min (130 samples $\times 15s$). This duration is reasonable as the test deals with images only (*i.e.* there are no motions as in the case of videos). In addition, the observer can stop the test any time based on his psychophysical state. After the first session, the observer fill out another SSQ so to record his state after experiencing omnidirectional images for approximately half an hour. After a sufficient break, the second session takes place with the remaining images from the database. Finally, at the end of the second session, another SSQ is filled out. In order to collect reliable results, we ensured that all observers followed the exact same protocol. The images playlists are randomly constructed, and each observer watches a random one in order to avoid rating biases.

The HMDs considered in this study are from different manufacturers, and have specific characteristics each. Table II.2 summarizes the ones that may contribute to the quality assessment task, and Fig. II.5 shows them.

The observers were recruited from our university, and they are all naive. Due to the sanitary situation during this study (Covid-19 pandemic), running subjective experiments become very challenging. In our case, the experiment with four HMDs lasts for about six hours for a single observer. This is why, the results exposed here are based on eight valid observers per HMD.

**Table II.2:** Characteristics of the considered HMDs.

| HMD | Resolution per eye | FoV | PPD |
|---|---|---|---|
| Varjo Vr-2 | 1920 × 1080 | 87° | 60 |
| HTC Vive Pro | 1440 × 1600 | 110° | 13.09 |
| HP Reverb VR | 2160 × 2160 | 114° | 18.94 |
| Oculus Quest | 1600 × 1440 | 100° | 14.4 |




**Figure II.5:** HMDs used in the subjective study.

## II.4 Results and discussion

In the following, we provide and analyse the subjective quality evaluation results. First, we investigate whether there is a substantial difference in terms of quality ratings between HMDs on the one hand and particular distortions on the other hand. The last part of this section will concentrate on analysing the simulator sickness questionnaire results.

### II.4.1 Effects of HMD on subjective ratings

It is known that, the use of different devices for SQA may result in different outcomes since each device has unique properties. In the case of 360-IQA, the device is the HMD. It is critical to establish whether such a difference is substantial, especially if it impacts the overall QoE. To that purpose, various questions are framed in order to determine the impact of using HMDs for 360-degree image SQA. In this study, these questions are roughly summarized as follows:

- Would the use of various HMDs result in different ratings?
- What is the inter-observer difference?
- Is the impact of a single distortion the same regardless of the used HMD?
- Which HMD offers the best quality?
- What about comfort and cyber-sickness?

Thanks to a statistical analysis of the obtained scores, we aim to find answers to the above questions. The histograms of the gathered rating scores of all HMDs are shown in Fig. II.6. We can clearly observe that the ratings span across the five perceptual quality scales with a Gaussian shape regardless of the used HMD, on the

one hand. On the other hand, the MOS versus its 95% confidence interval (CI) in Fig. II.7 shows bigger CIs for MOS values when $2 < MOS < 4$, whereas smaller CIs with $2 > MOS > 4$ . It is acknowledged that bad or excellent qualities are easy to distinguish for the users compared to medium qualities where observers may disagree.

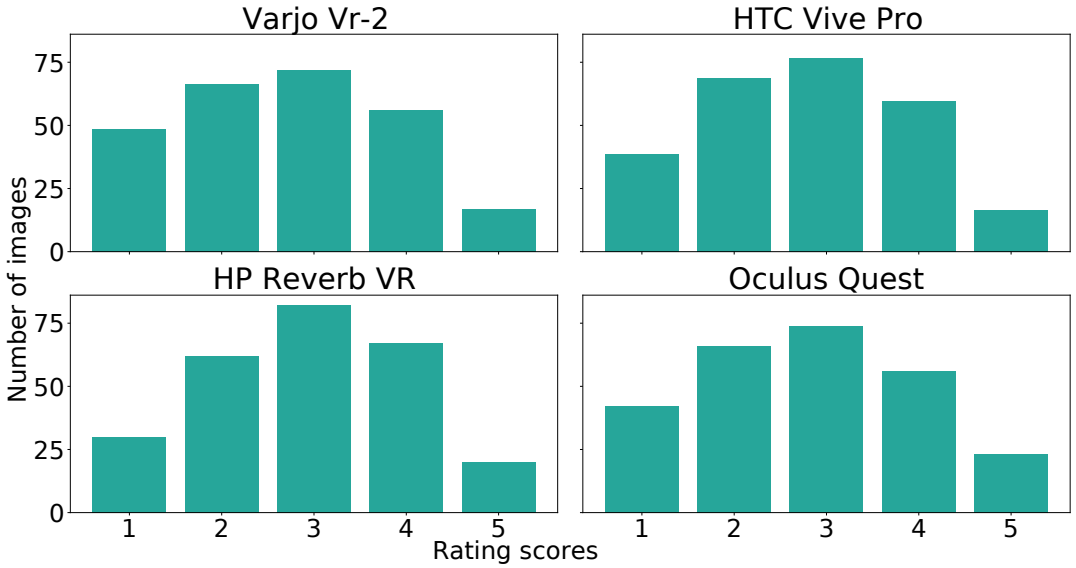



**Figure II.6:** Histograms of the rating scores obtained by the used HMDs.

Prior to investigating the effect of HMDs on ratings, we calculated the standard deviation of opinion scores (SOS) using the SOS hypothesis described in [115], which is defined by Eq. II.1 as follows:

$$SOS(x)^2 = -ax^2 + 6ax - 5a \tag{II.1}$$

The SOS parameter $a$ quantifies the uncertainty ratio among observers on a scale of 0 to 1. It reflects the inter-observer reliability, where a value of 0 denotes a full agreement among all observers, and 1 indicates a maximum variance. Table. II.3 provides the $a$ values of the obtained scores by the four HMDs. As it can be seen, the SOS parameter for all HMDs is in the range $0.03 < a < 0.06$. Based on the description given above, this interval of values demonstrates an inter-observer agreement and reliability of approximately 90%. This observation substantiates the overall efficacy of the constructed experiments and the adopted procedure.

**Table II.3:** SOS parameter $a$ of all HMDs' rating scores.

| Varjo Vr-2 | HTC Vive Pro | HP Reverb VR | Oculus Quest |
|---|---|---|---|
| 0.0361 | 0.0414 | 0.0338 | 0.0504 |

In order to statistically assess the impact of HMDs on the quality rating, we analyse

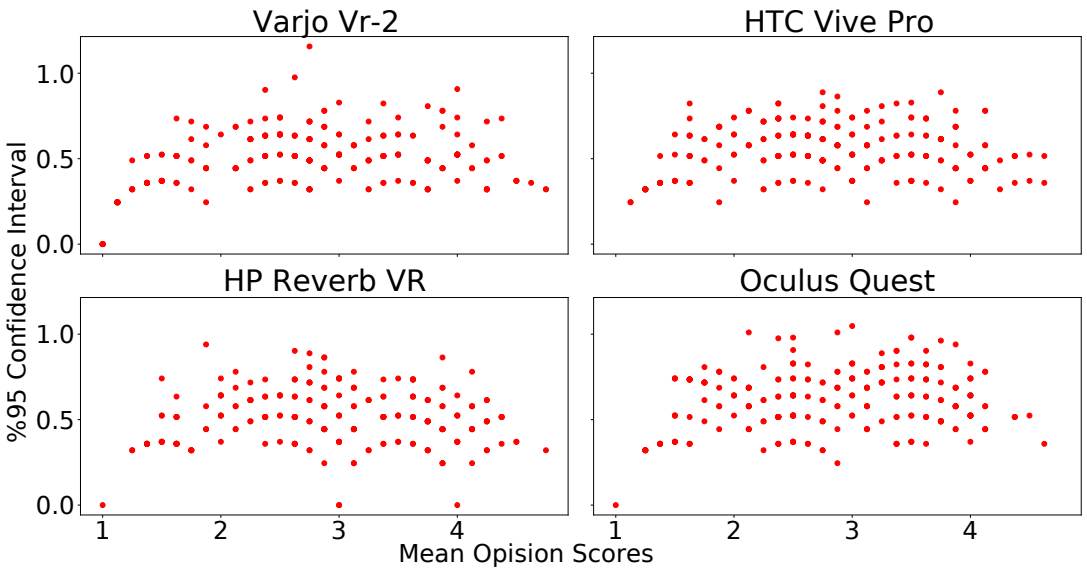



**Figure II.7:** Confidence interval of the MOS generated by the used HMDs.

the variance between the obtained MOSs. The following are the null hypothesis $H_0$ and the alternative one $H_1$:

**$H_0$:** There is no significant difference between the four HMDs.

**$H_1$:** At least one HMD is significantly different from the others.

To analyse the variance, the use of ANOVA [116] is a good choice. However, the ANOVA assumes that the sample data is normality distributed. Therefore, a normality check is performed, and the probability distribution for each HMD is illustrated in Fig. II.8. The formula used for the theoretical quantiles (horizontal axis of the probability plot) is the Filliben's estimate [117]. Looking at the plots, we see an upward sloping linear relationship. Deviations by the dots from the line can be observed around both extremities. The sample data (*i.e.* MOS) partially fits the diagonal line, which shows a deviation from the expected normal distribution. This demonstrates that the distribution of the gathered MOSs is not perfectly normal but very close. Based on this observation and in order to reliably analyse the variance, a non-parametric test is applied in addition to ANOVA. Here, the Kruskal-Wallis H-test [118] is used.

The ANOVA showed a $p-value$ of 0.035 while Kruskal a $p-value$ of 0.038, leading to the rejection of $H_0$, implying that the HMD has a statistically significant influence on the quality ratings. One possible explanation could be the screen door effect explained previously. This observation contrasts with the results reported in [21], which found that the effect of HMDs is not significant compared to other factors. Since a statistical difference is found, we further analyse specific differences between HMDs. A post hoc test [119] is performed in this case, and a significance plot is provided in Fig. II.9 (a). It appears that, the source of the identified differences is significant between HP Reverb VR and HTC Vive Pro, and more emphasized with Varjo Vr-2. This demon-

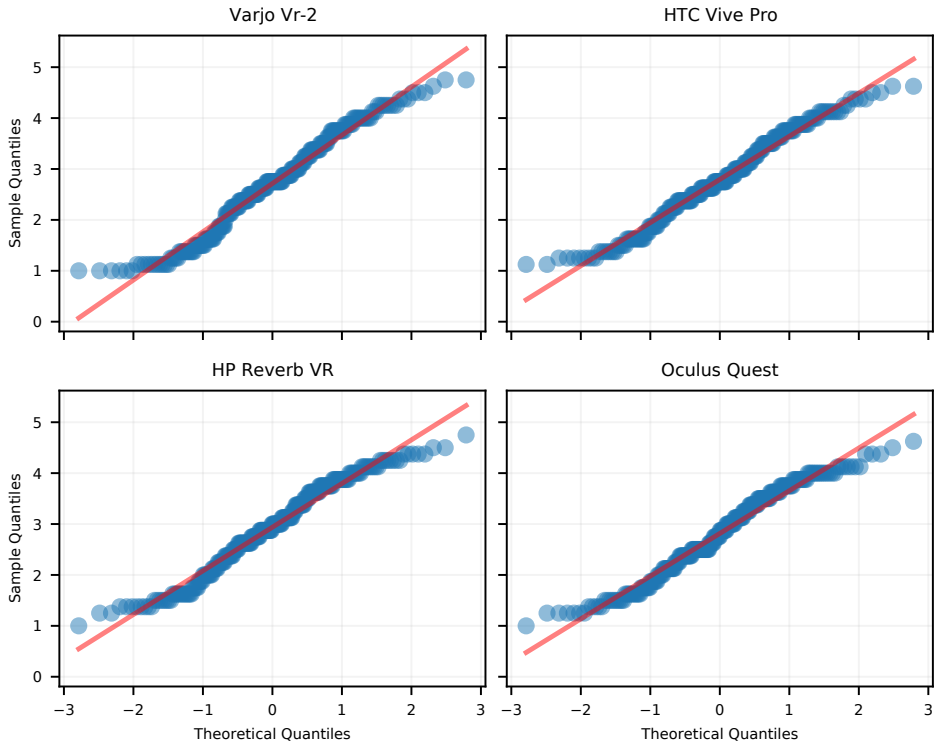
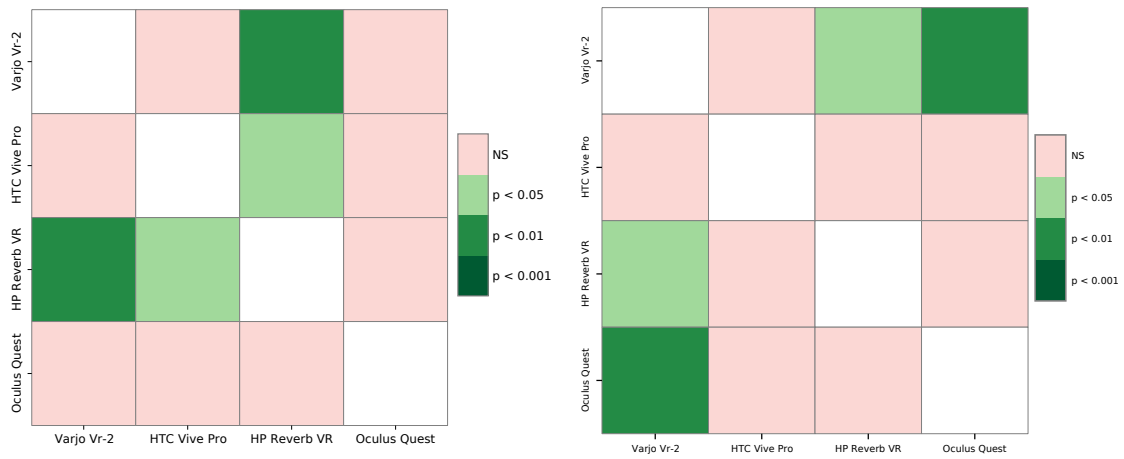



**Figure II.8:** Probability plot of MOS against the normal distribution quantiles.

strates that the variance occurs from multiple HMDs, suggesting that the collected subjective scores are most likely to be significantly different with respect to the used HMD. Regardubg the Oculus Quest, it seems there is no significant difference when compared any of the selected HMDs.

In addition to the analysis of variance on the overall scores, we performed an analysis of variance using the Kruskal-Wallis H-test between HMDs per individual distortions, with the aim of evaluating the effect of a single distortion independently of the used HMD. A $p-value$ of 0.023, 0.274, and 1 are obtained for JPEG, GB, and GN, respectively. In this case, observers noticed a difference between HMDs for JPEG but not for the remaining distortions. One may question the link between JPEG artefacts and the SDE, presenting some similarities in terms of distortion type (*i.e.* blocking artefacts). This observation backs up the previous one about the differences on the overall ratings. Additionally, we looked into the differences regarding the JPEG distortion, the significance difference is depicted in Fig. II.9 (b). One can notice that the difference here is between Varjo Vr-2 and Oculus Quest, as well as with HP Reverb VR.

When comparing among the statistical difference on the overall MOS and JPEG one, Varjo Vr-2 is identified as statistically different in both cases. The significant




(a) Overall MOS (b) JPEG MOS

**Figure II.9:** Pairwise multiple significance plot between HMDs. "NS" stands for no significance.

difference in PPD (*see* table II.2) between this HMD and the others, which greatly contributes to visual quality, may explain such a result.

We examined the MOS obtained for all HMDs to determine which one provides the best quality, and how the MOS per individual distortions is distributed. A box plot of MOSs from all images per single distortion is depicted in Figure II.10. Overall, we can notice that the MOSs for a certain distortion level are mostly within a limited range. This confirms the findings of the SOS parameter, which was previously discussed. One can also notice that, compared to JPEG and GN, GB was frequently rated as bad (1) and poor (2). Especially, levels 3 and 4 where the means fall in the same range. This clearly shows that the observers were annoyed by such a distortion regardless of the used HMD. For GN, the MOS mostly falls in the same range for level 2, 3 and 4, as if the observers did not perceive much difference between these levels. This is particularly true with HTC Vive Pro and Varjo Vr-2. In terms of which HMDs provides a better quality, we can observe that with Varjo VR-2 and HP Reverb VR more MOS greater than 3.5 were given. This suggests that these two offer better quality, and can be related to their resolution and PPD (*see* Table. II.2).

### II.4.2 Simulator Sickness Assessment

We computed the simulator-sickness scores, as mentioned previously, to measure the sickness level caused by each of the used HMDs. The scores are grouped by total scores (TS), oculomotor (O), and disorientation (D) as described in the VRSQ [20]. The VRSQ is derived from well-known SSQ [19] where 9 symptoms are selected among 16. Table II.4 summarizes these symptoms and classifies them into the oculomotor and disorientation categories. The scores of TS, O and D are computed as follow:

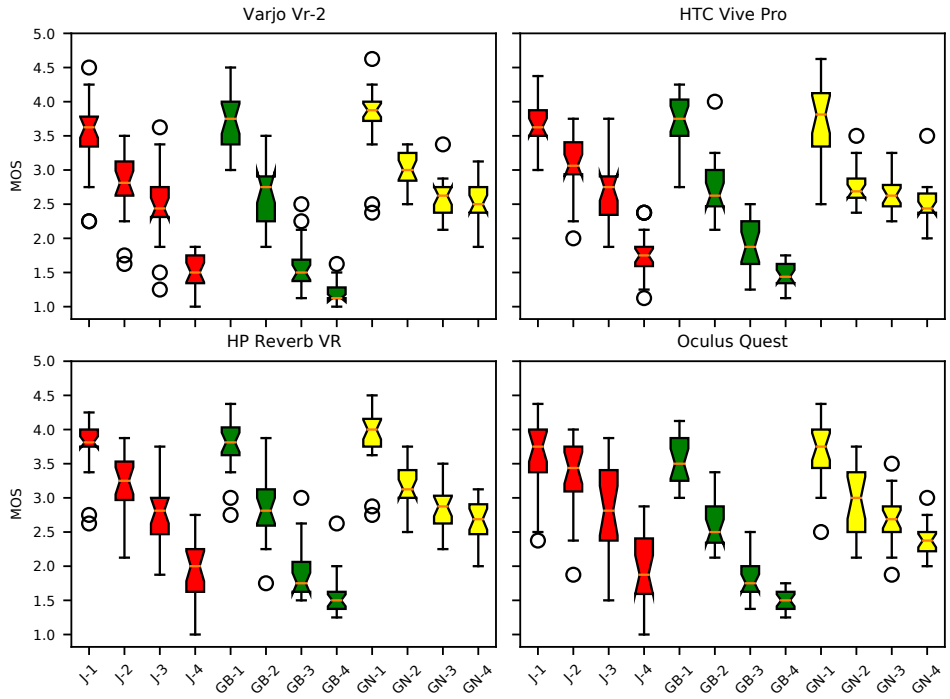



**Figure II.10:** Box plot of MOS per level of distortion, where J-*, GB-* and GN-* stand for JPEG, Gaussian Blur and Gaussian Noise with 4 different levels.

**Table II.4:** Virtual reality sickness questionnaire [20].

| Symptoms | Oculomotor (O) | Disorientation (D) |
|---|---|---|
| General discomfort | ✗ | |
| Fatigue | ✗ | |
| Eyestrain | ✗ | |
| Difficulty focusing | ✗ | |
| Headache | | ✗ |
| Fullness of head | | ✗ |
| Blurred visio | | ✗ |
| Dizzy (eyes closed) | | ✗ |
| Vertigo | | ✗ |
| Total | $T_O$ | $T_D$ |

$$O = \frac{T_O}{12} \times 100, \tag{II.2}$$

$$D = \frac{T_D}{15} \times 100, \tag{II.3}$$

$$TS = \frac{(O+D)}{2}, \tag{II.4}$$

where $T_O$ and $T_D$ stand for the sum of the O and D symptoms for each participant. Then, the average of all participant per HMD is taken. Here, a score around 40 is considered as severe. Fig. II.11 shows the histograms of the simulator-sickness scores obtained for the selected HMDs.

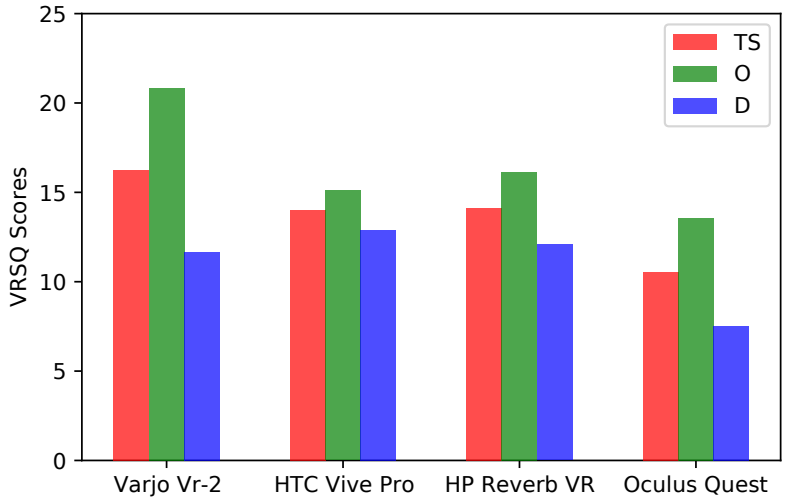



**Figure II.11:** Simulator-sickness scores for the considered HMDs in terms of total scores (TS), oculomotor (O), and disorientation (D).

For this experiments, we focused on analysing which HMD causes higher sickness in terms of TS, O and D. In comparison to the others, the Varjo VR-2 received the highest overall scores, while, the Oculus Quests received the lowest. Two explanations could convey these results. First, the weight of Varjo VR-2, reported by the observers as being high. Then, the double displays composing this HMD with two different resolutions, often requiring an adapted content. One can also see that the observers are more prone to oculomotor symptoms compared to those for disorientation and even the TS. The length of the sessions where the observers are subject to very close displays may be a reason for this. As the oculomotor involves eye strain, difficulty focusing, and fatigue, which can be increased with more exposition to omnidirectional images.

## II.5 Conclusion

In this chapter, we present a detailed evaluation of the impact of HMDs on the quality rating of 360-degree images. The provided analysis revealed a statistically significant difference between the used HMDs. This difference is mostly related to the distinct characteristics of each HMD. Especially the SDE, which can be confused with the distortions on the viewed scenes, and may lead to an unreliable assessment. This contrasts with previous observations in the literature. Furthermore, a significant difference on specific distortions was also observed with JPEG compared to GN and GB. The source of such difference was found between multiple HMDs supporting the observations regarding the device induced influence. Additionally, the simulator-sickness assessment revealed that the use of some HMDs lead to a higher simulator sickness scores compared to others, and oculomotor related symptoms induce significantly higher scores when compared to disorientation.

Despite the fact of conducting the presented study during the COVID-19 pandemic, we were able to draw interesting conclusions. A further analysis including additional factors and more participant would allow to verify and validate to what extent the HMDs influence the subjective quality ratings for 360-IQA. To this end, a holistic assessment is planned for future studies.

# Chapter III

# Convolutional Neural Networks for 360-IQA: A Benchmark

## III.1 Introduction

The absence of large, accurate, and representative perceptually annotated databases is a major issue when dealing with 360-IQA. The construction of such databases require arduous efforts in terms of scenes acquisition, device calibration, paradigm definition, subjective testing and data analysis [36, 90, 109]. As a consequences, the progress of 360-IQA is held back. As an alternative, well-known pre-trained models such as ResNet-18/34/50 [87], Vgg-16 [102] and DenseNet-121 [120] may cope with the lack of training data. It can be performed by the mean of transfer-learning (TL) and domain adaptation. The latter insure a transfer from a source domain to a target one. In the case of the aforementioned pre-trained models, the source domain is image classification. As mentioned in Sec. I.2.2.2, existing 360-IQA models have fine-tuned a well-known pre-trained model for a different task within the IQA framework. The main reason behind such choices lies in the fact that the used models are trained on very large databases, allowing them to reach a significant learning level. For instance, the ImageNet [121] database used to train existing pre-trained models contains over 14 millions images.

The use of pre-trained models for 360-IQA is becoming more current, owing to their popularity from other image processing tasks, several important questions are raised. In particular, questions related to the use of pre-trained models for 360-IQA and the exploitation of the rich state-of-the-art work dedicated to 2D content, such as:

- Prediction accuracy of pre-trained models for 360-degree images quality? And which model is performing the best?
- Radial vs. projected content-based training?
- Performance of projected format: CMP versus ERP?
- Performance of Patch-based training schemes?

- Would 2D quality databases improve the performance of CNN models?

In this chapter, an empirical and extensive analysis is conducted using different and widely used CNN models to answer the above-mentioned questions. Several models are considered, including ResNet-18/34/50 [87], Vgg-16/19 [102], DenseNet-121 [120] and Inception-V3 [99]. These models are compared under different configurations related to omnidirectional and spherical characteristics. The novelty of this work lies in the fact of providing answers to the above questions, for a very challenging type of content *i.e.* 360-degree images, as one cannot rely on conclusions drawn from standard 2D benchmarks [122, 123], not taking into account the targeted characteristics.

## III.2 The proposed benchmark: design and architecture

With the intent to provide a holistic study as well as recommendations on the use of CNNs for 360-IQA and to answer the questions raised in Sec. III.1, we designed a benchmark taking multiple considerations into account, related to the use of: 1) Content based splitting criteria for selecting training and validation sets, 2) Projected images as ERPs and CMPs, 3) Radial content rather than projected one, 4) multichannel CNN architecture, 5) Patch-based learning scheme, and 6) 2D benchmark IQA databases to train the selected models.

### III.2.1 Pre-trained CNN models

In this study, seven among the widely used models are exploited and compared. A brief description of their architecture is provided below in addition to Table III.1 giving their number of parameters and output size. All used models are fine-tuned by replacing the original top layers used for classification, with a quality regressor block (*see* Fig. III.1). The latter first performs a dimentionality reduction by applying a global average pooling (GAP) [124] on the extracted feature maps $\mathbf{F} \in \mathbb{R}^{D\times H\times W}$ by the pre-trained models, where $D$, $H$, and $W$ stand for the dimension, height and width, respectively. This operation produces a feature vector $V_{\mathbf{F}} \in \mathbb{R}^{D\times 1\times 1}$, and can be formally expressed as :

$$y^c = \frac{1}{N}\sum_{i,j}\mathbf{F}^D_{i,j}, \tag{III.1}$$

where $y^c$ is the output value of feature map F at channel $d$ .$(i, j)$ is the pixel index in the feature map $\mathbf{F}^d$. The GAP is useful to minimize overfitting. The generated $V_{\mathbf{F}} \in \mathbb{R}^{D\times 1\times 1}$ goes then to a fully connected (FC) layer followed by a rectified linear unit (ReLU) [125] activation function and a dropout layer. The latter is an effective

regularization method to reduce overfitting and improve generalization error in deep neural networks [126]. Finally, a FC layer with a single node and a linear activation function is used to deliver the quality score. The weights of the quality regressor are initialized according to [127]. During training, all layers of the pre-trained models are freezed to rely on the weights from ImageNet [121], and only the quality regressor block is trained for the IQA task.

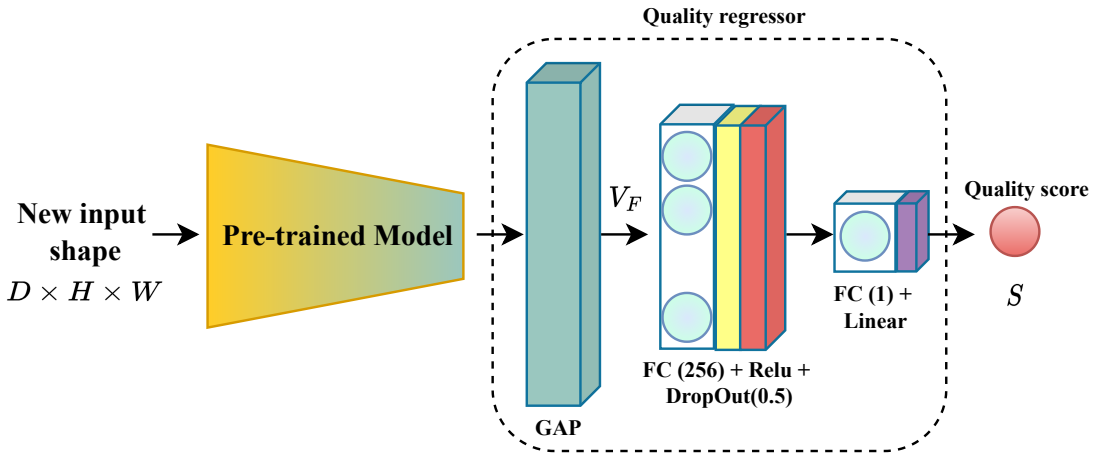



**Figure III.1:** Architecture of the CNN models: Top layers replaced by a regression block composed of a global average pooling (GAP) layer, a fully connected layer (FC), a dropout layer and a final FC layer to output the predicted score $S$. $V_F \in \mathbb{R}^{D\times1\times1}$ represents the extracted features vector.

In the following, we describe the used CNN models by giving the most important characteristics, leaving the readers to refer to the original cited works.

**ResNet :** residual networks are artificial neural networks introduced in 2015 [87]. The ResNet utilizes skip connections to jump over some layers. This helps training deeper networks without falling into the problem of vanishing gradients. ResNet employs residual learning to further deepen the CNN network, which can be interpreted by a number of deeper bottleneck architectures. Each bottleneck has three convolutional layers with kernel dimensions of $1 \times 1$, $3 \times 3$ and, $1 \times 1$ respectively. A shortcut connection is then added from the input of the bottleneck to its output. Several versions of this model were developed with the main difference lying in the number of layers. We use ResNet-18/34/50 in this study.

**VGG :** it is a convolutional neural network architecture proposed in [102]. This network is characterized by its simplicity and use only $3 \times 3$ convolutional layers stacked on top of each other in an increasing depth. It also includes $1 \times 1$ convolution filters acting as a linear transformation of the input, followed by ReLU [125] activation. The convolution stride is fixed to 1 pixel, so to preserve the spatial resolutions. Different versions of this network exists, but we only focus on the widely used ones *i.e.* Vgg-16/19 [91, 101, 122, 123].

**DenseNet :** it is a neural network composed of dense blocks introduced in [120]. In each block, the layers are densely connected, with $L(L+1)/2$ direct connections, where $L$ is the number of layers. Each layer in DenseNet receives additional input from all preceding layers and concatenates them with its own feature-maps before feeding them to the subsequent layers. This allows the model to reuse low-level features. The DenseNet-121 is considered in this study with the configuration used in [128].

**Inception :** this network architecture introduced in [99] is composed of convolutional blocks known as Inception modules. The latter contains $1 \times 1$, $3 \times 3$ and $5 \times 5$ convolutions as well as a pooling layer. The introduction of such module aims to allow for more efficient computation and deeper networks through a dimensionality reduction as well as the use of various convolutional filter sizes instead of using a single one. Several versions of this network also exist. The Inception-V3 model introduced factorized and smaller convolutions, helping to reduce the computational cost by decreasing the number of parameters involved in the network. This version of Inception is used in our comparative study.

**Table III.1:** Number of parameters (in million) in each selected model without their top layers, and the dimension of the output vector for feature representation (fv).

| Model | #Params (M)≈ | Size of $V_F$ |
|---|---|---|
| ResNet-50 | 23 | 2048 |
| ResNet-34 | 21 | 512 |
| ResNet-18 | 11 | 512 |
| Vgg-16 | 14 | 512 |
| Vgg-19 | 20 | 512 |
| DenseNet-121 | 7 | 1024 |
| Inception-V3 | 21 | 2048 |

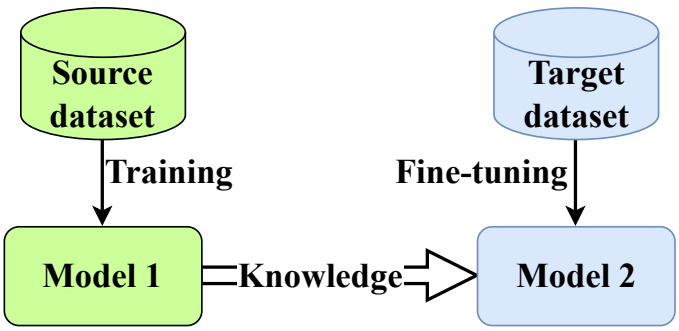



**Figure III.2:** Process of transfer learning from a source to a target domain.

To effectively conduct transfer learning from another domain, for instance image classification, using the abovementioned pre-trained models, fine-tuning is required.

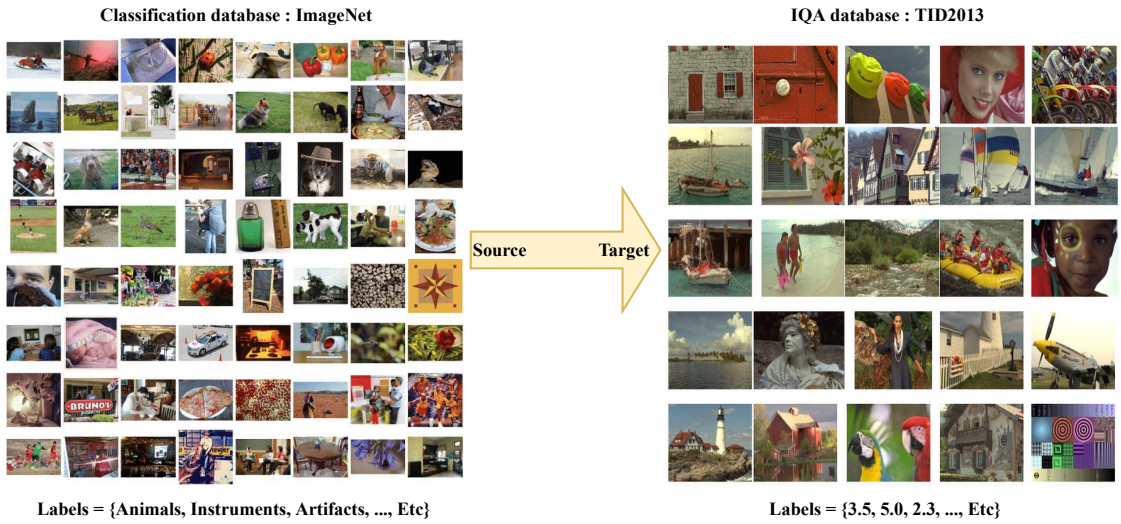



**Figure III.3:** Illustration of the source-to-target domain transfer. The labels of the classification database (ImageNet [121]) are classes, whereas the IQA database (TID2013 [130]) are the MOSs (continuous values).

To be more explicit, the knowledge acquired by the model after training for a specific task, may be exploited by a new target task. This is possible by means of knowledge transfer and fine-tuning as depicted on Fig. III.2. The latter allows removing the constraints on the label spaces of the source and target domains, *i.e.* from object classes to MOSs. An illustration of the source-to-target domain transfer between Imagenet and an IQA database (TID2013) is provided in Fig. III.3. Following the formulation suggested in [129], transfer learning can be expressed as follows:

$$f_s^* = \underset{f_s \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{N_s} \sum_{i=1}^{N_s} l_s(f_s(x_s, i, q_s, i)) + \alpha R(D_s, f_s), \tag{III.2}$$

where $(x_s, i, q_s, i)$ is the $i$-th tuple of the data sample and label of the source domain, $N_s$ represents the number of samples in the source domain, $R(.)$ is a regularization term controlled by the weight $\alpha$, and $f_s$ is a function that lies in a Hilbert space $\mathcal{H}$. $f_s$ is optimized by means of the loss function $l_s$ using the data from the source domain $D_s$.

The presented benchmark is carried out using the CVIQ [90] and OIQA [108] databases described in Sec.II.2. Samples from each one are provided in Fig. III.4

### III.2.2 Content-based splitting strategy

Machine learning-based IQA tasks are typically learning a predictive model from quality assessment databases. When training data-driven models, one must ensure the accuracy, representativity, and reliability of the databases. Data biases are a major



**Figure III.4:** Samples from the used databases: (top) CVIQ and (bottom) OIQA.

issue for learning-based IQA that is often overlooked. The consequences of such an issue are significant. It implies that, regardless of the used model, any computational prediction would have the same biases as the training data. Furthermore, the performance of a trained model is reported only on the testing set in which the selection may induce biases related to the content. A popular and straightforward approach is to split the training and testing sets based on pristine images. This means that the model is evaluated on unseen content independently of the existing distortions in the database. However, the obtained sets may lack diversity in terms of spatial complexity and colourfulness and may induce representativity biases, resulting in a test set that is not illustrative of the used database. Biases are mostly present, whether the data is split arbitrarily or based on more qualified criteria. However, minimizing those biases guarantees a validation on representative sets of the trained model.

For this benchmark, we first tackle the issue of content induced bias. To minimize such a bias, we use spatial information (SI) and colorfulness information (CFI), described in the Chapter II, as criteria for the splitting strategy to make sure that, the performance of the models are reported on a limited-bias set of images. Fig. III.5 shows SI versus CFI plots of pristine images on CVIQ and OIQA databases. As it can be seen, the variability of SI is higher in OIQA than in CVIQ, indicating that the latter database lacks diversity of content in terms of spatial complexity in comparison to OIQA. A similar conclusion holds in the case of CFI.

To select the training/testing sets, we used the Euclidean distance. For a couple of pristine images $I_1$ and $I_2$ characterized by $(CFI_{I_1}, SI_{I_1})$ and $(CFI_{I_2}, SI_{I_2})$ respectively, the distance $D(I_1, I_2)$ is expressed as follows:

$$D(I_1, I_2) = \sqrt{(CFI_{I_1} - CFI_{I_2})^2 + (SI_{I_1} - SI_{I_2})^2} \tag{III.3}$$

Based on the previous observation, we intend to demonstrate the existence of data biases when testing the effectiveness of a trained deep learning model on a given set from a database. The selection of the testing set influences the prediction correlation independently of the used database. Three splitting strategy are compared in this work. The first one is a random splitting by taking 20% on each iteration. The second one splits the databases in such a way that the images in the testing set are clustered in terms of SI/CFI (will be referred to as SI/CFI (a)). The third strategy takes the

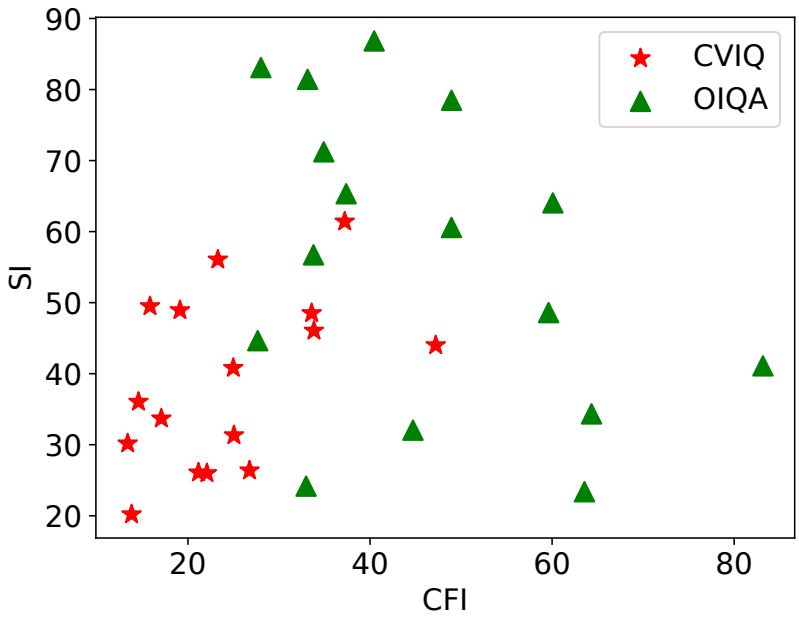



**Figure III.5:** Spatial information (SI) / colorfulness information (CFI) plot of pristine images in CVIQ and OIQA databases.

images that are the most spread-out in terms of SI/CFI (will be referred to as SI/CFI (b)). For all strategies, we ensure a complete separation of the training and testing sets, *i.e.*, the distorted images linked to the same pristine image are allocated to the same set.

### III.2.3 Projection-based training

Within this configuration, we first investigate the use of ERP images as inputs to the selected models. It is rather straightforward and aims at evaluating CNN models on high-resolution ERP images. The input ERP are down-sampled into a resolution of $1024 \times 512$. This implies an adaptation of the model in order to match the shape of the input images. The output feature maps are provided to the quality regression block described in Sec. III.2.1. The use of ERPs as direct input may be thought of as estimating global quality rather than local to specific regions on the scene [91]. Despite the geometric distortions occurring on this type of projection, investigating the effect of using high resolution content with CNN models seems appropriate. Also, the models will learn from additionally distorted content (*i.e.* distortion from the databases as well as the projection-induced ones). Providing an analysis regarding the impact of the latter is within the scope of this benchmark. We will refer to this configuration as $C_{ERP}$.

In addition to the use of ERPs, we intend to provide a performance analysis on the use of cube-map projection (CMP). The CMP introduces less distortions compared to

ERP. However, it provides separate content in form of cube faces. In fact, this projection requires a re-projection from ERP to CMP. It uses the six faces of a cube as the projection shape. The CMP is generated by first rendering the scene six times from a viewpoint. So, from each ERP image $I$, six faces are obtained $\{Left_I, Front_I, Right_I, Back_I, Top_I, Bottom_I\}$. An illustration of the re-projection is provided in Fig. III.6.

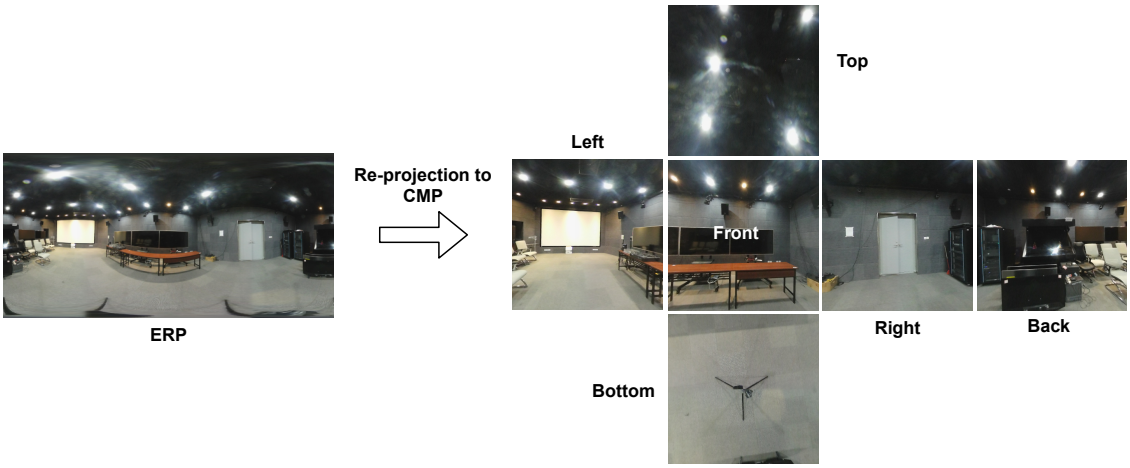



**Figure III.6:** ERP to CMP re-projection resulting in six faces: left, front, right, back, top and bottom.

One way to deal with the CMP as input to CNN models is to build a multichannel CNN as introduced in [90]. This way implies multiple CNNs in parallel where each is fed with one of the six obtained faces $\{Left, Front, Right, Back, Top, Bottom\}$. The output feature maps from these channels are concatenated, regressed and used to derive a quality score. The optimization of the model as well as the prediction is made on the six channels simultaneously and not individually. The use of CMP under a multichannel paradigm will be referred to as $C_{CMP}$ and the adopted architecture is depicted in Fig. III.8. A different way consists of taking each face as a separate content which involves a patch-based training scheme. Details on this approach are provided in Sec. III.2.5.

### III.2.4 Radial-based training

As mentioned previously, the viewing experience of 360-degree images is quite different from traditional ones. A user can only see the actual rendered FoV from the spherical representation. The next rendered FoV (viewport) is determined by his head movement around the x, y, and z axes. A slight head rotation will change the rendered viewport. The most important part of the actual viewed viewport is the content surrounding its center. Therefore, we only consider this latter to predict the quality on. To avoid any confusions, we will not call it viewport as it only represents a portion of it and, most of the time, this region is extracted as a square shape as in [90, 91,

131]. Indeed, a viewport is not square and using this term to describe square patches or regions could be misleading. As a result, we will refer to it as region.

By focusing on possible regions to predict the quality of 360-degree images, we seek an agreement with the viewing experience of this kind of images. Also, in this case, geometric distortions created by the previously described sphere to plane projection, will be avoided. Another avoided type of distortion, is content discontinuity, artificial borders and oversampling created by the CMP projection [132]. This can lead to a loss of the semantic content. One solution to avoid such unwanted results, is the use of the radial content rather than the projected one. It can be done by mapping the ERP content to the sphere (*i.e.* from plane to 3D space). Then projecting back the viewed content, which consists of important regions from possible viewports, to 2D representation (*see* Fig. III.7).

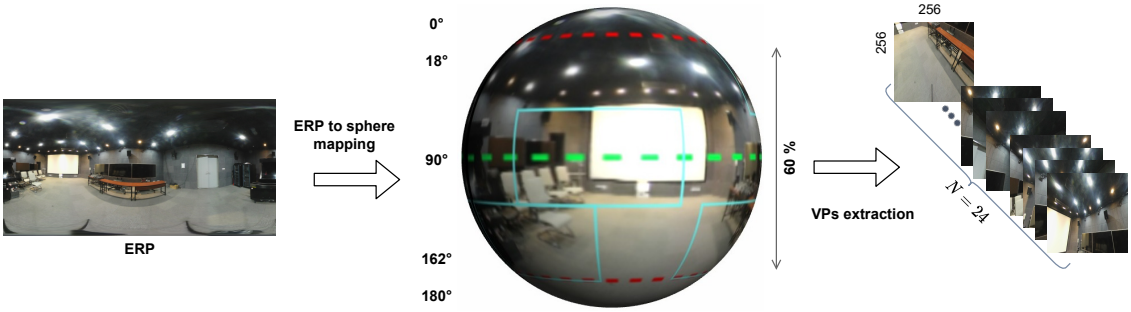



**Figure III.7:** Viewport selection for the spherical content configuration. Blue areas represent the selected viewports. In total, 24 regions surrounding the equatorial line (From 18° to 162°) are extracted from the spherical content.

In addition to the used exploration behavior described above, it is now admitted that the human gaze is biased towards the equatorial line when viewing 360-degree images [133]. Inspired by this and the fact that more than 30% of the content is often not viewed [134], we generate viewports surrounding the equatorial line representing more than 60% of the input content. Each center of a possible viewport $R_i$ from the possible candidates $k$ (up to $k = 24$) is extracted and projected from the spherical representation to the 2D plane. Then, the extracted contents are used as an input of a pre-trained model (among the seven selected networks). Similarly to $C_{CMP}$, this configuration implies a multichannel paradigm. The number of parallel channels depends on the number of extracted content. Accordingly, the complexity at this stage is proportionally increased. The output feature maps generated by the different channels are concatenated before feeding them to the quality regression block described in Sec. III.2.1. The training and prediction flow is depicted in Fig. III.8. We will refer to this as $C_{Radial}$ in the remaining of the paper.

For this configuration, we first train the models with eights inputs before increasing their number by 8 until 24. This involves expanding the architecture of the models by adding more channels to fit the additional inputs. Such a strategy is motivated by

the intent to analyze the impact of increasing inputs for a multichannel paradigm by finding the trade-off between accuracy of the models and the induced complexity.

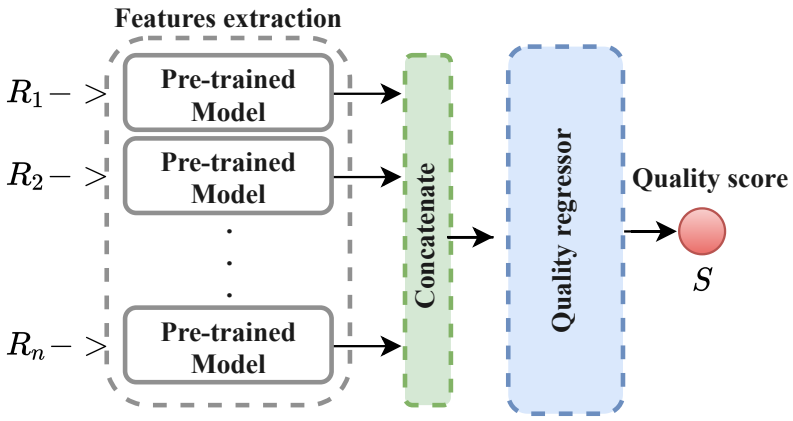



**Figure III.8:** Architecture of the multichannel CNN. $Ri$ with $i \in \{1, 2, .., n\}$ stand for the extracted regions. Architecture adopted for $C_{CMP}$ ($n = 6$) and $C_{Radial}$ ($n \in \{8, 16, 24\}$).

### III.2.5 Patch-based training

Differently from $C_{CMP}$ and $C_{Radial}$, this configuration adopts a patch-based learning scheme. This means all considered regions from the 360-degree images are seen as individual content, necessitating distinct labelling. Unfortunately, ground truth label (MOS) for individual patches are unavailable, since only the 360-degree image-level ground truth MOS is provided. This heavily increases the challenge of IQA when adopting a patch-based training. A straightforward solution is to assign the same MOS of the 360-degree image to the derived patches. This was first introduced in [135] and adopted by other researchers in [136, 137].

Within this configuration, two different approaches to extract patches from 360-degree images are used. First, the regions extracted for $C_{Radial}$ are considered as individual patches, 24 from each image. Second, the six faces from $C_{CMP}$ where each face is taken as a separate patch. This configuration involves the use of a single channel CNN rather than a multichannel one. An overview of this configuration is provided in Fig. III.9. By using different approaches to extract patches, we aim to provide a better understanding on how the extraction method can influence the training and the prediction accuracy of the selected models. The quality score of the entire 360-degree images is obtained by an average pooling of patches scores belonging to the same image. We will refer to this configuration as $C_{Patches}$.

### III.2.6 Training on 2D IQA databases

The lack of databases for IQA of 360-degree images hinders the promotion and development of CNN-based IQA models. In fact, designing a deep neural network and

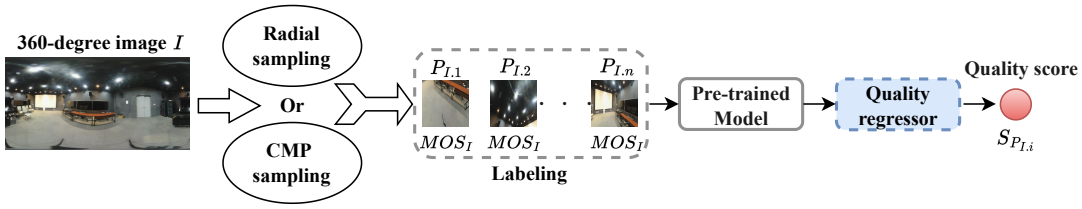



**Figure III.9:** Architecture of the $C_{Patches}$. $P_{I.i}$ with $i \in \{1, 2, 3, ..., n\}$, $n = 24$ for the radial sampling and $n = 6$ for the CMP sampling. $S_{P_{I.i}}$ represents the predicted quality score of patch $i$ from the 360-degree image $I$.

training it requires large-scale and representative databases. This is the main reason for adopting fine-tuning pre-trained models. Yet, pre-trained models have their limits, specially when used for a different task that may require specific type of ground truth. IQA is one of the most sensitive image processing tasks. The state-of-the-art for 2D IQA is well-developed compared to the 360-degree one. Exploiting what exists may benefit to 360-IQA. One of the aspects that we can exploit is 2D benchmark databases to train CNN models for IQA. Hence, the models weights will be optimized according to this specific task, from earlier layers to the top layers. A similar approach was used in [91] where they trained ResNet-18 [87] on the LIVE [138] database to further improve its accuracy. Other databases can be exploited to compensate for the lack of available data, such as the categorical image quality (CSIQ) database [139] and Tampere Image Database (TID2013) [130]. Other databases may also be useful.

In this study, two strategies of training on 2D-IQA databases are investigated. The first consists of training the selected models separately on each database. Each model is then trained from scratch using the ground truth provided by LIVE, CSIQ, and TID2013. The obtained knowledge is then transfered to 360-IQA by fine-tuning the obtained weights on CVIQ and OIQA. This way, all models are trained and fine-tuned for the same exact task. The second strategy consists of combining 2D-IQA databases in a large training dataset. This strategy is inspired by the work proposed in [140]. As combining IQA databases is rather difficult, requiring additional subjective experiments to ascertain the homogeneity of ratings according to the levels of degradation, the authors proposed a smart approach by using image pairing based on the Thurstone model. The ground truth labels are computed as the probability $P_{(x,y)}$ of the quality of $x$ being higher than $y$, *i.e.* quality ranking task rather than visual quality prediction. By doing so, a large scale training dataset can be obtained. However, it requires a Siamese network with $x$ and $y$ as inputs and $P_{(x,y)}$ as outputs. The reader should refer to the original paper [140] for more details. After training the model on the combined dataset, the weights are saved and used to perform transfer learning on CVIQ/OIQA.

For training on 2D databases, we unfreeze the trainable layers of the used models. The new weights are optimized according to the regression of extracted features to visual quality scores for the first strategy, and quality rankings for the second one. The $C_{Patches}$ under $CMP$ is used to perform the fine-tuning. For LIVE and TID2013,

**Table III.2:** Characteristics of the used 2D IQA databases.

| Databases | LIVE [138] | CSIQ [139] | TID2013 [130] |
|---|---|---|---|
| # of pristine images | 29 | 30 | 25 |
| # of distorted images | 779 | 866 | 3000 |
| Distortion types | JP2K, JPEG, WN, GB, FF. | JP2K, JPEG, WN, GB, FF, Contrast. | JP2K, JPEG, WN, GB and others. |
| Levels of distortions | 5 | 6 | 5 |

we cropped the regions surrounding the center with a resolution of $256 \times 256$ for all images, as they contain heterogeneous resolutions or rectangular shapes. This way, we avoid altering the content due to inappropriate resampling. Additionally, the input images are not normalized, which enables the proposed method to also cope with distortions introduced by luminance and contrast changes [136]. For both training starategies, all models are trained for 300 epochs and early stopping by monitoring the validation loss. We will refer to this configuration as $C_{2D}$ in the following.

For the end-to-end training and transfer learning of all configurations, the error between predicted and target scores is computed using the $L_2$ loss function between the ground truth label $y$ and the predicted one $\hat{y}$ over the batch of size $N$.

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})^2 \tag{III.4}$$

## III.3 Results and discussion

### III.3.1 Experimental setup

The proposed benchmark is implemented using TensorFlow [141]. The training of the considered configurations is performed on a server equipped with an Intel Xeon Silver 4208 2.1GHz CPU, 192GB of RAM, and an Nvidia Telsa V100S 32GB GPU. We use the RMSProp [142] optimizer for training the models. The learning rate is set to 0.001 with exponential decay. All models are trained with a batch size of 8 according to [143] for 25 epochs. We set the input dimension of all models to $(256, 256, 3)$ for the $C_{CMP}$, $C_{Radial}$, $C_{Patches}$ and $C_{2D}$. As for the $C_{ERP}$, we set it to $(1024, 512, 3)$.

The databases are split using the well-known Pareto principle and the criterion discussed in Sec. III.2.2, 80% is dedicated for training, and the remaining 20% for testing. For the sake of a fair comparison, all configurations are trained/tested using the same splitting scheme. Five-fold cross-validation is used for a complete evaluation within the selected database.

### III.3.2 Data biases evaluation

To analyze the performance of splitting strategies and demonstrate the influence of content-induced biases, we compared three schemes as discussed in Sec. III.2.2. We trained the selected CNN models on ERP images for this assessment, since both databases come with this format. The performance results in terms of correlation accuracy (PLCC) and monotonicity (SRCC) are summarized in Table III.3 for both databases. The mean values of five-folds are given to provide a complete and fair assessment.

As can be observed, each splitting scheme resulted in a different performance, regardless of the database. This actively demonstrates the existence of biases when splitting databases for training and testing. Besides, it shows the impact of the used strategy on the reported validation performances. Since we are dealing with IQA which is a delicate task compared to classification or object detection for instance, one should consider the selection of a representative set of data for testing the efficiency of CNN models. Content representativeness for IQA may be expressed by a variety of attributes such as those we used, *i.e.* spatial complexity and colorfulness.

**Table III.3:** Performance evaluation of the splitting strategies on CVIQ/OIQA databases. The best performing models are highlighted in **bold** for rows and underlined for columns. (a) and (b) stands for the SI/CFI-based schemes

| | | CVIQ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Splitting | Metric | ResNet-50 | ResNet-34 | RNet-18 | Vgg-16 | Vgg-19 | DeneNet-121 | Inception-V3 |
| Rand | PLCC | 0.900 | 0.733 | 0.799 | 0.813 | 0.772 | **0.903** | 0.809 |
| | SRCC | 0.831 | 0.672 | 0.729 | 0.734 | 0.667 | **0.832** | 0.740 |
| SI/CFI (a) | PLCC | 0.844 | 0.727 | 0.787 | 0.789 | 0.791 | **0.862** | 0.810 |
| | SRCC | 0.770 | 0.639 | 0.728 | 0.734 | 0.696 | **0.782** | 0.707 |
| SI/CFI (b) | PLCC | 0.837 | 0.725 | 0.800 | 0.803 | 0.735 | **0.858** | 0.719 |
| | SRCC | 0.774 | 0.634 | 0.705 | 0.691 | 0.639 | **0.767** | 0.622 |
| | | OIQA | | | | | | |
| Rand | PLCC | **0.749** | 0.590 | 0.565 | 0.586 | 0.534 | 0.725 | 0.732 |
| | SRCC | **0.710** | 0.512 | 0.505 | 0.568 | 0.505 | 0.686 | 0.696 |
| SI/CFI (a) | PLCC | **0.837** | 0.641 | 0.803 | 0.662 | 0.608 | 0.818 | 0.803 |
| | SRCC | **0.826** | 0.624 | 0.771 | 0.610 | 0.592 | 0.783 | 0.775 |
| SI/CFI (b) | PLCC | **0.899** | 0.695 | 0.779 | 0.613 | 0.694 | 0.860 | 0.790 |
| | SRCC | **0.877** | 0.666 | 0.762 | 0.576 | 0.665 | 0.829 | 0.764 |

From Table III.3, the random splitting scheme resulted in the best performance in terms of PLCC and SRCC for CVIQ. However, the observation is different for OIQA where the random splitting performance is outperformed by both SI/CFI based splitting schemes for all models. A possible reason could be related to the content composing each database, *i.e.* diversity of the images in terms of visual content and spatial complexity of the scenes. Based on this assumption and the previously discussed ob-

servation about the distribution of the characteristics regarding CVIQ images (*see* Fig. III.5), one can conclude that OIQA is more diverse than CVIQ. For the SI/CFI based splitting strategies, the (b) resulted in a more representative set since it selects diverse content and, intuitively represents approximately the used database in terms of content. In addition, it allows a more reliable evaluation of the performance accuracy within databases.

As observed for CVIQ, the random splitting scheme resulted in the best performance overall, except for ResNet-18 and Vgg-19. We believe that it is strongly related to the nature and diversity of the content. Additionally, between SI/CFI-based strategies, a very slight difference can be observed for CVIQ as they appear to be competing with each other. The opposite is observed on OIQA, where a noticeable difference can be reported in terms of correlation and monotonicity. On the same database, the SI/CFI (b) resulted in a better performance compared to the (a) strategy. Based on the above observations, we adopt the SI/CFI (b) strategy to train/test the considered configurations. By doing so, the performances will be reported on the most representative sets of the selected databases.

### III.3.3 Projection-based evaluation

#### III.3.3.1 $C_{ERP}$

To assess the performances of selected pre-trained models on high-resolution ERP images, we provide in Table III.3 (SI/CF based splitting (b)) the PLCC and SRCC scores obtained for both databases. Knowing that no omnidirectional peculiarities have been considered with this configuration, its performances are still satisfactory for almost all models. On average, the best performing model within this strategy is ResNet-50 followed by DenseNet-121, while the least performing one is Vgg-19. This is valid for both accuracy (PLCC) and monotonicity (SRCC) of the predictions. In fact, ResNet-50 obtained a PLCC (resp. SRCC) value of 0.844 (resp. 0.770) on CVIQ and 0.899 (resp. 0.877) on OIQA. DenseNet-121 achieved 0.862 (resp. 0.782) on CVIQ and 0.860 (resp. 0.829) on OIQA. These two models outperformed the other CNN models, regardless the used database. ResNet-50 is more popular compared to DenseNet-121, especially within the IQA community. DenseNet-121 model is underrepresented for IQA tasks, and most of the recent works adopted either ResNet or Vgg [101, 122, 144] as backbones. These choices are often made based on previous conclusions derived from other image processing tasks.

Comparing the results on CVIQ and OIQA, one can notice better correlations on OIQA, supporting the previous assumption on the nature of content in this database. The diversity of content helps models to better train and generalize. However, with the $C_{ERP}$, only 422 images are used for fine-tuning on CVIQ and 256 on OIQA. The more diverse the training data, the more examples to train on are required. Therefore, one can conclude that achieving a siginificant generalization ability on diversified databases requires larger training sets.

**Table III.4:** Performance evaluation of cross-database validation under the $C_{ERP}$. Best performing models in **bold**.

| Train/Test | Dist. | Metric | ResNet-50 | ResNet-34 | ResNet-18 | Vgg-16 | Vgg-19 | DenseN-121 | Inception-V3 |
|---|---|---|---|---|---|---|---|---|---|
| OIQA / CVIQ | Overall | PLCC | **0.820** | 0.403 | 0.410 | 0.309 | 0.485 | 0.750 | 0.687 |
| | | SRCC | **0.751** | 0.305 | 0.437 | 0.254 | 0.485 | 0.716 | 0.651 |
| | JPEG | PLCC | **0.903** | 0.339 | 0.360 | 0.436 | 0.556 | 0.813 | 0.761 |
| | | SRCC | **0.751** | 0.250 | 0.325 | 0.331 | 0.540 | 0.717 | 0.644 |
| | AVC | PLCC | **0.811** | 0.434 | 0.447 | 0.320 | 0.521 | 0.695 | 0.681 |
| | | SRCC | **0.769** | 0.396 | 0.423 | 0.314 | 0.497 | 0.672 | 0.653 |
| | HEVC | PLCC | **0.741** | 0.432 | 0.604 | 0.215 | 0.385 | 0.731 | 0.633 |
| | | SRCC | 0.700 | 0.401 | 0.588 | 0.200 | 0.374 | **0.713** | 0.618 |
| CVIQ / OIQA | Overall | PLCC | **0.476** | 0.256 | 0.268 | 0.295 | 0.320 | 0.472 | 0.474 |
| | | SRCC | **0.433** | 0.279 | 0.256 | 0.304 | 0.302 | 0.386 | 0.431 |
| | JPEG | PLCC | **0.768** | 0.264 | 0.404 | 0.324 | 0.281 | 0.754 | 0.285 |
| | | SRCC | **0.762** | 0.326 | 0.351 | 0.345 | 0.178 | 0.732 | 0.278 |

We conducted a cross-database validation to provide a better understanding of the use of ERP images with CNN models. We first trained the models on OIQA before testing their performance on CVIQ and *vice versa*. The performance results are summarized in Table III.4 in terms of PLCC and SRCC. We provide results for both the overall databases and on individual distortions. For training on CVIQ and testing on OIQA, we only provide results on the JPEG distortion according to [90].

Despite the satisfactory results obtained by the selected models on each database separately, one can observe very low performances on the cross-database validation. This depicts the limitation of $C_{ERP}$ when used with different CNN architectures, except for ResNet-50 and DenseNet-121. The latter models achieved good results in both cases. Training on OIQA and testing on CVIQ gave better results compared to the reverse case. One can observe a PLCC (resp. SRCC) value of 0.820 (resp. 0.751) on the overall database obtained by ResNet-50 when trained/tested on OIQA/CVIQ compared to 0.476 (resp. 0.433) when trained/tested on CVIQ/OIQA. A similar behavior is noticed with DenseNet-121 and the other models. This could be explained by the heterogeneousness of the distortions in OIQA, combining compression artifacts with Gaussian blur and white noise. Testing the performances of fine-tuned CNN models, primarily trained for classification on unseen distortions, resulted in poor performances. Among the models, the performances of ResNet-50 show a significant difference compared to ResNet-18/34 and Vgg-16-19. A possible explanation could be in the fine-tuning strategy [145]. It is known that the hyperparameters are key factors in achieving the best performance. These parameters are usually tuned according to the model, its architecture and depth, and the training datasets. However, as the focus of the study is rather benchmarking omnidirectional related configurations, the used hyperparameters are fixed for all models.

When comparing the performance on individual distortions, it can be seen that training on OIQA yields better results. Even though JPEG is present in both databases, training on OIQA resulted in significantly higher PLCC and SRCC scores. This finding holds for all seven models. A possible explanation is that the levels of JPEG distortion applied in both databases are different (five in OIQA and eleven in CVIQ). Still, the same class of artifacts should not result in such a significant difference. Perhaps compressing with eleven levels is not the best option because it results in less discernible differences between some stimulus (impaired images). When tested on CVIQ, Resnet-50 and DenseNet-121 also performed well regarding AVC and HEVC. These distortions are not available in OIQA, demonstrating the efficacy of these models in generalizing to comparable distortions.

#### III.3.3.2 $C_{CMP}$

With the intent to provide a comparison of pre-trained models' performances when used on CMP projection format and assess the influence of this type of projection, we provide in Table III.5 results in terms of PLCC and SRCC. Overall, the prediction performances are more correlated on CVIQ compared to OIQA. This is because CVIQ contains only compression artifacts, while OIQA contains various ones. The diversity of distortions may lead to a less generalized correlation across the entire database. This observation is applicable irrespective of the architecture of the model.

**Table III.5:** Performance evaluation in terms of PLCC and SRCC of pre-trained models using $C_{CMP}$. Best performances are highlighted in **bold** for each database.

| Database | Metric | ResNet-50 | ResNet-34 | ResNet-18 | Vgg-16 | Vgg-19 | DenseNet-121 | Inception-V3 |
|---|---|---|---|---|---|---|---|---|
| CVIQ | PLCC | **0.835** | 0.751 | 0.786 | 0.743 | 0.776 | 0.825 | 0.739 |
| | SRCC | **0.814** | 0.657 | 0.760 | 0.726 | 0.714 | 0.730 | 0.653 |
| OIQA | PLCC | **0.775** | 0.562 | 0.583 | 0.493 | 0.493 | 0.673 | 0.607 |
| | SRCC | **0.722** | 0.532 | 0.561 | 0.498 | 0.498 | 0.596 | 0.548 |

From Table III.5, it can be noticed that ResNet-50 outperforms the other models in terms of prediction accuracy and monotonicity on both databases. DenseNet-121 ranked second, but as it has fewer parameters compared to ResNet-50 (*see* Table III.1). Its performance can be considered as a trade-off between accuracy and complexity. One can also observe that Vgg-16 and Vgg-19 performed the worst among the seven models on OIQA. It is also the case of Inception-V3 on CVIQ.

A cross-database assessment was performed using CVIQ and OIQA to demonstrate the generalization ability of selected pre-trained models under the CMP configuration. Firstly, we trained the models on OIQA and tested them on CVIQ. The performance results on the overall database as well as per distortion types are provided in Table III.6. As it can be seen, the performances on the overall database are below 0.7, except for ResNet-50 and DenseNet-121 when tested on JPEG, achieving second-best performance. Is it worth mentioning that, none of the models were dedicated

**Table III.6:** Performance evaluation of cross database validation under the $C_{CMP}$. Best performing models in **bold**.

| Train/Test | Dist. | Metric | ResNet-50 | ResNet-34 | ResNet-18 | Vgg-16 | Vgg-19 | DenseNet-121 | Inception-V3 |
|---|---|---|---|---|---|---|---|---|---|
| OIQA / CVIQ | All | PLCC | **0.804** | 0.429 | 0.598 | 0.400 | 0.106 | 0.697 | 0.374 |
| | | SRCC | **0.738** | 0.308 | 0.566 | 0.389 | 0.080 | 0.625 | 0.394 |
| | JPEG | PLCC | **0.914** | 0.525 | 0.649 | 0.381 | 0.206 | 0.867 | 0.617 |
| | | SRCC | **0.819** | 0.318 | 0.469 | 0.283 | 0.208 | 0.752 | 0.472 |
| | AVC | PLCC | **0.743** | 0.464 | 0.680 | 0.552 | 0.188 | 0.598 | 0.349 |
| | | SRCC | **0.705** | 0.371 | 0.641 | 0.539 | 0.127 | 0.551 | 0.407 |
| | HEVC | PLCC | **0.703** | 0.382 | 0.690 | 0.442 | 0.314 | 0.506 | 0.385 |
| | | SRCC | **0.647** | 0.247 | 0.673 | 0.448 | 0.256 | 0.498 | 0.376 |
| CVIQ / OIQA | All | PLCC | 0.304 | 0.252 | 0.308 | 0.211 | 0.172 | **0.487** | 0.227 |
| | | SRCC | 0.287 | 0.261 | 0.306 | 0.149 | 0.158 | **0.431** | 0.228 |
| | JPEG | PLCC | 0.506 | 0.227 | 0.405 | 0.409 | 0.254 | **0.687** | 0.484 |
| | | SRCC | 0.470 | 0.233 | 0.346 | 0.407 | 0.250 | **0.649** | 0.388 |

to quality assessment as they were trained on ImageNet. Only the regression block is trained for the IQA task. Besides, the only common distortion between OIQA and CVIQ is JPEG. This is reflected in the same table, where an improvement of PLCC and SRCC for JPEG could be observed compared to the overall performance. The correlation performances shifted from 0.80 to 0.91 for ResNet-50, and from 0.69 to 0.86 for DenseNet-121. The performances of the other models improved as well, but remains below the 0.7 threshold. Regarding AVC and HEVC distortions, the performances dropped compared to JPEG and even to the overall scores, yet still acceptable.

Then, we trained on CVIQ and tested on OIQA. The correlation results are summarized in the lower part of Table III.6. One can observe low performances compared to previous results. The training on CVIQ seems to lead to less generalize models. The overall performances are very low as the models are trying to predict on unlearned distortions (*i.e.* WGN and GB). Besides, the performances on JPEG are low too compared to those obtained when trained on OIQA. Despite the low performances, the contrast to training on OIQA regarding the best-performing model can be noticed. The DenseNet-121 outperformed the other models, even ResNet-50.

### III.3.4 Radial-based evaluation

In this section, we discuss the performance evaluation of the radial content-based configuration $C_{Radial}$. Table. III.7 gathers the scores for CVIQ and OIQA. The previous observation regarding the best-performing models is still valid for this configuration. Overall, ResNet-50 and DenseNet-121 performed the best, with DenseNet-121 ranking first on CVIQ and ResNet-50 on OIQA.

Overall, one can notice that the performances obtained on CVIQ are mostly better compared to OIQA. A minimum PLCC (resp. SRCC) value of 0.72 (resp. 0.69) is obtained on CVIQ, while 0.39 (resp. 0.36) on OIQA. The reason might be the distortions

**Table III.7:** Performance evaluation of pre-trained models with the $C_{Radial}$ on CVIQ/OIQA databases in terms of PLCC/SRCC. Best performing model is highlighted in **bold** for CVIQ and underlined for OIQA.

| # inputs | Metric | ResNet-50 | | ResNet-34 | | ResNet-18 | | Vgg-16 | | Vgg-19 | | DenseNet-121 | | Inception-V3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CVIQ | OIQA | CVIQ | OIQA | CVIQ | OIQA | CVIQ | OIQA | CVIQ | OIQA | CVIQ | OIQA | CVIQ | OIQA |
| R = 8 | PLCC | 0.788 | 0.795 | 0.720 | 0.606 | 0.829 | 0.688 | 0.801 | 0.387 | 0.801 | 0.567 | **0.841** | 0.772 | 0.757 | 0.704 |
| | SRCC | 0.713 | 0.758 | 0.698 | 0.559 | **0.794** | 0.661 | 0.707 | 0.361 | 0.707 | 0.517 | 0.747 | 0.727 | 0.740 | 0.644 |
| R = 16 | PLCC | 0.807 | 0.842 | 0.770 | 0.544 | 0.751 | 0.725 | 0.772 | 0.482 | 0.809 | 0.482 | **0.857** | 0.822 | 0.793 | 0.791 |
| | SRCC | 0.747 | 0.816 | 0.689 | 0.530 | 0.722 | 0.705 | 0.700 | 0.508 | 0.716 | 0.508 | **0.769** | 0.780 | 0.755 | 0.752 |
| R = 24 | PLCC | 0.830 | 0.851 | 0.726 | 0.663 | 0.764 | 0.769 | 0.802 | 0.394 | 0.821 | 0.608 | **0.859** | 0.890 | 0.775 | 0.749 |
| | SRCC | 0.781 | 0.809 | 0.687 | 0.629 | 0.740 | 0.740 | 0.747 | 0.389 | 0.743 | 0.623 | **0.782** | 0.876 | 0.722 | 0.723 |

contained in OIQA. With this configuration, additional distortions due to projection can be avoided. Therefore, the reported results are more representative as they were obtained based on the actual viewed content.

The fact of increasing the number of inputs improved the performances of the selected models. One can notice that, in average, the performance increases with increased inputs for all models, and declines with $R = 24$ for Resnet-34 and Inception-V3. This behavior does not apply to ResNet-18, as we notice (the best score with $R = 8$) and then a decrease with additional inputs. This actively demonstrates an overfitting behavior. A similar behavior is shown by Vgg-16. Increasing the number of inputs leads to a higher number of channels where CNN models become more prone to overfitting. It is worth mentioning that the variation of the number of regions results in a variation of the quality prediction accuracy as well as the prediction monotonicity.

### III.3.5 Patch-based evaluation

To compare the performance of the multichannel paradigm versus the patch-wise training scheme, we trained the selected models using the output of the CMP as patches in addition to the regions generated for $C_{Radial}$ (*see* Sec. III.2.5). Table III.8 summarizes the obtained results.

In average, the radial-based method performed better. A PLCC (resp. SRCC) value of 0.821 (resp. 0.760) on CVIQ and 0.778 (resp. 0.756) on OIQA compared to 0.800 (resp. 0.751) and 0.708 (resp. 0.668) with CMP on CVIQ and OIQA respectively. A PLCC difference of approximately 2.6% on CVIQ and 9.4% on OIQA is observed when using patches obtained on the sphere. This illustrates the usefulness of using radial content against the projected one. Another possible reason is the number of extracted patches, providing the models with more training examples. Looking into individual performances, DenseNet-121 ranked the best for radial patches and ResNet-50 for CMP patches. Despite the heterogeneity of the distortions on OIQA compared to CVIQ, training the DenseNet-121 using a patch-wise lead to a better accuracy. Another noteworthy observation is related to the Vgg-16/19 performances. They achieve comparable performance to ResNet-50 for radial configuration on CVIQ. Knowing that the

pre-trained version of Vgg-16/19 scored among the worst in the previous configurations, their performances under $C_{Patches}$ prove to be satisfactory.

**Table III.8:** Performance evaluation of pre-trained models with the $C_{Patches}$ on CVIQ/OIQA database in terms of PLCC, SRCC. Best performances are highlighted in **bold** for columns and underlined for rows.

| Input type | | Radial | | CMP | |
|---|---|---|---|---|---|
| Metric | | PLCC | SRCC | PLCC | SRCC |
| ResNet-50 | CVIQ | 0.861 | 0.820 | **0.857** | **0.833** |
| | OIQA | 0.836 | 0.810 | **0.867** | **0.848** |
| ResNet-34 | CVIQ | 0.604 | 0.533 | 0.756 | 0.725 |
| | OIQA | 0.671 | 0.626 | 0.678 | 0.667 |
| ResNet-18 | CVIQ | 0.792 | 0.716 | 0.827 | 0.783 |
| | OIQA | 0.787 | 0.774 | 0.755 | 0.713 |
| Vgg-16 | CVIQ | 0.861 | 0.816 | 0.795 | 0.711 |
| | OIQA | 0.816 | 0.787 | 0.557 | 0.465 |
| Vgg-19 | CVIQ | 0.859 | 0.791 | 0.785 | 0.709 |
| | OIQA | 0.771 | 0.734 | 0.552 | 0.518 |
| DenseNet-121 | CVIQ | **0.907** | **0.851** | 0.807 | 0.768 |
| | OIQA | **0.925** | **0.917** | 0.845 | 0.818 |
| Inception-V3 | CVIQ | 0.864 | 0.792 | 0.772 | 0.726 |
| | OIQA | 0.641 | 0.640 | 0.701 | 0.645 |
| Average | CVIQ | 0.821 | 0.760 | 0.800 | 0.751 |
| | OIQA | 0.778 | 0.756 | 0.708 | 0.668 |

An in-depth analysis shows a significant difference in terms of performances among different models. This could be related to the pre-trained version of the models. Performing transfer learning with various amount of training examples is affecting the deeper and shallower models in different ways. For instance, deeper models such as DensNet-121 and Vgg-16/19 achieved good performances with radial compared to CMP. Whereas, with ResNet-18/34/50 the reverse can be observed. The models are fine-tuned using augmented databases of varying sizes. The radial configuration generates 10128 patches on CVIQ (resp. 6144 on OIQA) while the CMP configuration generates 2532 (resp. 1536 patches). Four times less the amount of training data is generated with CMP compared to the radial starategy with an impact on the model's achievable performances. This configuration and the training strategy appear to influence different backbones in different ways.

We performed a cross-database validation with the patch-wise configuration to verify the generalization ability of the selected models when trained using a patch-wise scheme. The performance results are gathered in Table III.9. The same observation regarding the best performing model still valid, ResNet-50 and DenseNet-121 achieved the best performance overall and per-distortion. Good results are obtained

**Table III.9:** Performance evaluation of pre-trained models with the $C_{Patches}$ on CVIQ/OIQA database in terms of PLCC, SRCC. Best performances are highlighted in **bold** for rows and underlined for columns.

| Training / Testing | Distortion | Metric | ResNet-50 | | ResNet-34 | | ResNet-18 | | Vgg-16 | | Vgg-19 | | DenseNet-121 | | Inception-V3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Radial | CMP | Radial | CMP | Radial | CMP | Radial | CMP | Radial | CMP | Radial | CMP | Radial | CMP |
| Train: OIQA & Test: CVIQ | Overall | PLCC | **0.886** | **0.843** | 0.705 | 0.593 | 0.637 | 0.607 | 0.789 | 0.710 | 0.767 | 0.665 | 0.859 | 0.841 | 0.764 | 0.684 |
| | | SRCC | **0.846** | 0.790 | 0.672 | 0.552 | 0.560 | 0.547 | 0.734 | 0.682 | 0.711 | 0.652 | 0.810 | **0.791** | 0.721 | 0.631 |
| | JPEG | PLCC | **0.948** | 0.928 | 0.834 | 0.709 | 0.735 | 0.755 | 0.929 | 0.854 | 0.905 | 0.786 | 0.945 | **0.929** | 0.901 | 0.846 |
| | | SRCC | **0.865** | **0.835** | 0.721 | 0.567 | 0.569 | 0.535 | 0.816 | 0.761 | 0.795 | 0.719 | 0.845 | 0.817 | 0.819 | 0.737 |
| | AVC | PLCC | **0.846** | **0.782** | 0.722 | 0.656 | 0.620 | 0.678 | 0.726 | 0.715 | 0.645 | 0.657 | 0.800 | 0.775 | 0.676 | 0.625 |
| | | SRCC | **0.842** | **0.768** | 0.705 | 0.622 | 0.599 | 0.616 | 0.704 | 0.691 | 0.642 | 0.645 | 0.780 | 0.754 | 0.651 | 0.586 |
| | HEVC | PLCC | **0.814** | 0.721 | 0.645 | 0.570 | 0.546 | 0.640 | 0.632 | 0.599 | 0.649 | 0.648 | 0.749 | **0.733** | 0.648 | 0.524 |
| | | SRCC | **0.804** | 0.716 | 0.636 | 0.521 | 0.522 | 0.575 | 0.627 | 0.579 | 0.646 | 0.659 | 0.730 | **0.728** | 0.625 | 0.509 |
| Train: CVIQ & Test: OIQA | Overall | PLCC | 0.558 | 0.441 | 0.336 | 0.453 | 0.469 | 0.455 | 0.318 | 0.330 | 0.436 | 0.309 | **0.570** | **0.598** | 0.469 | 0.497 |
| | | SRCC | **0.534** | 0.430 | 0.309 | 0.453 | 0.468 | 0.439 | 0.287 | 0.320 | 0.370 | 0.270 | 0.523 | **0.564** | 0.438 | 0.419 |
| | JPEG | PLCC | **0.858** | 0.631 | 0.484 | 0.363 | 0.720 | 0.492 | 0.637 | 0.587 | 0.709 | 0.631 | 0.855 | **0.825** | 0.759 | 0.700 |
| | | SRCC | 0.790 | 0.554 | 0.414 | 0.297 | 0.679 | 0.423 | 0.579 | 0.533 | 0.664 | 0.554 | **0.801** | **0.723** | 0.724 | 0.607 |

on the JPEG distortion with a PLCC (resp. SRCC) values of 0.94 (resp. 0.86) using radial patches, and 0.92 (resp. 0.83) using CMP patches by ResNet-50 when trained on OIQA and tested on CVIQ. Parallelly, satisfying results are obtained on AVC and HEVC distortions. Vgg-16/19 and Inception-V3 achieved satisfactory results with radial when trained on OIQA. PLCC/SRCC values above 0.90/0.80 on JPEG are obtained. One can also observe that the models trained on OIQA demonstrate a stronger generalization ability when tested on CVIQ compared to the opposite. This supports the previous observation concerning the richness of OIQA versus CVIQ in terms of content and distortions. Besides, the radial-based method resulted in the best performance compared to CMP regardless of the used database. This depicts the importance of using radial rather than projected content on the one hand. On the other hand, generating more examples for the models to train on, improved the accuracy, demonstrating the impact of having a large amount of data.

### III.3.6 Training on 2D IQA databases evaluation

**Table III.10:** Performance evaluation of $C_{2D}$ in terms of PLCC, SRCC. The best performing models are highlighted in **bold** for columns and underlined for rows on each dataset. 'All' stands for combined datasets.

| 2D database | | LIVE | | CSIQ | | TID2013 | | ALL | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC |
| ResNet-50 | CVIQ | 0.915 | 0.852 | 0.912 | 0.849 | 0.909 | 0.847 | **0.938** | **0.895** |
| | OIQA | 0.884 | 0.873 | 0.876 | 0.865 | 0.879 | 0.863 | 0.882 | 0.862 |
| ResNet-34 | CVIQ | 0.910 | 0.847 | 0.934 | 0.887 | 0.918 | 0.859 | 0.728 | 0.669 |
| | OIQA | 0.920 | 0.907 | 0.923 | 0.914 | **0.905** | **0.889** | 0.381 | 0.371 |
| ResNet-18 | CVIQ | 0.898 | 0.836 | 0.903 | 0.839 | 0.898 | 0.829 | 0.821 | 0.774 |
| | OIQA | **0.923** | **0.912** | 0.919 | 0.910 | 0.894 | 0.878 | 0.342 | 0.322 |
| Vgg-16 | CVIQ | 0.877 | 0.804 | 0.873 | 0.810 | 0.849 | 0.768 | 0.889 | 0.832 |
| | OIQA | 0.610 | 0.587 | 0.832 | 0.813 | 0.816 | 0.799 | 0.804 | 0.773 |
| Vgg-19 | CVIQ | 0.861 | 0.791 | 0.892 | 0.823 | 0.813 | 0.737 | 0.886 | 0.841 |
| | OIQA | 0.805 | 0.782 | 0.847 | 0.833 | 0.721 | 0.704 | 0.692 | 0.679 |
| DenseNet-121 | CVIQ | **0.948** | **0.918** | **0.943** | **0.906** | 0.906 | 0.842 | 0.869 | 0.838 |
| | OIQA | 0.837 | 0.827 | **0.931** | **0.921** | 0.880 | 0.858 | **0.908** | **0.917** |
| Inception-V3 | CVIQ | 0.897 | 0.832 | 0.908 | 0.886 | **0.923** | **0.875** | 0.751 | 0.735 |
| | OIQA | 0.905 | 0.893 | 0.911 | 0.898 | 0.853 | 0.833 | 0.696 | 0.834 |

With the intent to evaluate whether the use of 2D IQA databases improves the performance of CNN models compared to performing transfer learning, we trained all selected models on LIVE, CSIQ, TID2013, and combined databases (All). As it is

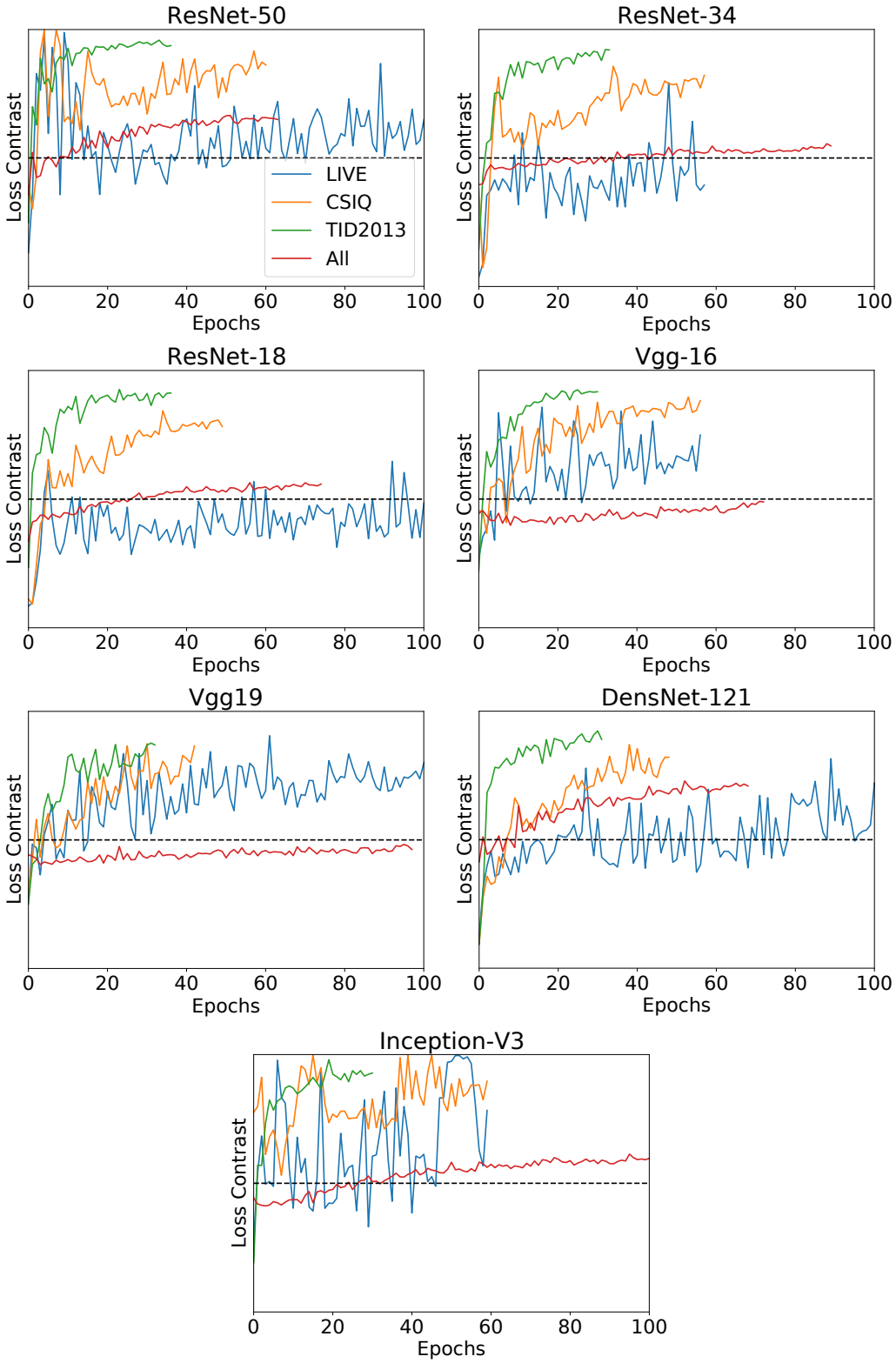



**Figure III.10:** Contrast $(val_loss - loss)/(val_loss + loss)$ between training and validation losses for all models trained on 2D-IQA databases ($0 \rightarrow$ equal loss between training and validation losses). 'All' stands for combined datasets.

known, deep neural networks require large-scale databases in order to achieve better accuracy as well as avoiding overfitting. To analyze the learning behavior, we provide in Fig. III.10 the contrast ($val_loss - loss / val_loss + loss$) between training and validation losses for the five folds (F-1 to F-5). A contrast equal to 0 depicts an equal loss between training and validation, whereas a contrast equal or close to 1 suggests an important gap between both losses, with $val_loss$ being higher and the opposite if equal or close to $-1$. We can see that training on LIVE is leading to a better generalization to the prediction of MOS, but with a non-smooth behavior. Training on the combined datasets (All) led to the best behavior of the training losses, as the contrast is stable and close to 0. This is a generalization to predict the probability ranking, as discussed in Sec. III.2.6, suggesting a robust performance during training. Training on TID2013 has a higher contrast, meaning that the models have difficulty to generalize. The gap between training and validation losses is much higher than those of LIVE and CSIQ. This is also demonstrated by the provided curves (*see* Fig. III.10). A possible reason is that TID2013 contains many diverse distortions, a total of 24 types, and it may need more examples to learn from in order to demonstrate a better generalization ability. From the provided curve, we can notice that the progress of training/validation loss is more stable on TID2013, despite the previous observation. This led to a quicker convergence for all models. Indeed, training on TID2013 required fewer epochs when compared to training on CSIQ, LIVE, and the combined datasets. Among the models, ResNet-34 converges quicker on each database, followed by Vgg-16.

In addition to the training behavior shown above, we provide the performance accuracy of the weights obtained from training on 2D databases, *i.e.* LIVE, CSIQ, and TID2013 individually and combined together. Table III.10 gathers the performance results on CVIQ/OIQA in terms of PLCC/SRCC. On average, the performances are quite satisfactory on both databases. A maximum PLCC (resp. SRCC) value of 0.948 (resp. 0.921) is achieved by DenseNet-121. Overall, the latter scored the best among the selected models when trained on each database separately and ResNet-50 when trained on the combined datasets. Training the models on 2D IQA databases appears to improve their performances in both correlation accuracy and monotonicity. The achieved efficiency is competitive except for Vgg-16 on OIQA when trained on LIVE. Despite the small size of IQA databases, the obtained performances actively demonstrate the usefulness of training CNN models on them. Acquiring knowledge about quality after being pre-trained to predict it is increasing the performances.

Among the selected 2D databases, training on TID2013 results in a poor performance compared to LIVE and CSIQ. It could be explained by the lack of generalization due to the limited number of instances per distortion. Indeed, insufficient amount of data may lead to overfitting, especially when training from scratch. When trained on LIVE and TID2013, ResNet-50 is outperformed by its smaller variants, ResNet-18/34. The difference is greater on OIQA than on CVIQ, which solely contains compression artifacts. It is known that deeper models require large databases in order to reach a generalization ability and sufficient accuracy, especially on databases with diverse content.

Regarding pre-training on the combined datasets, some improvement can be observed when compared to the use of imageNet weights reported in Table III.8. For example, ResNet-50 performances in terms of PLCC/SRCC shifted from 0.857/0.833 on CVIQ to 0.938/0.895, representing 9%/7.2% of improvement. On the same database, similar improvements can be seen with other models such as Vgg-16/19, as well as slight improvements with DenseNet-121, Inception-V3, and ResNet-18. Analyzing the performance on OIQA, ResNet-18/34 performed poorly, with accuracy and monotonicity scores below 0.5. As demonstrated by the lower performances, fine-tuning on databases with diverse content and degradation is less efficient for different models with varying depths. This indicates less generalization compared to training on individual databases.

### III.3.7 Computational complexity

With the aim to compare the computational complexity of the selected models under different configurations, we measure the required prediction time per input image. Since the inference analysis is independent from the training, we used a different hardware configuration. A computer equipped with an Intel® Core™ i9-9880H @ 2.30GHz, 32GB of RAM, and an Nvidia Quadro T2000 MAX-Q 4GB GPU is selected to measure the computational complexity. Fig. III.11 represents the average of the computational time required over ten images. Overall, DenseNet-121 requires the longest time, followed by Vgg-16/19. Considering this, one can conclude that DenseNet-121 and the Vgg-based models are heavier in terms of computational complexity, followed by Inception-V3, and finally ResNets. The training time is definitely not proportional to the complexity of the used model in terms of number of parameters (*see* Table III.1). DenseNets concatenations require high GPU memory and therefore more training time [146], while ResNet models implement skip-connections, allowing to jump over some layers and reducing the computational time [87, 147]. Despite the number of parameters of ResNet-50, the latter spent less time than VGG-16/19. DenseNet-121 has fewer parameters among the selected models, and yet it requires more time than ResNets.

Among the configurations, the multichannel appears to demand more computational time, except with Vgg-16/19. The computational time required by Vggs is highly impacted by the input shape. As it can be seen, Vgg-16/19 required the longest time when used with ERP images, suggesting that the architecture of the model plays a major role in the computational complexity.

In addition to the computational time, we measured the number of floating-point operations (FLOPs) with regard to the input shape. The latter determines the number of FLOPs providing insight on the computations required by the model. A large FLOPs number implies a higher complexity, suggesting a longer calculation time. The FLOPs are reported in Table III.11. One can observe that having an input shape of 1024*512*3 resulted in larger FLOPs. However, according to the computational time associated with each configuration, the first observation that emerges is that the

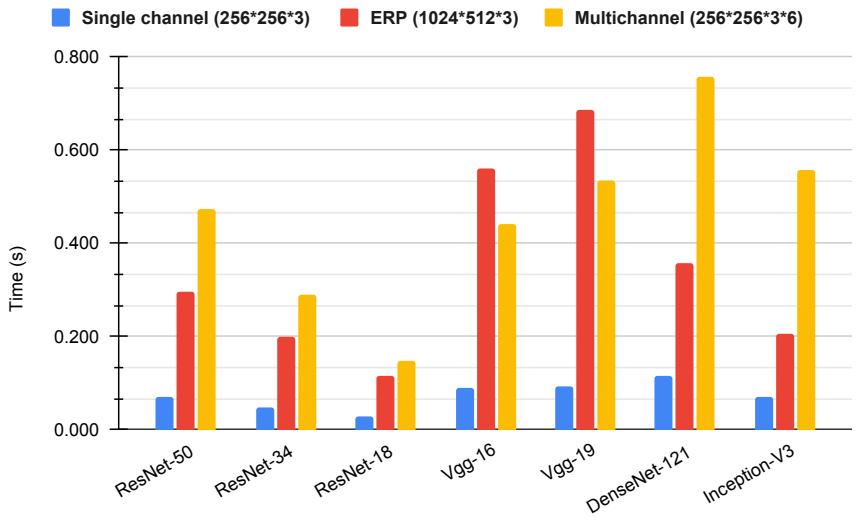



**Figure III.11:** Computational complexity in terms of required prediction time per image.The average over ten samples is provided.

**Table III.11:** The number of FLOPs with regard to the input shapes.

| Input shape | 256*256*3 | 1024*512*3 | 256*256*3*6 |
|---|---|---|---|
| ResNet-50 | 5.04 | 40.35 | 30.27 |
| ResNet-34 | 4.80 | 38.42 | 28.82 |
| ResNet-18 | 2.39 | 19.09 | 14.32 |
| Vgg-16 | 20.1 | 160.5 | 120.3 |
| Vgg-19 | 25.5 | 203.9 | 152.9 |
| DenseNet-121 | 3.70 | 29.61 | 22.21 |
| Inception-V3 | 3.97 | 36.23 | 23.14 |

FLOPs is not proportional with the required computational time. This could be explained by the fact that other operations are involved, especially memory-based ones, as discussed previously regarding DeseNet-121. In addition, some architectures implement skip-connections such as ResNets, allowing a more optimized utilization of the computational resources. Vgg-16/19 have the largest FLOPs independently of the used configuration. In contrast to the other models, both Vggs required significantly higher computational time. This confirms that the computational time is strongly affected by the architecture.

### III.3.8 Overall performance evaluation

The training and validation of a CNN model are often made on randomly selected sets. The adopted splitting scheme may result in biased sets. Limiting these biases helps to improve the reliability of the models as well as the reported performance. With this idea in mind, we first conducted a comparison of three splitting strategies (*see*

Sec. III.2.2). The results demonstrated the existence of content-induced biases, as the performances were different for each splitting strategy. In addition, it showed the difference in terms of content diversity in the available omnidirectional IQA databases. CVIQ appears to contain less diversity compared to OIQA. In our case, in order to provide accurate and reliable results, we adopted for all configurations, the splitting scheme that uses scene complexity and colorfulness as splitting criteria.

Predicting visual quality on projected content (*i.e.* ERP and CMP images) for omnidirectional IQA is straightforward and does not require an additional pre-processing step for extracting viewports or patches. However, the achieved performances are quite poor, except for ResNet-50 and DenseNet-121. This observation is confirmed when conducting cross-database validation, in which the performances decreased substantially. The limitations of using projected content, as well as the limitations of CVIQ in offering enhanced generalization ability, were demonstrated. When we trained the models on OIQA and tested on CVIQ, the results were better than when the reverse was performed. Because OIQA comprises a variety of distortions, it benefited the models in achieving higher correlation accuracy and monotonicity when compared to CVIQ, which solely incorporates compression artifacts.

The use of radial-content (*i.e.* spherical content) helps to mimic the exploration behavior of users, predicts visual quality on the actual viewed content, and avoids geometric distortions due to projection. This approach results in a set of regions for quality predictions. The challenge is to determine the number of these regions as well as their locations. Regarding the latter, it is preferable to focus on the equator, as demonstrated in [133]. We utilized 8, 16, and 24 extracted regions to investigate the influence of their number on the performance of the selected models. An improvement was observed with the increase of number of inputs, showing the importance of feeding CNN models with more content to learn from. An overall improvement was also observed regarding all models compared to the use of projected content. Except for Vgg-16/19 on OIQA where a low performance is observed. The performance of DenseNet-121 (resp. ResNet-50) stood out from the rest of the models on CVIQ (resp. OIQA).

Training a CNN model on selected regions from an omnidirectional image either implies the use of a multichannel CNN or a patch-wise training. The multichannel CNN learns from multiple inputs that are linked to a single ground truth (*i.e.* MOS), while patch-wise learning involves labeling each extracted patch independently. The multichannel strategy is investigated with the CMP- and Radial content-based configuration. For the patch-wise, two techniques were evaluated, the use of radial content and faces from the cube-map projection as patches. Overall, the superiority of using the radial content was observed. With this configuration, the DenseNet-121 and ResNet-50 still outperform the other models. The cross-database evaluation supports the idea of using radial content as well as generating larger training sets. Good performances were obtained when we trained the models on OIQA and tested on CVIQ, especially for JPEG distortion. Despite the difference in the used levels of JPEG, five on OIQA and eleven on CVIQ, PLCC/SRCC values above 0.90/0.80 were achieved.

Except for training on 2D databases, ResNet-50 and DenseNet-121 performed the best across all tested configurations. This actively demonstrates the effectiveness of these models for IQA tasks when used with the ImageNet weights. When trained on 2D databases, ResNet-18/34 and Inception-V3 achieved competitive performances noticeably better compared to those obtained with their original weights. This shows that deeper models need large databases, while less deep ones may achieve high accuracy with fewer data. In addition, actual 2D IQA databases are limited in comparison to ImageNet. Building IQA databases is time-consuming, which is why transfer learning is usually adopted; tiny databases would not allow CNN models to reach a substantial degree of accuracy. However, when we trained the selected models on LIVE, CSIQ, and TID2013 databases, we could observe an improvement over the pre-trained versions. Overall, the best performances were achieved when trained on LIVE and CSIQ. These databases share four distortion types with OIQA. In terms of loss contrast and training convergence, we discovered that the broader the database (*i.e.* various distortion type), the faster the model trains and less contrast it obtains (*see* Fig. III.10). This may be due to fewer examples to learn from when an important number of distortion is used. When we trained on the combined datasets, some improvement were observed, especially with CVIQ, while less generalization is achieved when fine-tuned on diverse database (OIQA).

## III.4 Summary

The main takeaways of this benchmark are:

- When training CNNs models for IQA, a complete separation of the training and testing sets should be performed. Otherwise, the validation would be biased as the model will have already seen the content. In addition, to avoid the representativeness bias a content-oriented splitting strategy should be considered.
- IQA datasets for training CNN models may suffer from diversity, either in terms of content or distortions. Consequently, the generalization ability and robustness of the model may be highly affected.
- The use of projected contents limits the achievable performances, especially the generalization ability. The fact that this content presents geometric distortions and less fidelity with the viewed content resulted in limited performances. In this case, the use of radial content could be more effective.
- Patch-based training is as efficient as multichannel models featuring several CNNs in parallel. With proper patches sampling and training strategy, the patch-based training should be considered since it drastically reduces the complexity. By doing so, the inference time is improved while maintaining promising accuracy.
- The design of multichannel models should properly consider the number of channels. The latter may influence the prediction performances in addition to being highly complex, leading to training difficulties.

- According to the experimental results, there is no linear relationship between the accuracy nor the monotonicity of the model and its complexity.
- Pre-training on 2D-IQA is helpful for increasing backbone performance over ImageNet weights. However, when databases are combined, some models perform poorly in terms of generalization, failing to account for the difference between pre-training and fine-tuning tasks.

## III.5 Conclusion

In this chapter, we explored the usage of well-known CNN models for IQA of 360-degree images. The reason for this choice is that these models were trained on large-scale databases, and transfer learning techniques may benefit IQA. We conducted an empirical and analytical evaluation by covering different CNN architectures, image representations, and training strategies to provide recommendations on the use of CNNs for 360-degree IQA. Seven pre-trained CNN models were fine-tuned and compared based on various configurations, including the use of projected and radial content, multichannel paradigm and patch-wise training, and retraining on well-known 2D IQA databases.

The obtained results showed the superiority of retraining CNN models on IQA databases over the use of ImageNet pre-trained versions. The use of radial content led to better performance and generalization ability compared to projected content, especially with the patch-wise training. Among the selected models, ResNet-50 and DenseNet-121 performed the best. We believe that this work sheds light on the usage of pre-trained CNN models for IQA and paves the way for further research. One critical factor is the scarcity of large-scale, accurate, and reliable 360-IQA databases. It can be viewed as the foundation of any quality assessment validation scenario, and such databases are in urgent need in order to promote the development of IQA models for such content.

# Chapter IV

# Pre- and Post-processing for CNN-based 360-IQA

## IV.1 Introduction

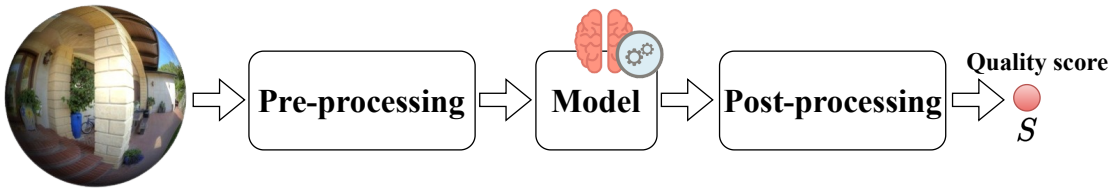



**Figure IV.1:** Steps in a typical deep-learning based 360-IQA framework.

360-IQA frameworks comprise various steps and processes, particularly the pre-processing, model's architecture design and training, and post-processing, as illustrated in Fig. IV.1. Each step contributes to the efficacy, reliability, and robustness of the IQA framework. Formally speaking, the inputs 360-degree image goes through a set of processes before being fed to the model to train on. As explained in Chapter III, the input image represented either in a projected format or as a set of viewports or patches, corresponding to specific regions from the image. The selection and generation of the latter must follows adaptive strategies in order to be consistent with IQA paradigms in general and 360-IQA in particular. Additionally, data representation and normalization retains highly influential information and aid the model to train faster. All these data preparation are paramount to achieve good and consistent results. Regarding the architecture design and training of the model, proper training strategies also boost the accuracy and adaptability of IQA models to the 360-IQA task. At the final stage, the post-processing focuses on deriving the final quality score from individual scores. This helps to enhance the correlation with the ground truth, *i.e.* MOS, and therefore improves the quality predictions.

With a focus on the pre- and post-processing steps for deep-learning based 360-IQA, the current chapter explores several contributions toward the design of predict-

ive and reliable 360-IQA models. Fig. IV.2 depicts the overview of the addressed issues in this chapter.

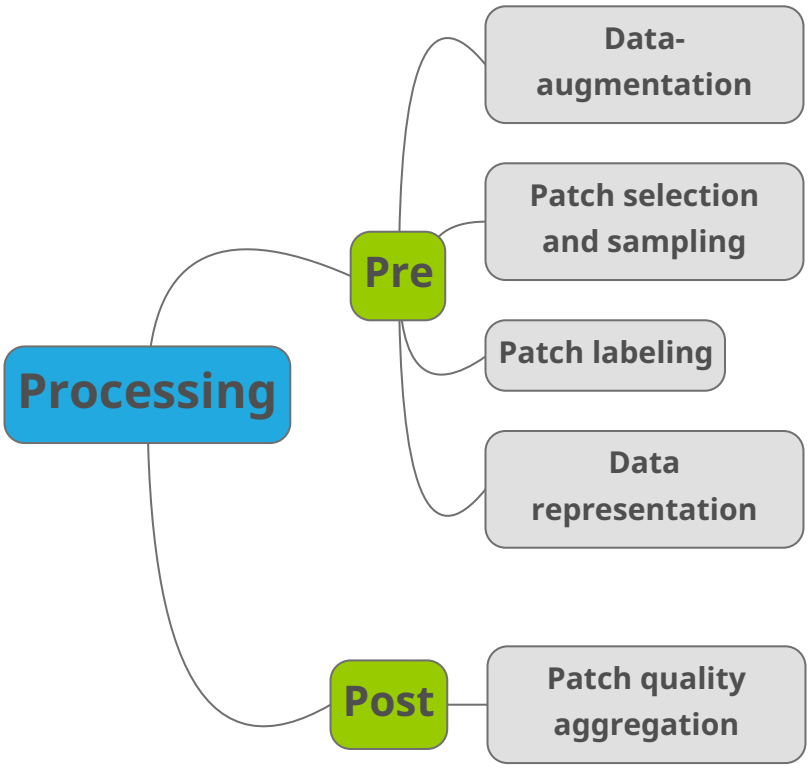



**Figure IV.2:** Overview of the processes addressed in this chapter.

## IV.2 Visual scanpath for patch-based 360-IQA

The multichannel paradigm is primarily used for 360-IQA based on deep learning as discussed in Sec. I.2.2.2. In contrast to multichannel models (*see.* Fig. IV.3 top), the patch-based ones take individual regions separately. Here, a single CNN is used (*see.* Fig. IV.3 bottom) which implies less complexity and leads to a faster training. In the literature, several works adopted the patch-based training for 2D IQA [136, 148–150], and good performances have been reported. The interest of such an approach lies in its proven performance in various image processing tasks, such image classification, object detection and recognition, etc. Also, the quality prediction tends to agree with the scene exploration by focusing on prominent parts of it that are translated into patches. However, the unavailability of MOSs for individual patches is considered as the main issue of this approach. Existing models label all patches extracted from the same image with the same MOS.

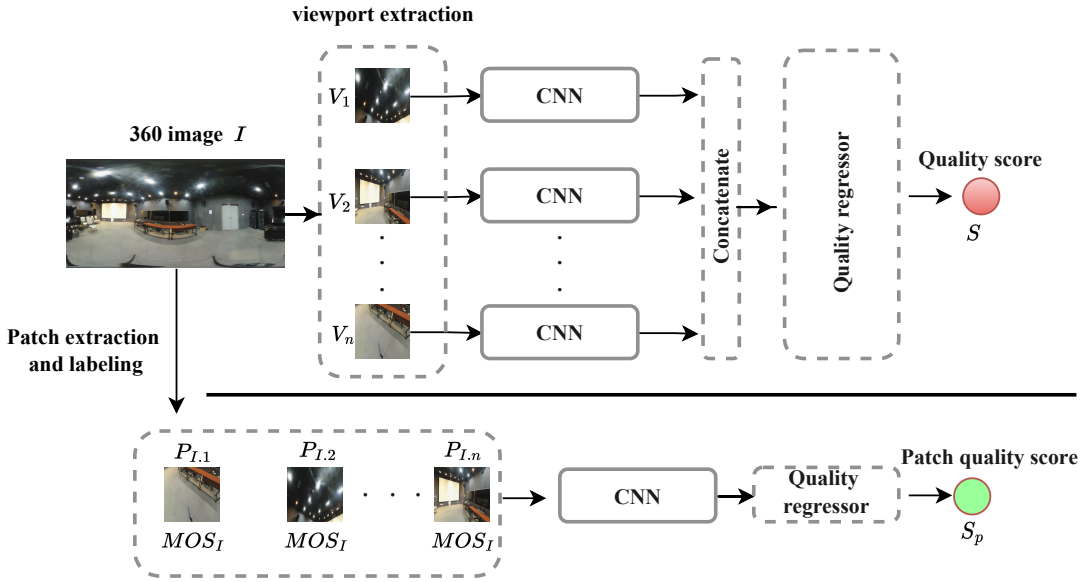



**Figure IV.3:** CNN models for IQA. (top) multichannel vs. (bottom) patch-based CNN.

In a patch-based IQA framework, two important aspects must be carefully considered. The first one corresponds to patches' selection and extraction. This is usually performed by using adaptive criteria such as visual saliency and scanpaths. In the extraction of these patches, the use of radial content (*i.e.* from the sphere) is highly recommended compared to the projected one [151, 152]. This way, the geometric distortion induced by the sphere-to-plane projection can be avoided. Fig. IV.4 depicts a 360-degree image viewing experience. We only consider chosen windows to predict the quality as it represents the way 360-degree images are generally viewed. The assumption that a person can only see the actual FoV from the spherical representation justifies this procedure. The next window is determined by his head movement around the x, y, and z axes. This way, the quality prediction scenario seeks to agree with the viewing experience of 360-degree images, and geometric distortions created by the previously described sphere to plane projection are avoided. The second aspect focuses on the aggregation of local qualities to a global quality score that should account for (i) the non-uniformity distribution of quality, and (ii) the variation among quality scores of individual patches.

The selection of prominent regions is performed using visual trajectory, *i.e.* scanpaths. The latter is used at two different stages. The first one consists of patches selection and data-augmentation. The second one consists of adaptive patch score aggregation. In the following we describe both use cases.

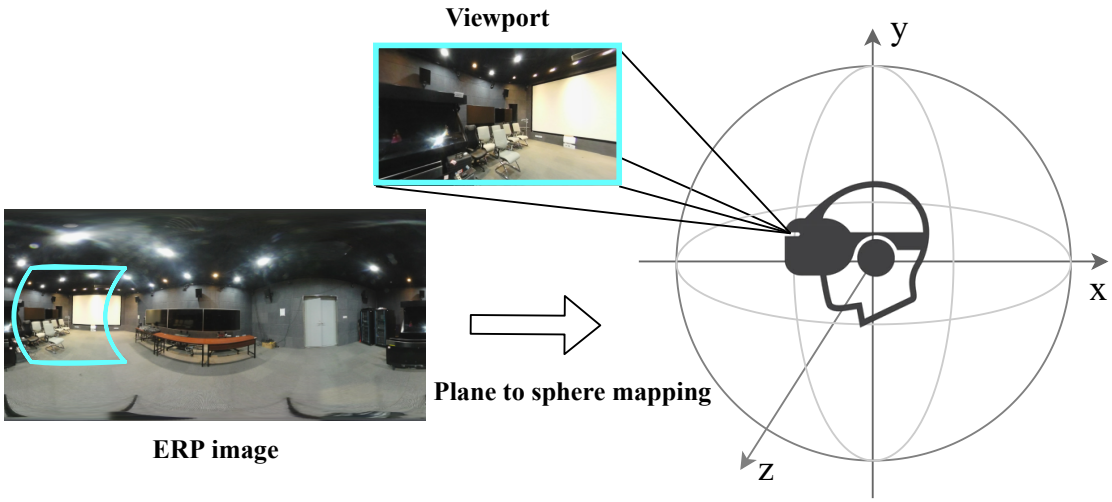



**Figure IV.4:** 360-degree images viewed using head-mounted displays. Blue area in the ERP represents the window extracted from the sphere.

### IV.2.1 Visual scanpath for data-augmentation

It is now widely admitted that when an image is viewed, the HVS gazes on salient details, which translates into eye fixations [153]. In our case, these regions are considered as relevant viewports and are detected using visual scan-path predictions model proposed in [154]. According to the use case and availability of the scanpath model, one may use other such as ScanGAN360 [155] and VPT360 [156], The selected model predicts a visual trajectory including eight relevant fixation points. In our model, ten trajectories are extracted representing ten virtual observers and used to account for the diversity of human scan-paths. These scan-paths are then considered for data augmentation, which results in a total of $N = 8 \times 10$ extracted viewports. This will help with the training of the model and avoid over-fitting caused by the lack of data. In fact, the efficiency of deep neural networks often increases as more data is available. Unfortunately, we still lack reliable and representative databases for 360-degree IQA that would allow deep learning models to assert their full capabilities. As stated in Chapter II, the construction of such databases requires important efforts in terms of scenes acquisition, device calibration, paradigm definition, subjective testing and data analysis [109]. Only two 360-degree image databases are being used to train and validate IQA models, namely CVIQ [157] and OIQA [108]. Consequently, the application of strategies to acquire more data with the existing one, is largely encouraged. The use of IQA-based data-augmentation is one option to accomplish this task.

Data augmentation is a method of creating new training data from existing one. This is accomplished by applying domain-specific (IQA in our case) strategies to elements from the training data in order to generate new and distinct training examples. Since IQA is more sensitive than other image processing tasks such as object detection

and classification, conventional approaches including shifting, rotating, flipping and brightness changing of an image, are counterproductive in our context. An illustration of standard techniques used in various image processing tasks is provided in Fig. IV.5. As it can be seen, the perceived content is altered and does not match the one viewed by the observers, making such techniques unusable for IQA. The particular reason for this, being that the images are labelled (rated) by human observers (MOS), and altering any visual attribute will make the actual rating incompatible. As a result, the use of data-augmentation techniques must be appropriate and concur with IQA. In the following, the viewport is referred to as a patch.




**Figure IV.5:** Illustration of standard data-augmentation [158].

The selection of the relevant regions using fixations from the scanpaths and extraction on the sphere is illustrated in Fig. IV.6. This way, we generate ten different instances of the database. To begin, the scan-path model is used to predict the ten VO potential trajectories and their gaze fixation positions. Then, rather than the projected format, each fixation point is located on the sphere, and the surrounding content is extracted and projected to a 2D plane. Following that, since we are using a patch-based learning scheme, each extracted region is fed to the model as a separate input. Previously mentioned, 360-degree images are rated based on multiply viewed regions. Giving the extracted patches the same MOS as their 360-degree image as firstly introduced in [135] for 2D content seems inefficient. Therefore, we applied well known and widely used 2D NR quality metrics to predict the quality of extracted patches, referred to in the following as the local quality. This is motivated by the fact that the extracted patches have a 2D representation and the model will consider them as sep-

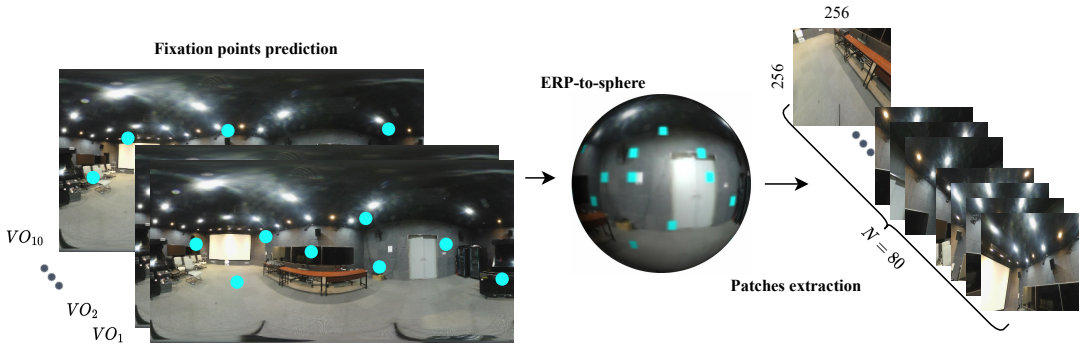



**Figure IV.6:** Patch selection using fixations from the scanpaths and extraction on the sphere.

arate images. Therefore, two labelling techniques are adopted including the MOS and 2D-IQA metrics namely BRISQUE [159] and NIQE [160]. This operation is described in Algorithm 1.

**Algorithm 1** Patches' labeling.

**procedure** LABELPATCH($P_i$, mode)
  $label_i = \emptyset$ ▷ Label to assign to $P_i$
  **if** mode = MOS **then**
    $label_i \leftarrow MOS_I$
  **else if** mode = NIQE **then**
    $label_i \leftarrow NIQE(P_i)$
  **else if** mode = BRISQUE **then**
    $label_i \leftarrow BRISQUE(P_i)$
  **end if**
  Return $label_i$
**end procedure**

## IV.2.2 Visual scanpath for patch qualities aggregation

Pooling strategies have been extensively investigated for 2D images [161–165]. Several strategies are considered, ranging from basic statistics and percentile pooling to content-based and information weighted spatial pooling. In [166], temporal pooling methods are compared for video quality assessment, where individual scores from different frames are pooled to a single quality score of the video. It is known that quality scores pooling is paramount in IQA frameworks, especially for patch-based CNNs.

It is true that patch-based CNNs have the potential to achieve robustness compared to multichannel ones. This is achievable with a proper data-augmentation technique and an adaptive patch-score aggregation strategy. Basically, the model is trained on

individual patches, meaning that the model only sees these patches, without having access to the whole 360-degree images. Therefore, $N$ scores associated to $N$ patches are predicted, and the mapping of these individual scores to a single quality score is important. This operation must be performed by adaptive aggregation to improve the correlation with the human judgment scores. Fig. IV.7 illustrates the mapping operation, where for each 360-degree image $I$, $N$ predicted scores $S = \{S_{p_1}, S_{p_2}, ..., S_{p_n}\}$ corresponding to $N$ patches $P = \{P_1, P_2, ..., P_n\}$ are aggregated together using the $Pooling(.)$ function.

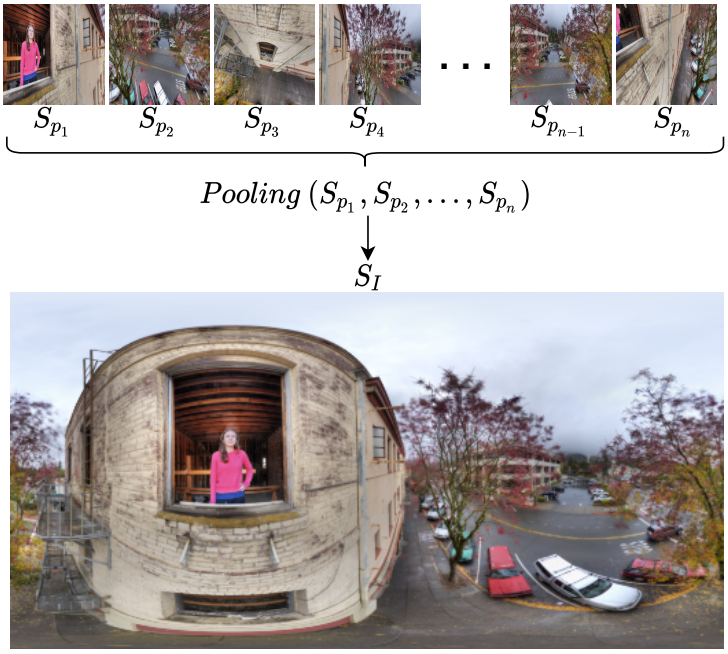



**Figure IV.7:** Mapping of predicted local qualities $S_{P_i}$ (per patch) to global quality $S_I$ (per 360-degree image).

For this specific task, one can find several methods used in the literature, mainly based on basic statistics. Table IV.1 describes existing methods.

It is known that aggregation strategies based on basic statistics tend to have poor correlations with subjective scores. It is especially the case of simple mean pooling that enforces an equal contribution of all patches to the global quality scores. By doing so, the non-uniformity distribution of quality is not taken into account. For this reason, a weighted mean pooling can reproduce this behavior by weighting each local score according to the importance of the patch's content. The estimation of these weights are usually based on perceptual properties such as visual attention [169], equator-bias [133] to incorporate the way the human gaze is biased toward the equator, making the computation of these weights handcrafted. Others opted for data-driven based estimation of the weights by adding subnetworks within a CNN model [136, 170].

Differently, we investigate a weighting strategy based on visual exploration. This is motivated by the fact that quality metrics are tuned and compared against the MOS

**Table IV.1:** Summary of basic statistic based aggregation methods.

| Method | Equation | Description |
|---|---|---|
| Arithmetic Mean | $S_I = \frac{1}{N}\sum_{i=1}^{N} S_{p_i}$ | The arithmetic mean is a straightforward method for pooling local qualities to a global one. By simply averaging the quality scores, the local qualities will contribute equally to the final score. |
| Harmonic Mean | $S_I = (\frac{1}{N}\sum_{i=1}^{N} S_{p_i}^{-1})^{-1}$ | The harmonic mean is one of the Pythagorean means. It is calculated by dividing the number of scores by the reciprocal of each score $S_{p_i}$ in $S$. Hence, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. It is known to emphasize the impact of small scores [167] reflecting the fact that subjective ratings are influenced by worst regions in terms of visual quality. |
| Geometric Mean | $S_I = (\prod_{i=1}^{N} S_{p_i})^{\frac{1}{N}}$ | The geometric mean is the third Pythagorean mean. It signifies the central tendency or typical values of $S$ by taking the root of the product of their values. |
| Five-Number Summary | $S_I = \frac{min + Q1 + median + Q3 + max}{5}$ | This method provides a description of $S$ using various descriptive statistics [168]. The five-number summary makes use of information on (i) the location given by the median, (ii) the spread of the scores given by the $Q1$ and $Q3$ quartiles representing the 25% and 75% percentile respectively, and (iii) the range of values expressed by the minimum and maximum of $S$. |
| Minkowski Mean | $S_I = (\frac{1}{N}\sum_{i=1}^{N} S_{p_i}^{p})^{\frac{1}{p}}$ | The Minkowski pooling has been widely used for IQA [42, 161]. The $P$ parameter emphasizes the lowest scores among $S$, *i.e.* the highly distorted patches. To understand the influence of the latter, we set it values to the most commonly used ones in the literature, including 1/4, 1/2, 2, 4, 8 and 16. |
| Percentile Pooling | $S_I = \frac{1}{\mid S \downarrow k\% \mid}\sum_{i \in S \downarrow k\%} S_{p_i}$ | The percentile pooling is considered as one of the most effective pooling methods. It is based on the fact that perceived quality is strongly affected by the most distorted regions [162]. This is accomplished by considering only the quality scores from $S$ that are lower than a $k$—th percentile. In order for us to study the impact of this latter, five percentiles are used as threshold including 5%, 10%, 20%, 25%, and 50%. |

collected by psychophysical experiments. By incorporating the way observers explore a scene before rating its quality could improve the pooling performance. Thus, an observer explores a visual scene by focusing on certain regions and usually not all parts of the scene. This behavior can be modeled using visual scanpaths, as explained previously. Ten scanpaths, composed of eights gaze fixations for each 360-degree image $I$ are all considered. Two important information associated with each fixation are considered as weights for each patch $P_i$ extracted from $I$. The first is the order of fixations, expressing the temporal progress of the visual trajectory. The second is the duration, representing the amount of time a region is likely to be focused on. The longer the gaze, the greater the influence on the observers' judgment. Finally, the pooling is performed as shown by Eq. IV.1, with $W_i$ is either the fixation duration or fixation order associated with patch $P_i$.

$$S_I = \frac{\sum_{i=1}^{N} W_i S_{p_i}}{\sum_{i=1}^{N} W_i}. \tag{IV.1}$$

Furthermore, to account for previous observations about the impact of most distorted regions on perceived quality, we combine the fixation-based pooling with the percentile threshold as given in Eq. IV.2.

$$S_I = \frac{\sum_{i \in |S \downarrow k\%|} W_i S_{p_i}}{\sum_{i \in |S \downarrow k\%|)} W_i}. \tag{IV.2}$$

To investigate the use of scanpath for 360-IQA, we used the ResNet-50 [87] as the base model to extract visual features from selected patches. We replace the top layers with a regression block in order to regress the learned features into a single quality score. ResNet employs residual learning to further deepen the CNN network, which can be interpreted by a number of deeper bottleneck architectures, as described in Sec. III.2.1.

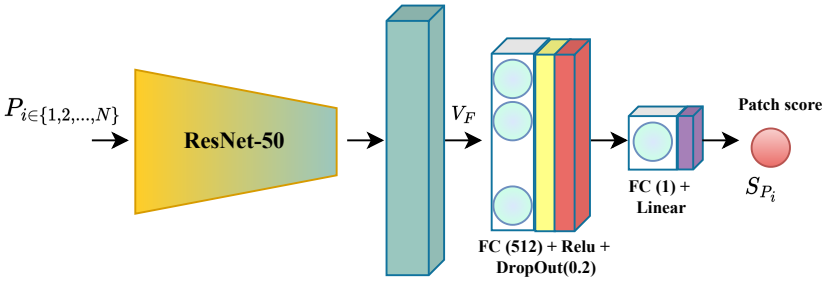



**Figure IV.8:** Architecture of the proposed model. Features are only extracted from individual patches $P_i$ by ResNet-50. $V_{\mathbf{F}} \in \mathbb{R}^{D\times1\times1}$ represents the extracted feature vector.

The output of ResNet-50 is fed to a quality regressor which is composed of a global average pooling so to reduce the spatial dimension of the extracted feature maps and help to minimize overfitting. Finally, two FC layers are then used to calculate

the quality score as depicted in Fig. IV.8. The weights for the quality regressor are initialized according to the method provided in [127]. For the end-to-end training, we used the $L_2$ loss function to compute the error between predicted and target scores.

## IV.3 Adaptive patch sampling

The use of visual trajectory as input sampling strategy is consistent with IQA paradigms. However, it is accomplished with a burden of using an effective scanpath prediction model. The latter are sometimes heavy in terms of computational time, on the one hand. On the other hand, the sampling performance is strongly affected by the accuracy of the used scanpath model. Therefore, it becomes urgent to investigate an adaptive and less complex sampling strategy. To do so, we propose an effective method based on (i) the latitude and (ii) content importance.

Most 360-IQA models are viewport-based [90, 100, 151], where the quality prediction is performed on specific regions. By doing so, an important part of the images is neglected. In the proposed strategy, we consider all the possible content from the sphere as important. However, the importance is varied according to the content's location on the sphere. The influence on the visual importance in implemented by considering different sampling ratios along the latitude.

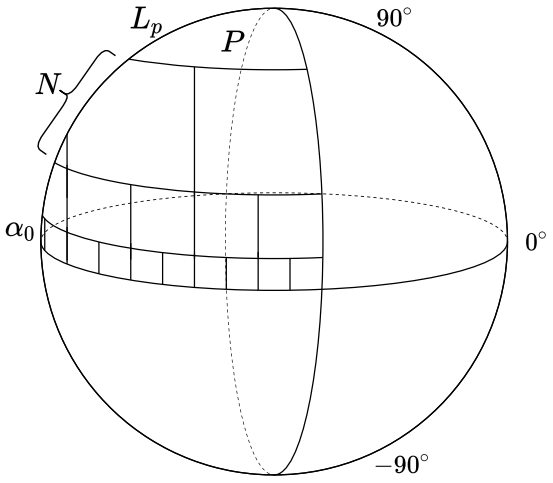



**Figure IV.9:** Latitude and content's importance based patches sampling from the sphere.

Let consider $r$ the position of latitude and longitude equal to 0. The hemisphere sampling starts by defining $\alpha_0$ the latitude of the initial square patch around the equator:

$$\exists\ \alpha_0 > 0\ :\ \frac{360}{\alpha_0} \in \mathbb{N}^+ and\ \frac{360}{\alpha_0.2^N} \in \mathbb{N}^+ \tag{IV.3}$$

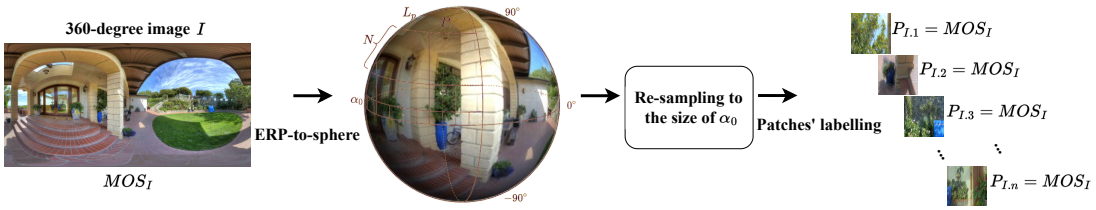



**Figure IV.10:** Overview of the adaptive sampling process on the sphere and patches' labelling.

where $N$ is the last level of sampling before the polar region $P$. The patches of the next level are of double size in latitude and longitude. The number of sampling levels is defined so as:

$$\exists\, N \geq 0 \ : (1 + \sum_{i=0}^{N} 2^i)\alpha_0 + L_p = 90 \ and \ L_p < \alpha_N \tag{IV.4}$$

where $L_p$ is the latitude of the polar region. Fig. IV.9 illustrates such an adopted sampling. All the extracted patches are re-sampled to the size of the equator patches *i.e.* $\alpha_0$. The resolution corresponding to $\alpha_0$ is set to 128 pixels. This operation is described by Algorithm 2.

**Algorithm 2** Patches' sampling.

**procedure** RESAMPLE($P_i$, $L_p$)
  $size_{\alpha_0} = 128px \times 128px$ ▷ Resolution corresponding to $\alpha_0$
  $L_{\alpha_0} = r$ ▷ Latitude position of $\alpha_0$
  **if** $L_p \neq L_{\alpha_0}$ **then**
    $size_{P_i} \leftarrow size_{\alpha_0}$
  **end if**
  Return $P_i$
**end procedure**

Therefore, $N_{patch} = 2 \times (4 + \sum_{i=0}^{N} \frac{360}{2^i.\alpha_0})$ non-overlapping patches are extracted from the sphere to avoid geometric distortions and provide the model with the actual viewed content [151]. An overview of the sampling process is provided in Fig. IV.10. Due to the unavailability of MOS per patches, the sampled ones from image $I$ are labelled using the MOS associated with $I$.

The efficiency of the proposed sampling strategy is verified under two configurations. We select two pre-trained CNNs, including the ResNet-50 [87] and EfficientNet-B3 [171] to this end. Both models are pre-trained on ImageNet. The choice of ResNet-50 and EfficientNet-B3 is motivated by their success and popularity with transfer

learning in various image processing tasks in general, and IQA in particular. We transfer the acquired knowledge from both models according to the previously described formulation in Sec. III.2.1

## IV.4 Input representation (normalization)

In general, input images for CNN models are pre-processed to ensure a better representation. The primary purpose of pre-processing is to improve image content by increasing specific visual components which contribute to the learning of the specified task [172]. Therefore, it is task-dependent, and it is usually referred to as normalization. Input data normalization prior to CNN models' training is greatly encouraged to help the model learn the useful information. In the case of IQA, normalization mainly consists of retaining high frequency information [169] over low frequencies, as the latter are less affected by distortions and are less perceivable by the HVS. In the literature, one can find several adopted normalization method for IQA frameworks. In particular, Kang *et al.* [135] used a method based on divisive normalization that was developed as a canonical computation implemented throughout the neocortex [173], and used to explain the response of neurons in the primary visual cortex. This method is called local contrast normalization and was adopted in several works afterward [174–176]. Kim *et al.* [169] used a simple low-frequency subtraction to only retain high-frequency components. Although input normalization has significant benefits, a part of the information is lost, especially luminance and contrast changes. The use of unnormalized images allows the model to learn from such additional information, as stated by Bosse *et al.* [136].

The contrasting opinions about the use or not of normalization prior to CNN training for IQA bring some confusion when developing a new framework based on CNNs. In addition, the variety of normalization methods used for IQA as well as for other image processing tasks, such as histogram equalization, zero component analysis whitening, difference of Gaussian, motivate the investigation of the usefulness of normalization. To this end, we analyze the influence of existing normalization methods on the performances of patch-based CNN IQA model.

The normalization methods found in the literature can be categorized into five categories: 1) basic scaling, 2) local normalization based methods, 3) difference based methods, 4) histogram based methods, and 5) whitening based methods.

In the following, we describe some of the important normalization methods within each category.

### IV.4.1 Basic Scaling

**Pixel Scaling** is a fundamental pre-processing step for deep learning algorithms. It consists of simply scale the pixel values of input images in a specific range, generally between 0 and 1. By doing so, luminance and contrast are not altered.

**Mean-Centering** consists of centering the distribution of pixel values on zero. This is performed by calculating the mean pixel value across the entire training dataset, then subtract it from each image. It is referred to as mean normalization where the input patch $P$ is converted to $P' = P - \mu$ where $\mu$ is the mean.

**Standardization** Standardization consists of mean centering patches followed by a division by their standard deviation (SD), making the mean and variance normalized and resulting in a zero-mean reduced Gaussian distribution of the training dataset.

### IV.4.2 Local Normalization based methods

**Local Contrast Normalization (LCN)** is used as a nonlinear preprocessing step in various image processing tasks [177], and to reduce statistical dependencies of visual signals for IQA tasks.

For each value of pixel $(i, j)$ from a patch $P$, the normalized value $P'(i, j)$ is computed using Eq. IV.5 where $\mu$ and $\sigma$ are respectively the mean and variance of intensity values in the normalization window. This latter is set to $3 \times 3$ according to [159] so as to avoid decreasing the performances with larger window sizes. $C \in \mathbb{N}$ is a positive constant used to avoid calculation instability.

$$P'(i,j) = \frac{P(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{IV.5}$$

$$\mu(i,j) = \sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} I(i+p, j+q) \tag{IV.6}$$

$$\sigma(i,j) = \sqrt{\sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} (I(i+p, j+q) - \mu(i,j))^2} \tag{IV.7}$$

**Local Response Normalization (LRN)** is related to LCN, however it aims more at normalizing the images in terms of brightness rather than contrast [178]. It was initially designed to normalize feature maps. It is defined as follows:

$$P'(i,j) = \frac{P(i,j)}{\left(k + \alpha \sum_{x=\max(0,i-n/2)}^{\min(W,i+n/2)} \sum_{y=\max(0,j-n/2)}^{\min(H,j+n/2)} P(x,y)^2\right)^{\beta}} \tag{IV.8}$$

where the constants $(k, \alpha, \beta, n)$ defined here as $(1, 1, 2, 2)$ are hyperparameters, $k$ being used to avoid instability.

### IV.4.3 Difference based methods

**Low-Frequency Subtraction (LFS)** is a simple normalization [169], performed by subtracting each patch $P$ from it low-pass filtered version. The low-frequency patch

is obtained by downscaling/upscaling $P$ by four. LFS is based on the assumption that HVS is less sensitive to changes in low-frequency bands.

**Difference of Gaussian (DoG)** [179] is obtained by the subtraction between two Gaussian filtered images at different SDs. This technique highlights details lost between the two Gaussian-blured versions of the image. This can be achieved as follows:

$$P' = (w_1 \cdot G_{\sigma 1} - w_2 \cdot G_{\sigma 2}) * P, \tag{IV.9}$$

with $G_{\sigma *}$ a Gaussian kernel of SD and $w_*$ its weight.

### IV.4.4 Histogram based methods

**Histogram Equalization (HE)** [179] aims to take full advantage of the range of pixel values by enhancing the image contrast. Consider $\lambda_i$ an intensity value, and a histogram $h(\lambda_i)$. $R_{\min}$ and $R_{\max}$ are the desired value bounds, and $\tilde{\lambda}_i$ the intensity level applied to $\lambda_i$.

$$\tilde{\lambda}_i = R_{\min} + \lfloor (R_{\max} - R_{\min}) \sum_{j=0}^{i} h(\lambda_j) \rfloor \tag{IV.10}$$

**Contrast Limited Adaptive Histogram Equalization (CLAHE)** [179] is based on HE. Rather than acting on the image intensity distribution equally, it uses spatial constraints and attempts to avoid noise amplification. Here, HE is applied locally with a fixed threshold $\lambda_{\text{th}}$.

### IV.4.5 Whitening based methods

**Zero-phase Component Analysis Whitening (ZCA) [180]** is a whitening technique that aims to normalize illumination by decorrelating features within images. This can be achieved as follows:

$$P' = \mathrm{U} \cdot \mathrm{diag}\left(1/\sqrt{\mathrm{diag(S)} + \varepsilon}\right) \cdot U^T \cdot P, \tag{IV.11}$$

where diag(.) is the diagonal matrix, $U$ the Eigen vectors and $S$ the Eigen values of singular value decomposition of covariance matrix. $U^T$ is the transposed matrix of $U$. $\varepsilon$ is the whitening coefficient.

**Simplified Whitening** is quite similar to standardization, but per-pixel mean and per-pixel SD are computed instead of feature-wise mean and SD.

In order to conduct the comparative study among the described normalization methods, a patch-wise training scheme is adopted. Patches of $64 \times 64$ are sampled from each input image. Each patch $P_i$ sampled from the image $I$ is labeled using the MOS associated with $I$. The normalization is applied to individual patches rather than the whole images, so as to account for local luminance and contrast, which vary in different parts of the image.

## IV.5 Experimental setup

TensorFlow [141] is used to implement the models used to conduct the described studies. During training, the datasets are split into a training set with 60%, a validation set with 20% of impaired images, and a testing set with the 20% remaining. To achieve total separation of the training and testing content, the impaired images related to the same pristine one are allocated to the same set. The same splitting scheme is used for all configurations to ensure a fair and reliable comparison.

**Table IV.2:** Summary of training and evaluation setup of the aforementioned studies.

| | Data-augmentation | Scores aggregation | Adaptive sampling | Normalization |
|---|---|---|---|---|
| Datasets | CVIQ / OIQA | CVIQ / OIQA | CVIQ / OIQA / MVAQD | CSIQ / LIVE / TID2013 |
| Server | **CPU:** Intel Xeon Silver 4208 2.1GHz / **RAM:** 192G / **GPU:** Nvidia Telsa V100S 32G | | | |
| Optimizer | Adam [181] / $lr = 1e-4$ / $\beta_1 = 0.9$ / $\beta_2 = 0.999$ / $\epsilon = 10^{-8}$ | | | |
| Batch size | 32 | 32 | 64 | 128 |
| 5-fold cross-evaluation | ✓ | ✓ | ✓ | ✓ |
| Evaluation metric | PLCC / SRCC | PLCC / SRCC / RMSE | PLCC / SRCC / RMSE /MAE | PLCC / SRCC / RMSE / Krasula *et al.* [45] |

## IV.6 Results and Discussion

In this section, we provide the obtained results and the corresponding discussion.

### IV.6.1 Data-augmentation

Table IV.3 summarizes the performance of individual VO-based training in terms of accuracy of prediction (PLCC) and monotonicity (SRCC), as well as the application of combined VOs on both databases. The latter refers to the data-augmentation based training. Regarding the performances of individual VOs, we can observe that the range

of performance is not significantly different. It is confirmed by the standard deviation given in the table. This actively demonstrates that, the various predicted scan-paths are almost of similar importance, and none of them can be considered as non-valid or outlying.

**Table IV.3:** Performance evaluation of the model. The Best performance is highlighted in **bold**. The mean of 5 folds is provided

| | MOS | | | | NIQE | | | | BRISQUE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CVIQ | | OIQA | | CVIQ | | OIQA | | CVIQ | | OIQA | |
| | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC |
| $VO_1$ | 0.829 | 0.773 | 0.898 | 0.884 | 0.756 | **0.791** | 0.445 | 0.382 | 0.771 | 0.673 | 0.732 | 0.714 |
| $VO_2$ | 0.815 | 0.753 | 0.877 | 0.860 | 0.791 | 0.685 | 0.426 | 0.398 | 0.759 | 0.707 | 0.747 | 0.713 |
| $VO_3$ | 0.836 | 0.762 | 0.907 | 0.892 | 0.793 | 0.677 | 0.452 | 0.406 | 0.743 | 0.693 | 0.751 | 0.731 |
| $VO_4$ | 0.835 | 0.759 | 0.911 | 0.895 | 0.792 | 0.686 | 0.419 | 0.372 | 0.772 | 0.716 | 0.690 | 0.661 |
| $VO_5$ | 0.830 | 0.765 | 0.879 | 0.868 | 0.752 | 0.620 | 0.432 | 0.415 | 0.673 | 0.626 | 0.779 | 0.740 |
| $VO_6$ | 0.820 | 0.748 | 0.916 | 0.898 | 0.781 | 0.653 | 0.498 | 0.457 | 0.792 | 0.723 | 0.710 | 0.662 |
| $VO_7$ | 0.838 | 0.759 | 0.888 | 0.872 | 0.738 | 0.605 | 0.450 | 0.423 | 0.729 | 0.656 | 0.777 | 0.749 |
| $VO_8$ | 0.845 | 0.783 | 0.898 | 0.880 | 0.801 | 0.700 | 0.462 | 0.399 | 0.768 | 0.711 | 0.767 | 0.736 |
| $VO_9$ | 0.817 | 0.760 | 0.902 | 0.884 | 0.743 | 0.616 | 0.446 | 0.398 | 0.735 | 0.683 | 0.758 | 0.712 |
| $VO_{10}$ | 0.835 | 0.754 | 0.893 | 0.872 | 0.722 | 0.591 | 0.479 | 0.412 | 0.722 | 0.662 | 0.743 | 0.714 |
| Avg | 0.830 | 0.762 | 0.897 | 0.881 | 0.767 | 0.662 | 0.451 | 0.406 | 0.746 | 0.685 | 0.745 | 0.713 |
| STD | 0.010 | 0.010 | 0.013 | 0.013 | 0.028 | 0.059 | 0.024 | 0.023 | 0.034 | 0.031 | 0.028 | 0.030 |
| All | **0.871** | **0.801** | **0.920** | **0.904** | **0.827** | 0.704 | **0.519** | **0.478** | **0.799** | **0.738** | **0.811** | **0.788** |

Between the MOS, local quality (NIQE and BRISQUE) for the model training, the former performed the best. In terms of difference, the range of correlation for the MOS is the smallest among the studied cases. The obtained results contradict our expectations. Assigning the same MOS value to small regions from the same 360-degree image looks at a first sight as not appropriate. Applying 2D models are adopted to account for local quality related to extracted regions. However, the used blind metrics did not improve the prediction accuracy globally in terms of PLCC and SRCC.

The proposed data-augmentation by the use of all VOs combined improved the performances for all the three cases, regardless of the used database, as can be seen in Table IV.3. The PLCC (resp. SRCC) value shifted from an average of 0.830 (resp. 0.762) to 0.871 (resp. 0.801) for the MOS based training on CVIQ. A similar behaviour is observed on OIQA where an improvement is achieved over the performance of individual observers. As for the use with NIQE and BRISQUE, an improvement is also observed both for PLCC and SRCC. The level of improvement should be put into perspective over the range of performances, which is higher for the blind metrics. Based on the previous correlation results, NIQE and BRISQUE do not appear as the best alternative to replace the MOS for data-augmentation. However, other performance data should be analyzed before drawing final conclusions.

When comparing between databases, one can observe a higher performance on

OIQA compared to CVIQ, except with NIQE. The difference is obvious with the MOS-based training supporting the previously discussed observation regarding the variety and diversity of the content present on OIQA. This led to a significant performance *i.e.* PLCC (resp. SRCC) value of 0.920 (resp. 0.904) compared to 0.871 (resp. 0.801) on CVIQ.

**Table IV.4:** Computational complexity in terms of training time for data-augmentation on CVIQ and OIQA databases. The mean of 5 folds is provided.

| | Database | MOS | NIQE | BRISQUE |
|---|---|---|---|---|
| Time (s) | CVIQ | 6285 | 3093 | 2577 |
| | OIQA | 3212 | 2360 | 3320 |

With the intent to compare the computational complexity of the proposed data-augmentation, we compute the training time for individual VOs as well as their combination (data-augmentation). It is given on Fig. IV.11 for CVIQ and Fig. IV.12 for OIQA where one can notice that, the MOS-based training has the lowest training time for all VOs except for $VO_{3,5,7}$ on CVIQ and $VO_1$ on OIQA. This could be explained by the lack of scores diversity during the learning process, leading to a faster convergence. At the contrary, BRISQUE generates a considerably higher training time, which is even more extensive for NIQE on both databases. This observation becomes invalid when it comes to the proposed data-augmentation, as shown by Table IV.4. Hence, the MOS-based case requires more than twice the NIQE/BRISQUE training time on CVIQ, and on OIQA the BRISQUE-based training took the longest time. This can be explained by the fact that, learning from a considerable amount of data that is associated with the same quality score (*i.e.* MOS) tends to make the model converge slowly. More data implies more diversity for the model to learn from. However, associating this diverse data with the same labels has a negative effect by increasing the computational cost. With NIQE and BRISQUE, the model is able to converge quickly as more data is available with distinct quality scores.

In addition to the computational time, we analyzed the evolution of the loss for the data-augmentation case. Figs. IV.13 and IV.14 plot the contrast (max-min) / (max+min) between training and validation losses for the five folds. A contrast equal to 0 depicts an equal loss between training and validation. On the contrary, a contrast equal or close to 1 indicates an important gap between both losses. In addition to the contrast, Fig. IV.13 and IV.14 provide the final loss values for both training (T) and validation (V) for each fold and each studied case on CVIQ and OIQA respectively. We can see that the MOS-based learning has more difficulties to generalize either with OIQA or CVIQ. In fact, the gap between T and V for MOS is much higher than those of NIQE- and BRISQUE-based cases. This is also demonstrated by the provided curves for both databases.

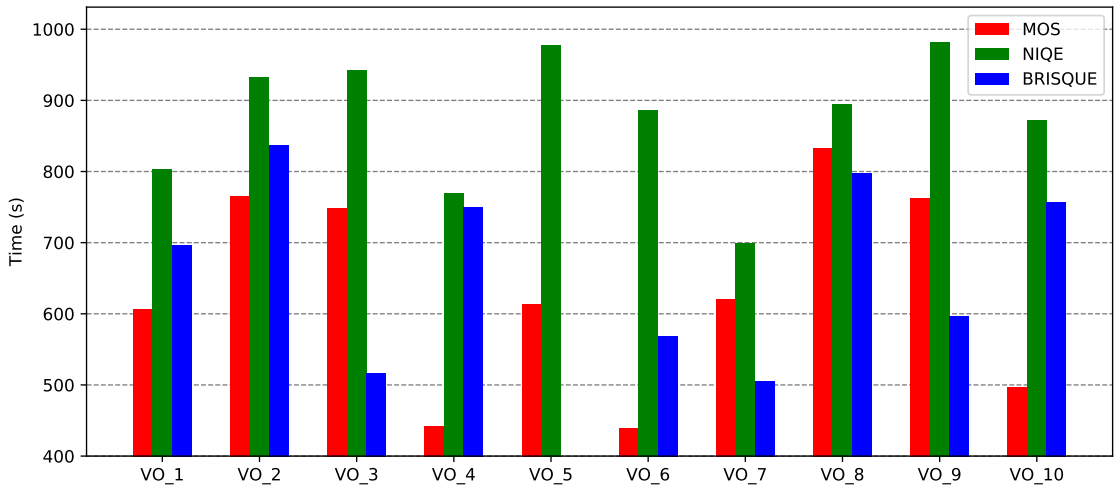



**Figure IV.11:** Computational time for VOs individually on CVIQ.

### IV.6.2 Aggregating patch qualities

As mentioned in Sec. IV.2.2, we use the data-augmentation configuration to further investigate the performances of different aggregation methods, as well as the proposed one.

We summarize the performances of all pooling strategies in terms of PLCC and SRCC in Table IV.5. The provided performances are computed as the median of the five-fold cross-validation. Overall, one can notice that the widely used arithmetic mean ranks among the worst approach on both databases, demonstrating its weakness when it comes to quality pooling. Pooling strategies accounting for the variability among quality scores should be considered in this case, as shown by the performance results in Table IV.5. One can observe that harmonic and geometric means outperformed the arithmetic one on both databases. For instance, the harmonic mean performances are approx. 0.8% PLCC, 1.0% SRCC, and 4.2% RMSE better than the arithmetic mean on OIQA, and approx. 0.6% PLCC, 1.2% SRCC, and 4.0% RMSE on CVIQ. The Minkowski mean and the five-number summary did not perform well compared to the arithmetic mean. A slight improvement can be observed with the Minkowski mean, whereas the five-number summary did not appear to express the nature of the variability among the local qualities scores. The percentile pooling achieved the best performance in terms of PLCC and SRCC on OIQA, and competitive results when combined with fixation orders and fixation durations on CVIQ. This shows that expressing the phenomena of perceived quality being impacted by the most distorted content improves the final quality pooling.

In the following, we analyze closely the performance of the Minkowski mean and the Percentile pooling. The evaluation of PLCC/SRCC scores is given in Fig. IV.15 and IV.16, respectively. For the Minkowski mean, one can observe a decrease in both accuracy and monotonicity with the increase of $P$. This observation is valid on both

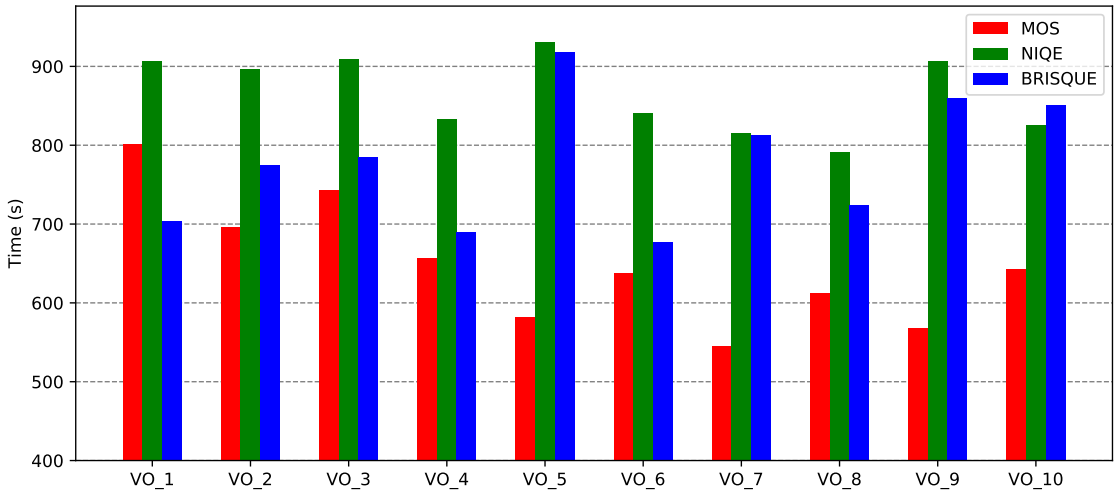



**Figure IV.12:** Computational time for VOs individually on OIQA.

databases, with a significant margin on CVIQ, approximately 6% with PLCC and 10% with SRCC. As for the Percentile Pooling, an increase of performances can be observed with a saturation at $k = 25$ on OIQA and $k = 10$ on CVIQ, followed by a decease of performance. Based on these observations, the parameter for both methods should be carefully chosen, as it is dependent on the variability and span of local qualities. In addition, the difference among OIQA and CVIQ is due to the nature and diversity of their content, as shown in [182]. This is also depicted by the provided curves, where an important gap between PLCC and SRCC values can be observed on CVIQ compared to OIQA independently of the used pooling methods.

With the intent to show the effectiveness of the patch-based CNN over multichannel models, we provide in Table IV.6 a performance comparison with three state-of-the-art models. These models adopt a multichannel paradigm using different strategies. From the table, one can observe that patch-based CNN with a simple arithmetic mean achieved competitive results compared to Sun *et al.* and Zhou *et al.*. However, it scored worse than Xu *et al.* (approx. 3.8% PLCC and 4.5% SRCC) on OIQA and (approx. 3.0% PLCC and 8.7% SRCC) on OIQA. When the adaptive pooling is used, a different behavior is observed. The patch-based model outperformed Sun *et al.* and Zhou *et al.* on both databases, and scored slightly lower compared to Xu *et al.* on OIQA and achieved the best accuracy on CVIQ. This slight difference of performance could be considered as insignificant when weighted by the complexity generated by the multichannel architecture. These performances support the previous observation regarding the usefulness of adaptive pooling of local qualities on the one hand. On the other hand, patch-based CNN is as effective as multichannel networks, and sometimes even better if proper training techniques are adopted.

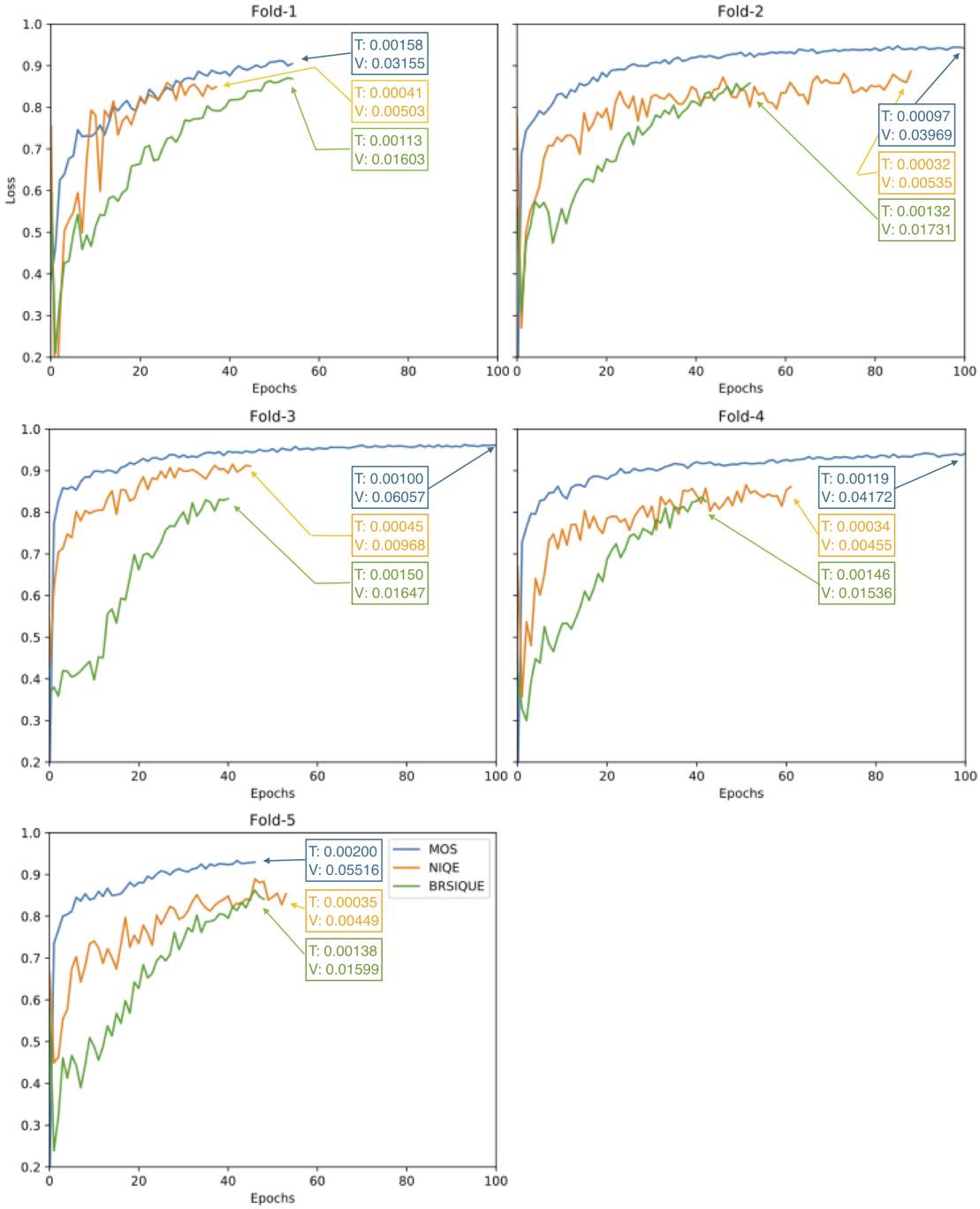



**Figure IV.13:** Contrast (max-min)/(max+min) between training and validation losses for the five folds (0 → equal loss between training and validation and 1→ important gap between both losses) on CVIQ. T and V represent the reached loss values for training and validation, respectively.

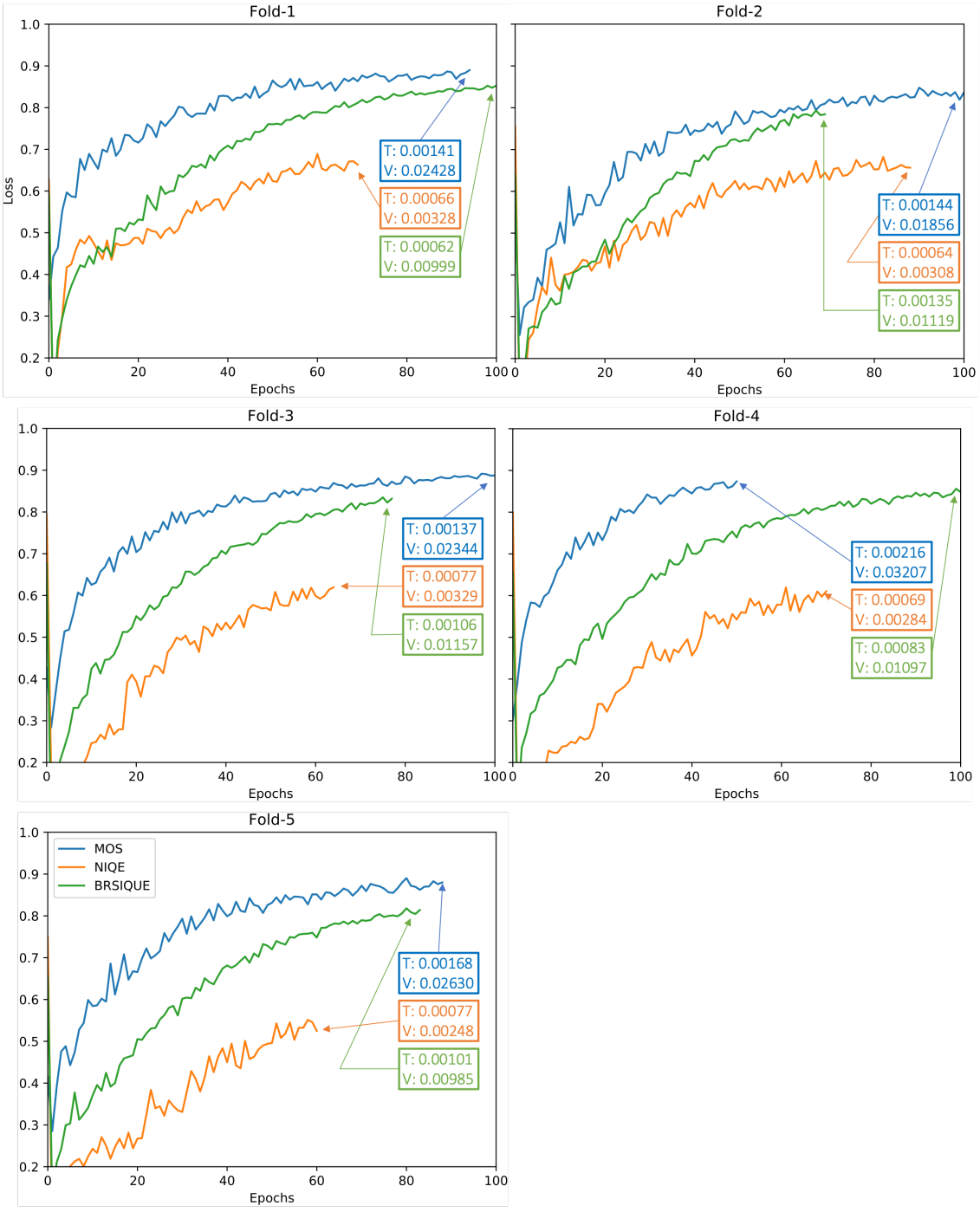



**Figure IV.14:** Contrast (max-min)/(max+min) between training and validation losses for the five folds (0 → equal loss between training and validation and 1→ important gap between both losses) on OIQA. T and V represent the reached loss values for training and validation, respectively.

**Table IV.5:** Performance evaluation of the pooling strategies in terms of PLCC, SRCC, and RMSE. The best performance is highlighted in **bold** and second-best underlined

| Database | OIQA | | | CVIQ | | |
|---|---|---|---|---|---|---|
| Metric | PLCC | SRCC | RMSE | PLCC | SRCC | RMSE |
| Arithmetic Mean | 0.9162 | 0.9017 | 5.8185 | 0.9297 | 0.8786 | 5.0537 |
| Harmonic Mean | 0.9235 | 0.9105 | 5.5685 | 0.9352 | 0.8891 | 4.8582 |
| Geometric Mean | 0.9200 | 0.9066 | 5.6876 | 0.9326 | 0.8841 | 4.9537 |
| Five-number summary | 0.9061 | 0.8971 | 6.1415 | 0.9233 | 0.8721 | 5.2683 |
| Minkowski Mean | 0.9196 | 0.9045 | 5.7034 | 0.9322 | 0.8833 | 4.9660 |
| Percentile Pooling | **0.9434** | **0.9340** | **4.8156** | 0.9623 | **0.9329** | 3.6790 |
| Fixation Order | 0.9063 | 0.8931 | 6.1357 | 0.9305 | 0.8787 | 5.0273 |
| Percentile Fixation Order | 0.9392 | 0.9296 | 4.8265 | 0.9621 | **0.9329** | **3.6287** |
| Fixation Duration | 0.9164 | 0.9028 | 5.8096 | 0.9296 | 0.8792 | 5.0564 |
| Percentile Fixation Duration | 0.9403 | 0.9291 | 4.8658 | **0.9625** | 0.9324 | 3.6883 |

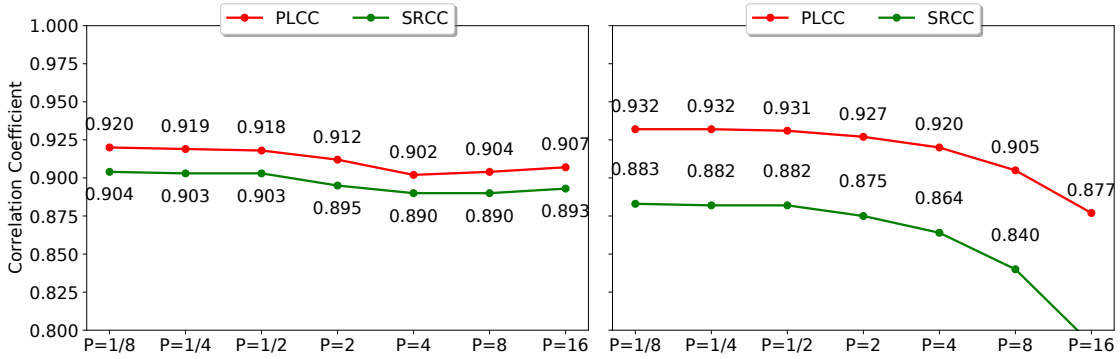



**Figure IV.15:** Performance of Minkowski mean in terms of PLCC/SRCC on OIQA (left) and CVIQ (right).

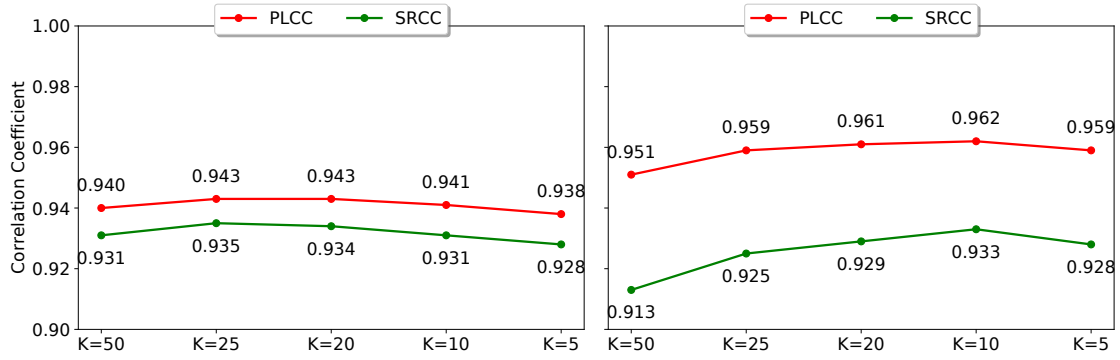



**Figure IV.16:** Performance of Percentile pooling in terms of PLCC/SRCC on OIQA (left) and CVIQ (right).

**Table IV.6:** Performance comparison with state-of-the-art mutlichannel-based models

| | Database | | OIQA | | CVIQ | |
|---|---|---|---|---|---|---|
| | Multichannel | Number (Backbone) | PLCC | SRCC | PLCC | SRCC |
| Xu *et al.* [100] | ✓ | 20 (Resnet-18) | **0.952** | **0.944** | 0.959 | **0.953** |
| Sun *et al.* [90] | ✓ | 6 (ResNet-34) | 0.924 | 0.918 | 0.950 | 0.914 |
| Zhou *et al.* [92] | ✓ | 6 (Inception-V3) | 0.899 | 0.923 | 0.902 | 0.911 |
| Ours $_{Arethmetic\ Mean}$ | ✗ | 1 (ResNet-50) | 0.916 | 0.902 | 0.930 | 0.879 |
| Ours $_{Adaptive\ Pooling}$ | ✗ | 1 (ResNet-50) | 0.943 | 0.935 | **0.963** | 0.932 |

### IV.6.3 Adaptive patch sampling

With the intent to evaluate the interest of using the proposed adaptive sampling and demonstrate the advantage of considering 360-IQA specific characteristics, we compare its performances with a standard sampling strategy. The latter takes patches from the ERP directly without considering any 360-IQA peculiarities. This means that the geometric distortion reflected by the stretched content in the polar regions is taken as is.

Table IV.3 lists the obtained performances on OIQA, CVIQ, and MVAQD. The first observation that emerges is that all configurations performed well on OIQA and CVIQ independently of the used sampling strategy, demonstrating the effectiveness of transfer learning. The performances on MVAQD lags behind those obtained on OIQA and CVIQ in terms of accuracy, monotonicity and prediction errors. This could be related to the dataset itself regarding the content variability. A more in-depth analysis shows that ResNet-50 performs robustly well by achieving the best performances on all datasets. This is reflected by the obtained overall correlations and associated SD over the five used folds.

According to the results on OIQA, ResNet-50 scored approx. 1.25% better with adaptive sampling in terms of accuracy compared to EfficientNet-B3. Similar margins can be observed with SRCC, RMSE and MAE. It appears that ResNet-50 is taking more advantage of the used peculiarity during patch sampling of 360-degree images. The opposite of this observation can be noticed on MVAQD where ResNet-50 achieved less prediction error in terms of RMSE and MAE with the standard sampling. The obtained performance by EfficientNet-B3 falls behind ResNet-50. It scored the worse on all datasets, independently of the used sampling strategy.

Overall, one can say that the adaptive sampling is improving the performances of both models as highlighted by the results, with important margins on MVAQD. The use of ERP content lags behind the effectiveness of the latitude and radial based sampling. Despite its popularity, ERP format is known for being geometrically distorted due to the projection process on the one hand. On the other hand, ERPs do not represent the subjectively actual rated content, as highlighted in the previous chapters. In this case, the use of radial content in addition to latitude-based sampling complies with (i) the

**Table IV.7:** Performance evaluation on OIQA, CVIQ, and MVAQD. The median and standard deviation (SD) over five-folds are provided. Best performance is highlighted in **bold**.

| | Model | ResNet-50 | | EfficientNet-B3 | |
|---|---|---|---|---|---|
| | Sampling method | Adaptive | Standard | Adaptive | Standard |
| OIQA | PLCC↑ (SD↓) | **0.9614** (0.034) | 0.9267 (0.025) | 0.9494 (0.045) | 0.9223 (0.038) |
| | SRCC↑ (SD↓) | **0.9529** (0.029) | 0.9352 (0.026) | 0.9474 (0.037) | 0.9338 (0.041) |
| | RMSE↓ (SD↓) | **0.0642** (0.010) | 0.0619 (0.007) | 0.0783 (0.008) | 0.0681 (0.007) |
| | MAE↓ (SD↓) | **0.0518** (0.008) | 0.0501 (0.005) | 0.0635 (0.006) | 0.0532 (0.006) |
| CVIQ | PLCC↑ (SD↓) | **0.9491** (0.025) | 0.9267 (0.025) | 0.9120 (0.027) | 0.9105 (0.032) |
| | SRCC↑ (SD↓) | **0.9562** (0.054) | 0.9352 (0.026) | 0.8951 (0.056) | 0.8941 (0.053) |
| | RMSE↓ (SD↓) | **0.0603** (0.028) | 0.0619 (0.007) | 0.0851 (0.014) | 0.0753 (0.018) |
| | MAE↓ (SD↓) | **0.0482** (0.024) | 0.0501 (0.005) | 0.0709 (0.013) | (0.0614 (0.016) |
| MVAQD | PLCC↑ (SD↓) | **0.9016** (0.086) | 0.8861 (0.118) | 0.8509 (0.027) | 0.8025 (0.028) |
| | SRCC↑ (SD↓) | **0.8941** (0.100) | 0.8653 (0.108) | 0.8242 (0.040) | 0.7715 (0.033) |
| | RMSE↓ (SD↓) | 0.1406 (0.025) | **0.1047** (0.030) | 0.1616 (0.030) | 0.1423 (0.030) |
| | MAE↓ (SD↓) | 0.1154 (0.020) | **0.0868** (0.023) | 0.1391 (0.026) | 0.1199 (0.026) |

way the human observers explore a 360-degree scene and (ii) the predictions made on the viewed content using HMDs.

In addition to the performance analysis, we recorded the computational time required for one step during training. A training step is one gradient update, meaning that a single batch size of training examples are processed. In our case, the batch size was set to 64. ResNet-50 required 48.0 ms/step while EfficientNet-B3 required 60.2 ms/step. The architecture of ResNet-50 allows it to efficient computation with the help of skip connections. In addition to the achieved performances by ResNet-50, it required less time, making it more robust and computationally efficient overall.

### IV.6.4 Patch normalization

Table IV.8 provides the list of the normalization methods used in this study.

**Table IV.8:** Labels for the used normalization methods.

| Method | Label | Method | Label | Method | Label |
|---|---|---|---|---|---|
| Scaling [0, 1] | N1 | LFS | N5 | Simp. whitening | N9 |
| Standardization | N2 | LRN | N6 | HE | N10 |
| Mean-Centering | N3 | DoG | N7 | CLAHE | N11 |
| LCN | N4 | ZCA | N8 | | |

### IV.6.4.1 Global performance

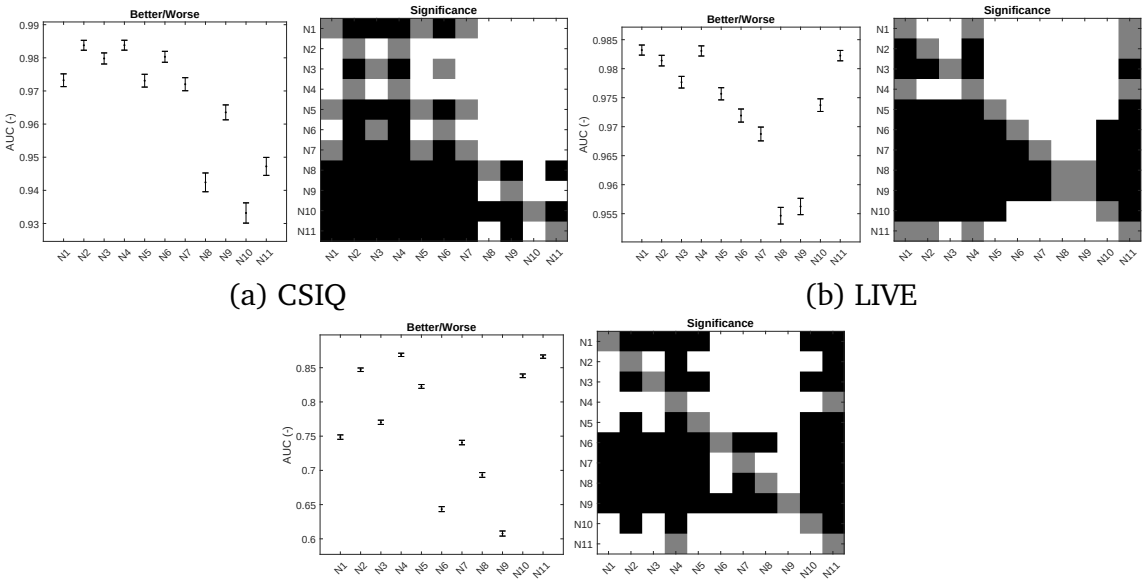
Table IV.9 summarizes the overall performances of the adopted architecture with the different normalization methods using CSIQ, LIVE, and TID2013. For simplicity, the normalization methods are labeled according to Table IV.8. The best performances are obtained on LIVE followed by CSIQ and then TID2013. It is known that the latter is quite challenging as it contains 24 different distortions, making the generalization ability of IQA models less robust. Also, it is of note that this database was collected in an uncontrolled manner [139] compared to CSIQ and LIVE. Regarding the best normalization, N4 outperformed the other methods by achieving the best performance in terms of accuracy (PLCC), monotonicity (SRCC) and error (RMSE). This observation is valid regardless of the used database, depicting the interest of input normalization, on the one hand. On the other hand, LCN has been considered as an effective normalization method for IQA [135, 176]. At the same time, methods N7, N8 and N9 performed poorly among the selected methods. These methods are not specifically designed for IQA. Regarding the use of basic pixel values representations such as N1, N2, and N3, one can observe that despite their simplicity, they obtain competitive performance better than more elaborated methods.

**Table IV.9:** Performance comparison of the selected normalization on CSIQ, LIVE, and TID2013. The median over five folds is taken. The best performance is highlighted in **bold** and second-best underlined

| | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 | N11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | CSIQ | | | | | |
| PLCC | 0.9009 | 0.9060 | 0.9139 | **0.9236** | 0.8737 | 0.9163 | 0.8935 | 0.7100 | 0.8498 | 0.8583 | 0.8893 |
| SRCC | 0.8470 | 0.8493 | 0.8686 | **0.8932** | 0.8050 | 0.8837 | 0.8272 | 0.7379 | 0.8165 | 0.7924 | 0.8158 |
| RMSE | 0.1105 | 0.1081 | 0.1139 | **0.1031** | 0.1227 | 0.1065 | 0.1147 | 0.1758 | 0.1424 | 0.1306 | 0.1165 |
| | | | | | | LIVE | | | | | |
| PLCC | 0.9536 | 0.9400 | 0.9401 | **0.9569** | 0.9440 | 0.9399 | 0.9235 | 0.9175 | 0.9119 | 0.9413 | 0.9538 |
| SRCC | 0.9463 | 0.9324 | 0.9392 | **0.9528** | 0.9430 | 0.9390 | 0.9194 | 0.9081 | 0.9060 | 0.9402 | 0.9466 |
| RMSE | 0.0480 | 0.0543 | 0.0557 | **0.0477** | 0.0548 | 0.0567 | 0.0638 | 0.0661 | 0.0660 | 0.0561 | 0.0478 |
| | | | | | | TID2013 | | | | | |
| PLCC | 0.6520 | 0.6780 | 0.6841 | **0.7406** | 0.6686 | 0.6728 | 0.6265 | 0.6354 | 0.6739 | 0.6969 | 0.6798 |
| SRCC | 0.5424 | 0.6150 | 0.6014 | **0.6348** | 0.5581 | 0.5925 | 0.5407 | 0.5607 | 0.5729 | 0.5531 | 0.5703 |
| RMSE | 0.1043 | 0.1006 | 0.0990 | **0.0953** | 0.1044 | 0.1013 | 0.1068 | 0.1057 | 0.1006 | 0.0986 | 0.1008 |

### IV.6.4.2 Statistical significance

We provide in Fig. IV.17 the overall statistical significance analysis on CSIQ, LIVE, and TID2013. On the left side, we provide the better or worse in terms of quality classification with growing difference in predicted score, providing insights on how many times does the model correctly recognize the stimulus of higher quality. On



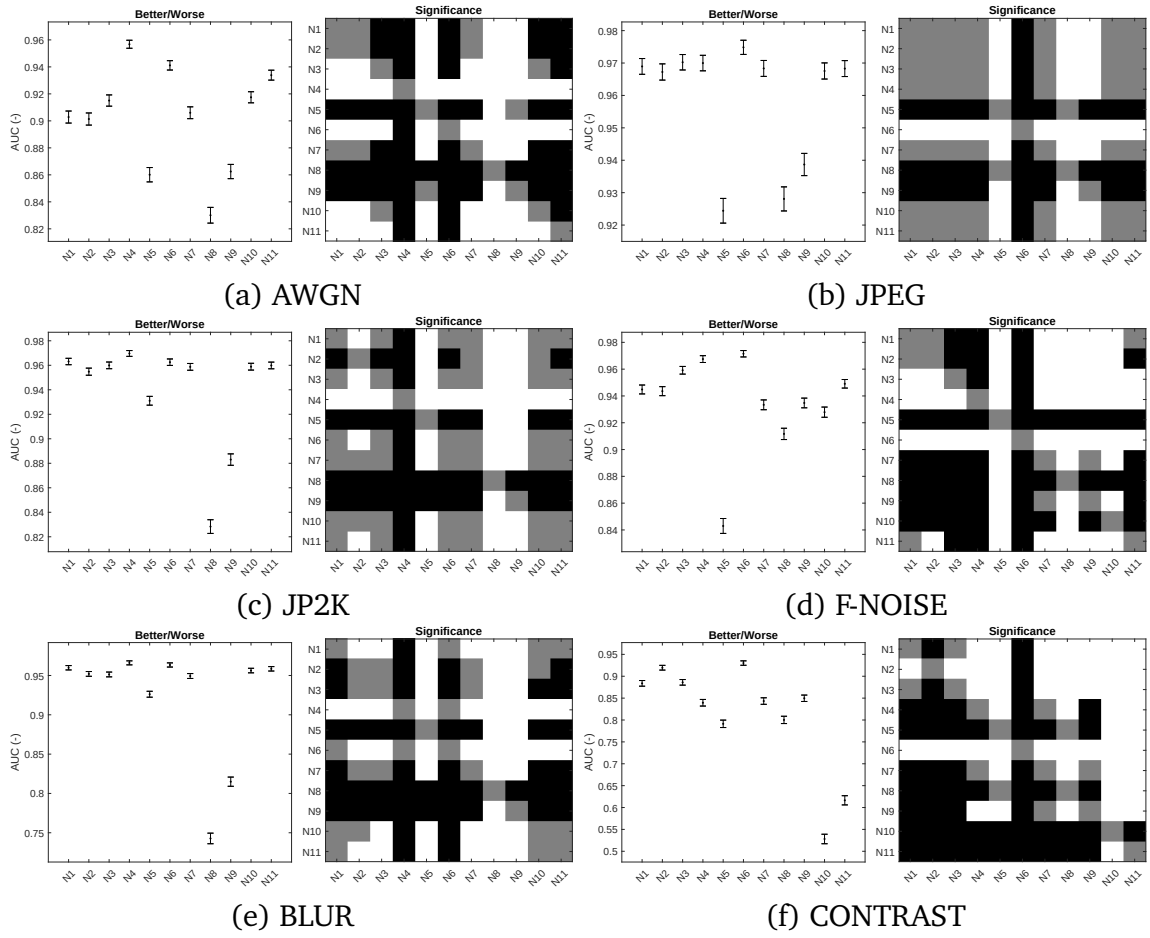

(a) CSIQ (b) LIVE

(c) TID2013

**Figure IV.17:** Overall Statistical Significance on CSIQ, LIVE, and TID. (Left) better vs. worse analysis and (Right) statistical significance. A white/black square: row method is statistically better/worse than the column one; gray square: statistically indistinguishable.

the right side, the statistical significance among the considered methods is provided to determine if the difference in performance is statistically significant. As it can be seen, different results are obtained with each normalization, depicting an influence of each method on the final performances. Overall, N2 and N4 stood out from the other methods. On CSIQ, one can notice that these two methods outperformed the others while being statistically indistinguishable compared to each other. On LIVE, N1, N4, and N11 scored the best performances. Finally, on TID2013, we can find N4 and N11 outperforming the rest of the methods. As N2 is the standardization, which is essentially a pixel value representation approach, one may claim that training CNNs without normalization is a good choice. However, N4 and N11, which represent the LCN and CLAHE, appear to improve the model's performance by being statistically superior to basic scaling methods, *i.e.* N1, N2, and N3. Among the three databases, N7 to N10 performed poorly compared to the others while the LCN performed the best, demonstrating its usefulness and explaining its popularity in the literature.

The overall statistical significance analysis highlighted the LCN, standardization, and CLAHE performances over the other methods. However, considering each degradation separately, interesting findings are reflected. The statistical significance in terms of better/worse and significance per degradation are provided in Fig. IV.18 and IV.19 for CSIQ and LIVE, respectively. Unfortunately, we could not include TID2013 in this analysis due to the pages limitation. With CSIQ, one can notice that N6 appears to be better than all the other methods on JPEG, F-Noise, and Contrast. This method is



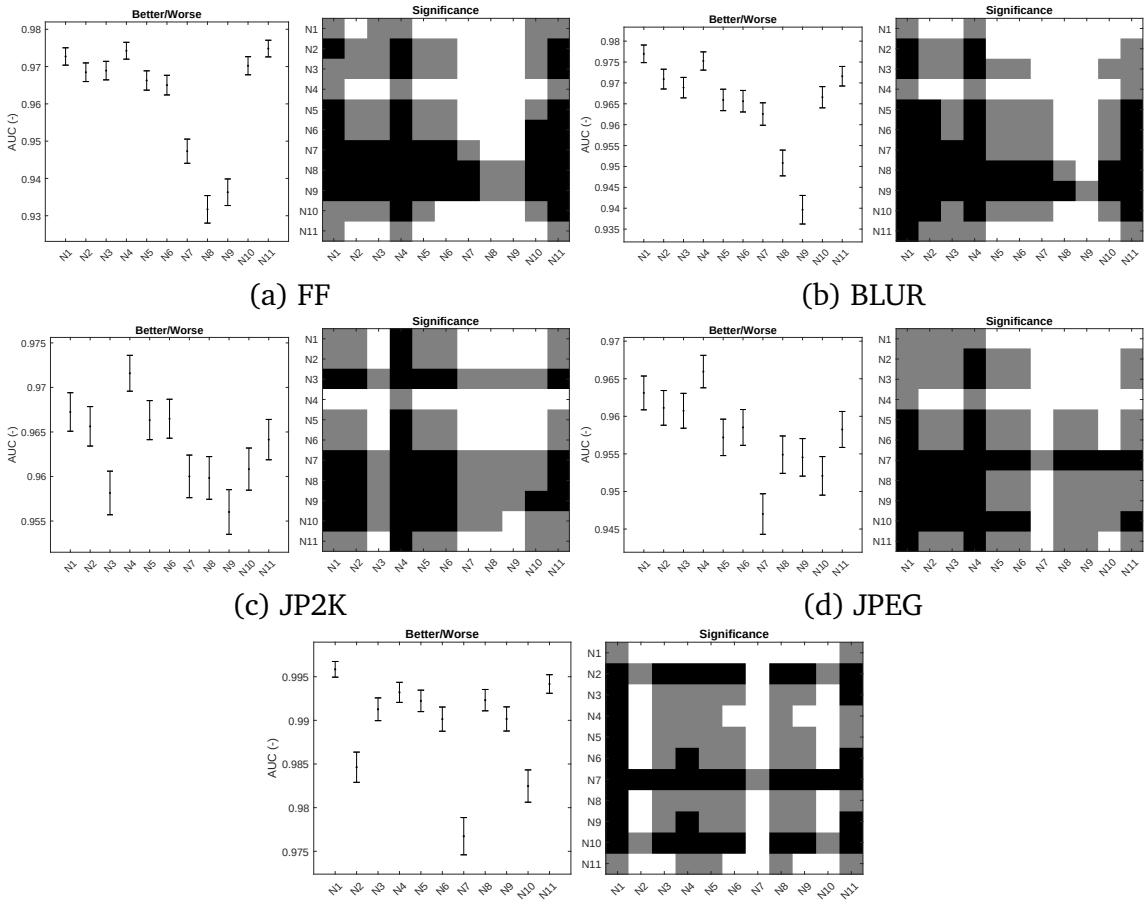

(a) AWGN (b) JPEG

(c) JP2K (d) F-NOISE

(e) BLUR (f) CONTRAST

**Figure IV.18:** Statistical significance on CSIQ per individual degradation.

performing well locally for some distortions and poor globally as seen on the overall performances. N4 achieved the best with AWGN, JP2K, and BLUR, and competitive results with N6 on F-NOISE. This shows the effectiveness of LCN locally per degradation. However, on contrast impairment the LCN scored worse compared to N1, N2, N3, N6, and N9. This is mainly due to the fact that contrast changes are not retained when normalizing images using LCN, leading to poor performances with regard to this degradation. One can also observe that satisfactory performances are obtained with the basic methods N1, N2, and N3 with JP2K and BLUR distortions. However, these performances were not enough to outperform N4 nor N6, supporting the idea to perform proper normalization prior to CNN training. In terms of the worst performances, N8 and N9 performed poorly among the selected normalization except on contrast. On the latter, N10 and N11 gave the worst performance. It appears that the histogram equalization based methods did not cope well with contrast distortions. It is worth noting that these methods enhance the image contrast, and by doing so on already contrast-distorted images, it results in a poor performance.




(a) FF (b) BLUR

(c) JP2K (d) JPEG

(e) White Noise (WN)

**Figure IV.19:** Statistical significance on LIVE per individual degradation.

As with LIVE (Fig. IV.19), the same observation on CSIQ best performing methods holds. N4 outperformed the other methods by providing significantly better results, except on WN where it performed worse compared to N1, which is also reflected by the associated better/worse analysis. Regarding the worst performances, N7-10 seem unsatisfactory on LIVE, which is also the case on CSIQ.

To summarize, the best performing normalization method on the overall databases is found to be the LCN, a widely used IQA-specific method, which was confirmed by the achieved performances in terms of accuracy, monotonicity and error as well as the statistical significance. When analyzing the performances per degradation, the best performance was achieved by LCN and LRN on CSIQ and LCN on LIVE over basic scaling methods and histogram equalization based ones. This demonstrates (i) the usefulness of normalization, and (ii) that the use of a proper method may enhance the performance and robustness of the CNN model.

## IV.7 Takeaways

The importance of pre-processing and data preparation in attaining good results cannot be underestimated. Considering deep-learning models typically learn a predictive model using training data, it is critical to ensure better consistency and representation of this data. From the described contributions in this chapter, one can conclude that each pre-processing step, including input sampling, normalization, data-augmentation, and labeling, has the potential to improve independently and jointly the learning rate of the model. Regarding the post-processing, adaptive aggregation of patch scores to derive the global quality can enhance the prediction correlation. The fact of weighting individual scores for patch-based CNNs with IQA-specific properties achieved comparative and even better performances compared to multichannel models.

The main observations from this chapter are:

- Better results and generalization can be obtained with IQA-specific data augmentation techniques. In our case, the visual scan-path yielded a good performances.
- Local quality labeling for patch-based training is a delicate task. The MOS-based labeling outperformed 2D metrics in terms of performances, whereas less training time was required when using 2D metrics.
- Adaptive aggregation of patch qualities contributes significantly to the performances of the model.
- Patch-based CNN for 360 images provides competitive performances to multichannel CNNs. This is achieved with less complexity. We believe, improvement can be made with more appropriate IQA-specific adaptations.
- Adaptive sampling improved the performances compared to standard sampling. By taking patches on the sphere with a latitude and importance based sampling strategy, the performances can be significantly improved.
- Adaptive data representation is more efficient compared to simple scaling of the input data. The performance of neural networks can be significantly improved with adapted normalization in general, and IQA-based ones in particular.

## IV.8 Conclusion

With a focus on the performances gain, we presented in this chapter several analysis related to pre- and post-processing for NR CNN based 360-IQA.

First we analyzed the use of visual scan-paths as data-augmentation for 360-degree IQA, mainly for reducing over-fitting and improving the prediction performances. To do so, ten different scan-paths (simulating 10 virtual observers) were generated with eight possible fixation points each, are used as centers of the generated viewports. Besides, a comparison is made between the use of blind metrics (BRISQUE and NIQE) and MOS for local quality (quality of patches) for training the model.

The obtained results demonstrated an improvement when using data-augmentation compared to individual virtual observers. The lack of diversity of the MOS values associated with the same 360-degree image, does not allow to reach sufficient generalization of the model. In addition, it requires more computational time than the cases using blind metrics.

We further investigate the use of adaptive aggregation strategies for 360-degree IQA using a patch-based CNN. We found that the use of a simple arithmetic mean, which is the most common and straightforward technique, does not account for the variability among the quality scores, and therefore, the correlation performance tends to drop. Adaptive pooling strategies are seen as a good answer to cope this limitation, especially when IQA-specific characteristics are incorporated. Moreover, patch-based CNN with adaptive pooling achieved competitive performances compared to state-of-the-art multichannel models. As patch-based CNNs introduce less complexity compared to multichannel, it makes it more appropriate with an adequate training strategy for 360-degree IQA.

Afterward, an adaptive patche sampling strategy is designed. The performances of the latter are compared to standard sampling under two configuration. To this end, two pre-trained models were used, including ResNet-50 and EfficientNEt-B3. The results demonstrated a robust performance of ResNet-50 in terms of prediction accuracy. A good generalization was observed on all datasets, particularly with the adaptive sampling strategy. This supports the use of 360-degree images peculiarities and the superiority of the proposed sampling strategy. Still, several challenges need to be resolved, especially patches' labeling with the same MOS leading to less diversity. The construction of a large and representative 360-IQA dataset would allow more robust and generalized models.

Finally, we conducted an empirical analysis on input image normalization methods prior to CNN training for IQA. Here, we focused on 2D-IQA in order to draw conclusion for immersive content such as 360-degree images. The motivation behind such choice lies in the need of large datasets, and 2D-IQA one are more large and diverse compared to 360-IQA's. Several methods are considered, ranging from IQA-specific ones to others designed for specific image processing applications. The overall results on three commonly used databases showed that normalizing input images is better than utilizing basic scaling methods. The statistical significance analysis of the evaluated methodologies revealed the same findings. The best results were obtained by the local contrast normalization, which outperformed the other methods by achieving the best results across all databases and per degradation except contrast. In this latter case, the loss of information affected negatively the performance. Other methods were also efficient in reaching satisfying results for some specific degradation. According to the findings, using adequate normalization to improve the performance of CNN models is favorable. Accounting for the loss of information due to normalization during training can improve the model's robustness.

# Chapter V

# Perceptually-Weighted CNN For 360-IQA Using Visual Scan-Path And JND

## V.1 Introduction

IQA generally considers the quality of an image as the property of its entire content [183]. Due to this, existing IQA databases come with a global ground truth label, *i.e.* MOS, per image. The unavailability of MOS per individual regions on the one hand and the higher resolution of 360-degree images on the other hand, constrained the community to adopt the multichannel paradigm for deep-learning based 360-IQA model, as mentioned in Sec. I.2.2.2. As the a multichannel model is basically trained to predict the global quality of 360-degree images, it partially solves the unavailability of MOS per individual regions. This chapter presents a multichannel CNN that considers different perceptual characteristics of HVS represented, including (i) JND probability maps and (ii) visual scanpaths.

First, we extract viewports on the spherical content of 360-degree images according to visual scan-path predictions. This way, we reproduce the actual viewed content. Then, motivated by the effectiveness of well-known pre-trained CNN models confirmed by our study in Chapter III, we use DenseNet-121 [120] to extract visual features from the selected viewports and predicts their visual quality. We use the JND probability map to account for HVS sensitivity to local distortions. The proposed model estimates the weight of each extracted viewport by fusing JND information, extracted visual features, and visual scan-path attributes (fixation duration and fixation order). In the following we describe the proposed model.

## V.2 Proposed model

The proposed approach involves two steps. The first focuses on data pre-processing including scan-path prediction, viewports extraction, and JND probability maps generation. The second step consists of an end-to-end training. Details on each step are given below.

### V.2.1 Pre-processing

Inspired by the way 360-degree images are generally viewed, *i.e.* only portions of the images called viewports are seen by the users through HMDs (*see* Fig. IV.4), we only consider selected viewports to predict the quality. This can be justified by the fact that a user can only see the current rendered field of view (FoV) from the spherical representation. The next viewport depends on his head direction along the x, y, and z axes. This way, quality prediction scenario tends to be in agreement with the viewing experience of 360-degree images and geometric distortions caused by the sphere to plane projection mentioned previously are avoided.

It is now widely admitted that when an image is viewed, the HVS gazes on salient details, which translates into eye fixations [153]. In our case, these regions are considered as our viewports and are detected using the visual scan-path model proposed in [154]. This model provides trajectories including the order and duration of fixations. This information giving valuable data about the exploration behavior is fed to the CNN model described in the next section. It corresponds to a sequence of $N$ ordered fixation positions and their corresponding duration, respectively denoted as $[F_{Or}, F_D] \in \mathbb{R}^+$.

In our model, the above-mentioned information is predicted for ten different virtual observers representing the diversity of human scan-paths. The predicted scan-paths are considered as data augmentation, not for the training stage but to increase the diversity and robustness of the cross-validation. This will help with the generalization analysis. The motivation behind such an approach is that each virtual observer (VO) will explore the same scene but will probably provide a different rating as in real subjective experiments. Therefore, from each image in the dataset, we extract eight viewports for each VO where fixation points are taken as the center of the viewports with $512 \times 512$ resolution. This way, we generate ten different instances of the training dataset. An illustration of the scan-paths prediction is given in Fig. V.1. As it can be seen, each predicted visual trajectory, composed of eight ordered fixations, is distinct. Besides, most fixations fall on the equatorial region where the human gaze is usually biased. This demonstrates the consistency of the viewports sampling with the human exploration behavior when using HMDs. During the end-to-end training, each VO is used separately.

With the aim to perceptually account for the sensitivity of viewport content to distortions, and give more cues to our model about distortion visibility, we used JND probability maps. We believe that training the model to learn about HVS sensitivity

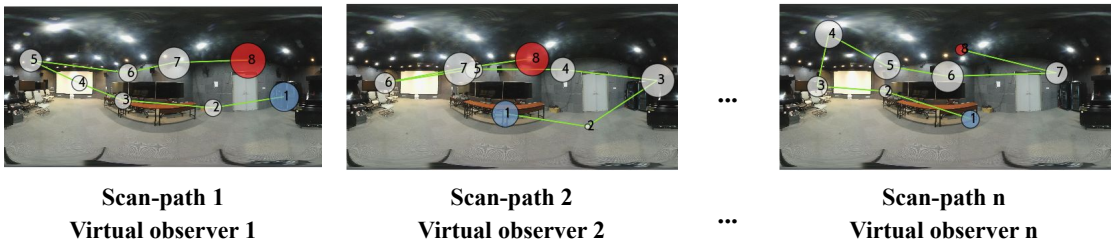



**Figure V.1:** Different scan-paths considered as virtual observers (VOs). Each scan-path is composed of eight ordered fixations. The radius of each fixation reflect the fixation duration. The color blue represent the first viewed viewport and the red one corresponds to the last viewport.

will perceptually improve the estimation of the weights to be given to each viewport when deciding about the quality of the 360-degree image. Fig. V.2 gives samples of extracted viewports and their respective JND probability maps. It shows the impairments detection probability values and their variation depending on the complexity of the region. Flat regions are prone to more visible distortions compared to more complex ones.

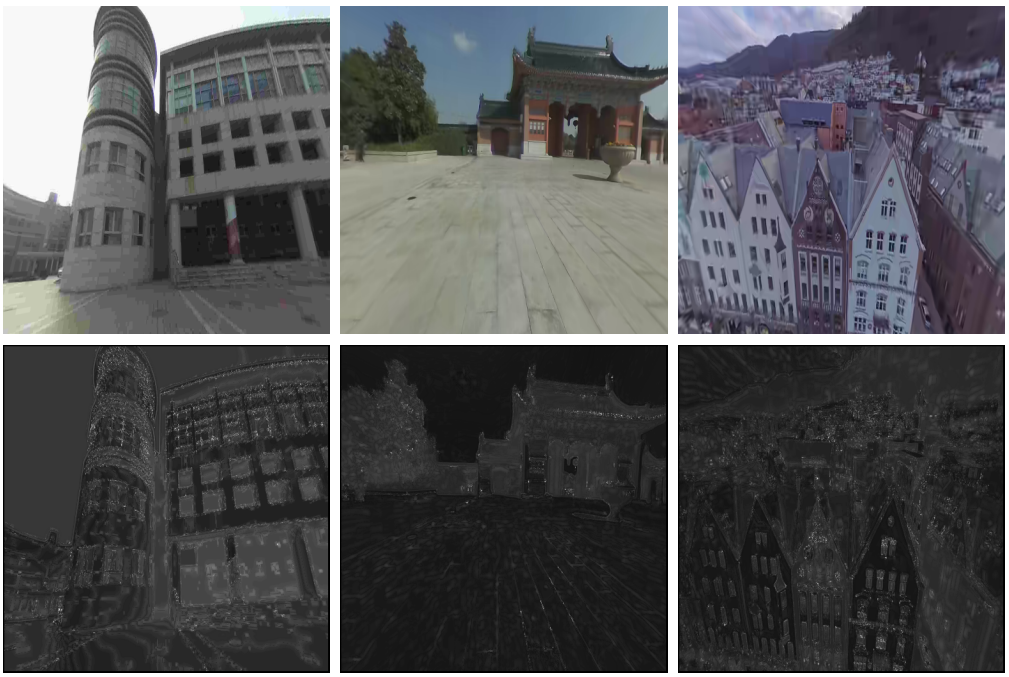


**Figure V.2:** (Top) Examples of extracted viewports and (Bottom) their corresponding JND probability maps.

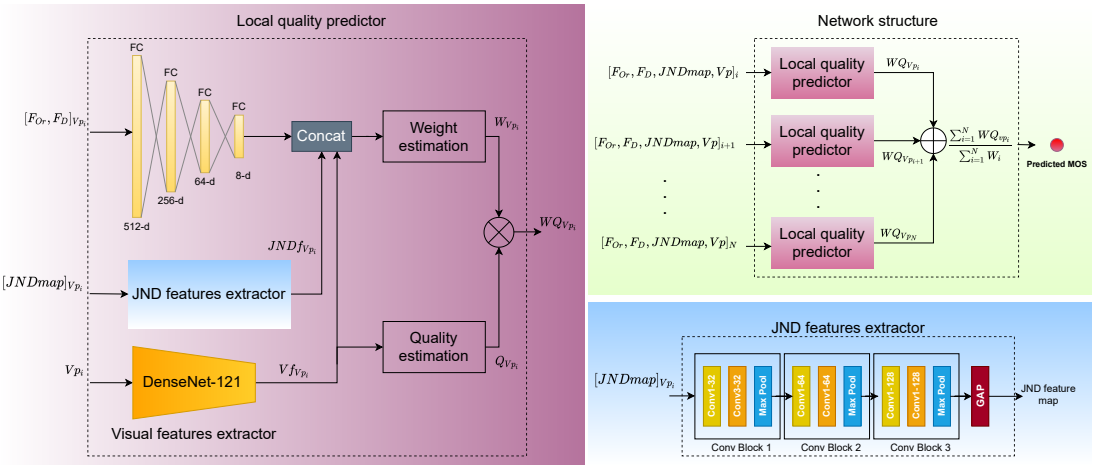



**Figure V.3:** Architecture of the proposed model: The green rectangle depicts the overall network structure, the magenta rectangle depicts the local quality predictor structure, and the blue rectangle depicts the JND features extractor network.

## V.2.2 Network Architecture

Fig.V.3 depicts the architecture of the proposed method with its different components. Given a set of viewports $Vp_i \in \mathbb{R}^+$ with $i \in \{0...N\}$ extracted from a 360-degree image, the model takes four inputs for each $Vp_i$ including visual content, JND probability map, fixations order and fixations duration. These inputs are fed to the local quality predictor (LQP) (green rectangle) resulting in $N\times$ LQP modules running in parallel. Then, the LQP module fuses different learned features and outputs a weighted quality score for each $Vp_i \in \mathbb{R}^+$ denoted as $WQ_{Vp_i}$. Finally, the model outputs the weighted arithmetic mean of the local quality scores as follow:

$$PredictedMOS = \frac{\sum_{i=1}^{N} WQ_{Vp_i}}{\sum_{i=1}^{N} W_{Vp_i}}, \tag{V.1}$$

where $WQ_{Vp_i}$ is obtained by:

$$WQ_{Vp_i} = W_{Vp_i} \times Q_{Vp_i}. \tag{V.2}$$

As shown in Fig. V.3, the main component is the LPQ which consists of three parts. The first is a visual feature extractor (VFE). Here, we use the DenseNet-121 [120] model with its original weights, *i.e.* ImageNet weights. The choice of the DenseNet model is made based on a previous comparative study that we conducted and for which it ranked first compared to VGG, ResNet, and Inception architectures. The study in question is detailed in Chapter III. The VFE provides a learned visual feature map $Vf_{Vp_i} \in \mathbb{R}^+$ that goes to a quality estimation module and is used also for the estimation of the weight $W_{Vp_i}$. The second part consists of JND features extractor that takes the JND probability map $JNDmap_{Vp_i} \in \mathbb{R}^+$ of $Vp_i \in \mathbb{R}^+$ and outputs a feature map that contributes to $W_{Vp_i}$ estimation. The Learned JND features account for the different

sensitivities of the HVS toward various distortion types and magnitudes. For the JND probability maps detection, we used the 2D model proposed in [184] as it is applied on the extracted viewports being assimilated to standard 2D images. Therefore $W_{Vp_i}$ is obtained as follows:

$$W_{Vp_i} = Concat([MLP([F_{Or}, F_D]), JNDmap_{Vp_i}, Vf_{Vp_i}]). \tag{V.3}$$

The proposed network used for JND features extraction aims to learn from HVS sensitivities [185]. It is composed of three convolutional blocks as illustrated in Fig. VI.2 (blue rectangle). Each block includes three layers, two convolutions ($1 \times 1$ and $3 \times 3$) kernels followed by a max-pooling layer. By adding a $1 \times 1$ convolutional layer before the $3 \times 3$ convolution, for the same height and width of the JND feature map, we reduce the number of operations. It also adds non-linearity to the network and allows to implement a smaller CNN while keeping a higher degree of accuracy [124]. Therefore, we are reducing the computational requirements and being more efficient at the same time.

At the final stage of the network, a GAP layer is used according to the recommendation in [124] to generate the feature vector. Finally, the third part is a MLP that takes as input the duration and order of fixations given by the visual scan-path predictor and encodes them to account for the visual exploration behavior. The MLP outputs a visual information vector used for the estimation of $W_{Vp_i}$. The fixation duration informs about which visual content is more likely to attract the user gaze. It also gives the time spent in visualizing a portion of the scene. As for the fixation order, it informs about the nature of the visual exploration path.

The weight estimation stage considers encoded duration and order of fixations, JND, and visual feature maps. These different features are fused and used to estimate the weights $W_{Vpi}$ of $Vp_i$ using four FC layers. In parallel, the visual feature map is also regressed to predict the quality score $Q_{Vpi}$ of $Vp_i$. For this, a GAP is performed on the output feature map of DenseNet-121 followed by an FC layer, a dropout layer, and another FC layer for score prediction. The final score is computed using Eq. V.1.

For the end-to-end training, we used the $L_2$ loss function to compute the error between predicted and target scores. The loss function is defined as:

$$loss = (q_{predicted} - q_{target})^2. \tag{V.4}$$

Three different versions of the proposed model are developed. The first version uses only fused visual features of the 8 extracted viewports from a given 360-degree image. It consists of eight pre-trained DenseNet-121 and a trained quality estimator. The second version accounts for scan-path features $F_{Or}$ and $F_D$ for weights estimation. Finally, the third version is built on top of version two by incorporating the JND probability maps for weight estimations.

## V.3 Results and discussion

### V.3.1 Experimental setup:

**Dataset:** The proposed model is trained and evaluated on the CVIQ [90] database, *see* Sec. II.2 for more details. We use the strategy discussed in Sec. V.2.1 where the proposed model is compared across ten predicted scan-paths. Hence, the model is trained under ten different iterations as illustrated in Fig. V.4. Each set associated with a virtual observer is divided into 60/20/20 for training, validation, and testing.

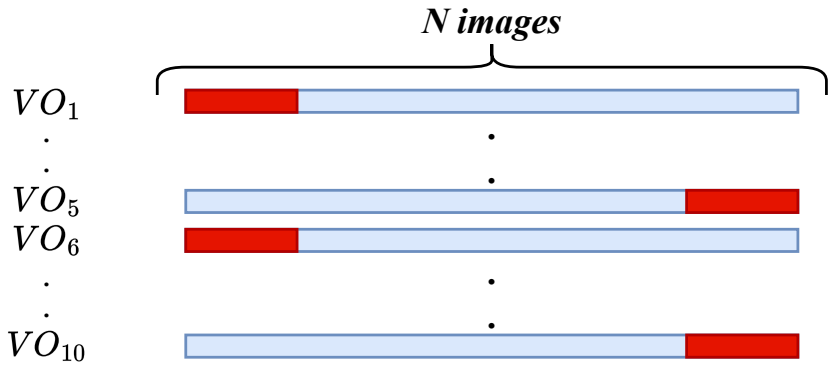



**Figure V.4:** Data splitting for training the proposed model. (Red) testing sets (blue) training sets.

**Implementation:** The proposed architecture is implemented using TensorFlow [141]. The training was performed using NVIDIA Tesla P100-PCIE-16GB and 26GB of RAM. We used the *early-stopping* to stop the training if no performance gain is observed by monitoring the validation loss.

**Performance evaluation:** The predicted scores are fitted using a five-parameter non-linear logistic function. The performance of the proposed model are computed using the ten VOs. The MIN and MAX represent respectively the least and best performance among VOs.

### V.3.2 Performance comparison

The performances of our model are compared with state-of-the-art quality models including: 1) 2D full reference (FR) metrics like PSNR and SSIM, 2) Learning-based NR 2D models such as BRISQUE [159], QAC [186], BPRI [187] and DipIQ [188], 3) PSNR-based 360-degree models WS-PSNR, S-PSNR and CPP-PSNR, and 4) learning-based NR 360-degree metrics SSP-BOIQA [32], and two version of the MC360IQA [90] model, including $_{origin}$ and $_{mean}$ trained respectively without and with data augmentation. The MC360IQA is a multichannel paradigm with six ResNet-34-[87]. Table V.1

**Table V.1:** Performance comparison with state-of-the-art quality models in terms of PLCC and SRCC. Best performance is highlighted in bold.

| | Metric | PLCC | SRCC |
|---|---|---|---|
| FR | PSNR | 0.7662 | 0.7320 |
| | SSIM | 0.8972 | 0.8857 |
| | S-PSNR | 0.7819 | 0.7574 |
| | WS-PSNR | 0.7741 | 0.7467 |
| | CPP-PSNR | 0.7755 | 0.7498 |
| NR | BRISQUE | 0.7641 | 0.7448 |
| | QAC | 0.8681 | 0.8299 |
| | BPRI | 0.8877 | 0.8576 |
| | DipIQ | 0.8065 | 0.7381 |
| Learning-based 360-degree | SSP-BOIQA | 0.9077 | 0.8614 |
| | MC360IQA$_{origin}$ | 0.9271 | 0.9069 |
| | MC360IQA$_{mean}$ | 0.9391 | 0.9153 |
| SP360IQA-F-JND | MIN | 0.900 | 0.866 |
| | MAX | **0.949** | **0.928** |

summarizes the performances of aforementioned metrics on the CVIQ database. We can notice that traditional 2D models and their extended versions have significantly lower performance compared to 360-degree models. Therefore, they are not well suited for this type of image as already demonstrated in benchmark studies [109]. SSP-BOIQA slightly improves the correlation with subjective MOS compared to SSIM that measures the structural similarity according to the HVS characteristics. MC360IQA versions provide good results. At its lowest performance (MIN), our model outperformed all state-of-the-art FR, NR, and 360-degree models except MC360IQA. Regarding the latter, the *origin* version is outperformed by the three versions of the model with the VO providing the maximum performance. The *mean* version is in turn outperformed by the F and F-JND version of the proposed model.

### V.3.3 Ablation study

To evaluate the effectiveness of the considered additional inputs (scan-path visual information and JND maps), we conduct an ablation study. It focuses on performance added to the model by the additional components. First, we predict the quality score using only regressed visual features on 8 viewports extracted based on the virtual observer scan-path denoted as SP360IQA. Second, we add the viewport weight estimation as described in Sec. V.2.2 by encoding scan-path visual information thought an MLP (see Fig. VI.2). This version is denoted as SP360IQA-F. Finally, we optimize the estimation of the weights by exploiting JND probability maps of the selected viewports to account for HVS sensitivity and provide perceptual distortion-ability to the

model, denoted as SP360IQA-F-JND.

**Table V.2:** Standard deviation, maximum and minimum performance in terms of PLCC, SRCC, and RMSE of virtual observers. Best PLCC values are highlighted in bold and SRCC underlined.

| | SP360IQA | | | SP360IQA-F | | | SP360IQA-F-JND | | |
|---|---|---|---|---|---|---|---|---|---|
| | PLCC | SRCC | RMSE | PLCC | SRCC | RMSE | PLCC | SRCC | RMSE |
| MAX ↑ | 0.929 | 0.911 | 0.150 | 0.945 | 0.921 | **0.111** | **0.949** | 0.928 | 0.113 |
| MIN ↑ | 0.780 | 0.750 | 0.090 | 0.889 | 0.863 | 0.084 | **0.900** | 0.866 | **0.080** |
| SD ↓ | ± 0.044 | ± 0.045 | ± 0.017 | ±0.020 | ±0.021 | **±0.009** | **±0.019** | ±0.023 | **±0.009** |

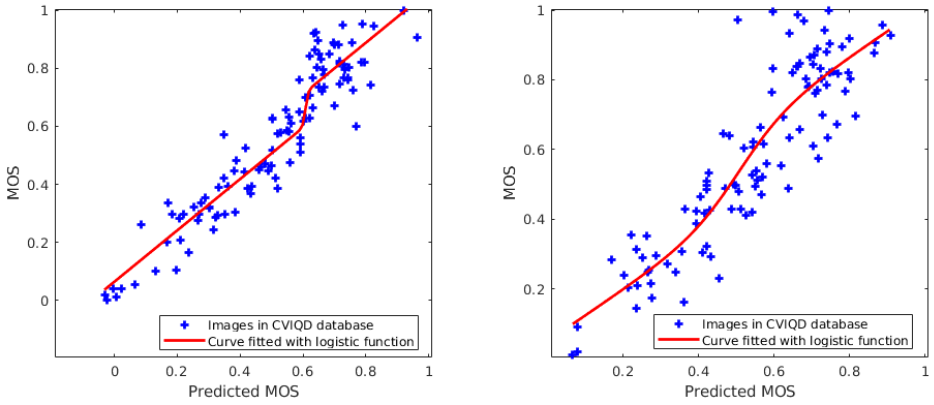
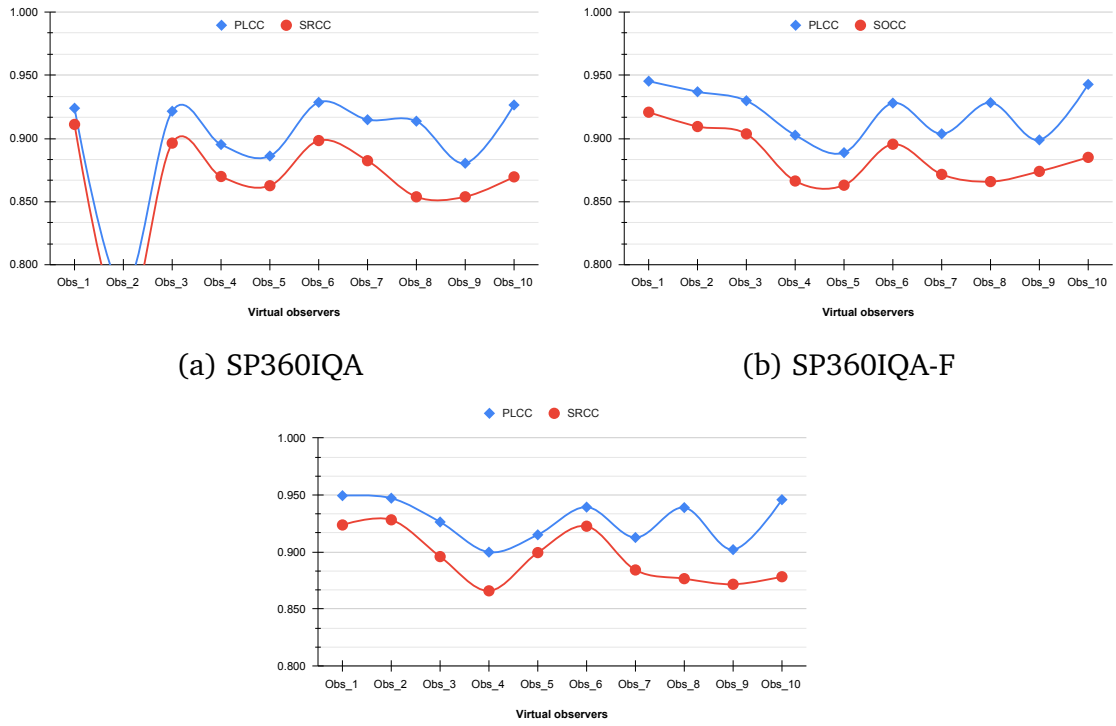



**Figure V.5:** Scatter plots of predicted quality scores versus MOS of the final model SP360IQA-F-JND (Best performance on the left and worst performance on the right among the *VOs*).

Table. V.2 provides the results of the conducted ablation study. The maximum and minimum values of PLCC, SRCC, and RMSE regarding all VOs are given, in addition to standard deviations. One can observe that the proposed weight estimation improves the performance when considering fixations order and duration for each viewport. The minimum PLCC/SRCC shifts from 0.78/0.75 to 0.89/0.86 showing that the model gained significantly in terms of accuracy and monotonicity. The prediction errors measured by the RMSE is also improved by the SP360IQA-F over the SP360IQA one. The incorporation of the JND further boosted the performances but with a slight shift. Therefore, we conclude that using scan-path visual information and JND features contribute to the prediction accuracy of our model. It also contributes to the generalization of our model as given by the SD values. Indeed, the latter are decreased explaining that VOs are providing better and less spread performances. Scatter plots of the predicted scores versus MOS of the best and least performing VO are given in Fig. V.5. It supports the aforementioned discussions and shows consistent distribution of the predictions.




(a) SP360IQA

(b) SP360IQA-F

(c) SP360IQA-F-JND

**Figure V.6:** PLCC and SRCC of individual VOs with regard to the version of the model.

Additionally, Fig. V.6 illustrate the achieved performances in terms of PLCC and SRCC of all VOs with regard to the version of the model. As it can be seen, the performances of the native version without weights estimation lag behind those obtained by the SP360IQA-F and JND versions, regardless of the used VO. This actively demonstrates the usefulness of adaptive weights estimation of each viewport, on the one hand. On the other hand, the incorporation of HVS properties, translated by the visual trajectory behavior and JND, are significantly improving the overall performances of the model. Among the VOs, there is a significant difference between Obs_2 and the others in the SP360IQA version. However, when considering weight estimation, Obs_2 appears to have performed well. Such behavior depicts the accuracy of various scanpaths, and considering only visual features may be dependent on the predictability of the scanpath. Hence, perceptual features are required so as to cope with such lack, as depicted with the provided curves.

## V.4 Conclusion

We presented in this chapter a CNN-based model for 360-degree IQA. This model relies on predicted scan-paths for the extraction of adapted viewports. In addition, to account for the HVS properties, fixations order and duration are used together with

JND to define weighting factors exploited for quality pooling. This adopted weighting strategy has shown a significant improvement of the prediction performances. Additionally, taking advantage of the variability of visual exploration of 360-degree scenes (visual trajectory) through virtual observers, is a significant added value for model generalization analysis. Our model showed the usefulness of predicting quality on the spherical content rather than projected one. We believe that using additional HVS properties may greatly contribute to the improvement of prediction accuracy. So, a more optimized network for learning HVS properties for 360-degree quality assessment tasks will be investigated.

The SP360-IQA is currently under optimization. We intend to optimize the model by adopting the weight sharing among the CNN channels. By doing so, the complexity of the model will be significantly reduced. In addition, the model will learn from all eights input simultaneously, and not each channel from its inputs. The weights are then updated according to all inputs, including the visual features, JND information, and the attributes associated with the visual trajectories.

# Chapter VI

# Attention-aware Patch-based CNN for Blind 360-degree Image Quality Assessment

## VI.1 Introduction

As discussed in the previous chapters, the proposed CNN-based 360-IQA models adopt the multichannel CNN paradigm. As the latter increases significantly the complexity of the model with regards to the number of channels, it becomes of paramount important to achieve a trade-off between complexity and robustness. In the meantime, patch-based CNNs offer good performances while being less complex. The fact of having a single channel CNN instead of multichannel one helps training the model more efficiently with less computational requirement. Besides, patch-based training implies patch sampling prior to training. The latter are used as data augmentation to cope with the lack of training data for IQA tasks, as highlighted in Chapter IV. With a proper, adaptive, and consistent sampling strategy, the robustness of the model can be significantly enhanced [126].

To address several issues related to 360-IQA with deep neural networks, we propose in this chapter the use of deep CNN with a patch-based training scheme. The proposed model is trained and evaluated from scratch on three available 360-IQA databases. In order to resolve the lack of perceptually annotated data, we artificially augment the available training sets by training the model on sampled patches obtained from the 360-degree images. Here, an adaptive sampling is proposed by considering the importance of the content according to its location. Hence, a latitude-based sampling from the radial content is adopted rather than a random sampling on the projected representation. By doing so, we incorporate the exploration behavior and the way human gaze is biased toward the equatorial region [133]. Prior to training the model, the extracted patches are normalized using a local contrast based normalization to only retain information that is perceptually relevant to the HVS, and to speed up the training process. Then, an adaptive aggregation strategy for pooling

local patch qualities to 360-degree image quality is employed. The main contributions within the proposed model are listed as follows:

- Inspired by the exploration behavior of 360-degree images by the users, we propose an adaptive patch sampling on the sphere to tackle the issue of geometrically distorted content of the projected versions. The ablation analysis proves its superiority over standard sampling.
- We design a CNN model with spatial attention integrated to efficiently learn weight maps, which represent the relative importance of activations within feature maps. Furthermore, we incorporate features from the earliest layers at the final stage through long skip-connections.
- We present a method for aggregating patch quality to image quality using outlier rejection and saliency. To begin, the standard deviation is utilized to exclude scores that fall beyond an agreement range. The saliency is then used to weight the selected local qualities based on their visual relevance. The relevance of the associated content is considered using a weighted average pooling of local patch qualities.

## VI.2 The proposed model

The proposed method involves several steps. 1) data pre-processing, including patch sampling and representation, 2) an end-to-end training, and 3) patch predicted qualities aggregation to image quality by means of an adaptive strategy. Details on each step are given in the following.

### VI.2.1 Inputs generation

Regions surrounding the equator of 360-degree images are known to be more visually important than polar regions [133], on the one hand. On the other hand, adaptive sampling is proven to be more effective than standard sampling, as demonstrated in Chapter IV. Therefore, we adopt the sampling strategy described in Sec. IV.3 to generate non-overlapping patches from the sphere.

### VI.2.2 Patches normalization

Adaptive data representation is of paramount importance for any machine learning algorithm. For CNNs, image normalization prior to feeding the model boosts the training and helps the model to learn the useful information for the specified tasks. Based on the conclusions from the data representation study described in Chapter IV, we adopt the LCN normalization to normalize the sampled patches. The normalization is performed on individual patches rather than the full 360-degree images. The particular purpose for this, is to account for local luminance, which might vary in different parts of the image. It is worth noting that, 360-degree images are in fact multiple snapshots

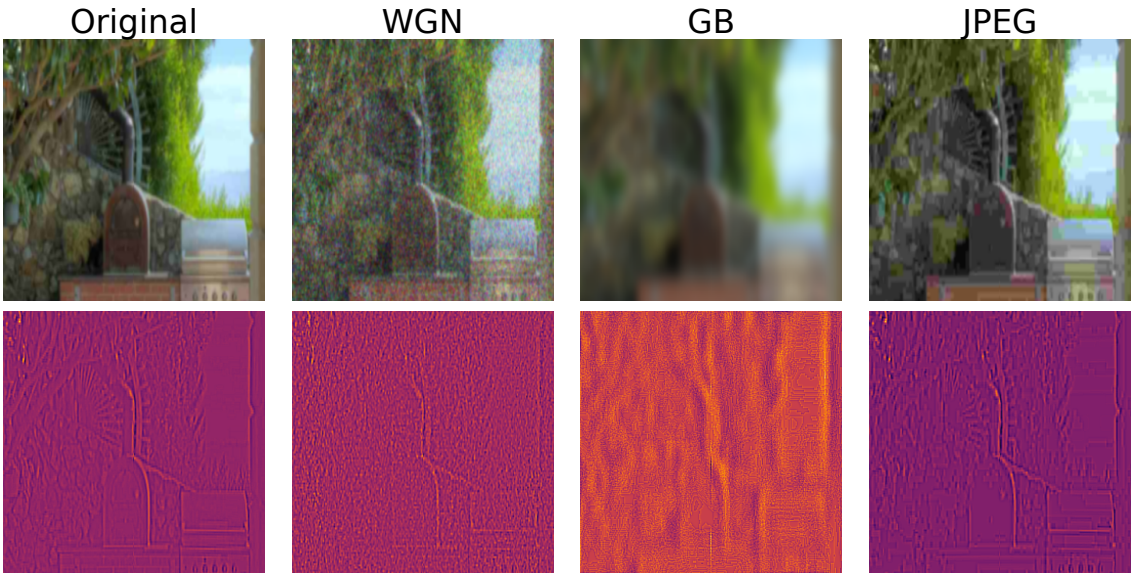



**Figure VI.1:** Local normalization applied on patches with different distortions. (Top) extracted patches, (bottom) normalized versions.

taken by multiple cameras and then stitched together. The setting of the cameras is made in such a way to cover a 360-degree range even if we consider that the stitched images has been corrected in terms of luminance uniformity. Consequently, the range of the latter may vary spatially.

The proposed model takes locally contrast normalized patches as input. For each value of pixel $(i, j)$ from a sampled patch $P$, the corresponding normalized value $P'(i, j)$ is computed using Eq. IV.5 described in Sec. IV.4.2. The visual results of this step are illustrated in Fig. VI.1. As it can be seen, high-frequency details are preserved, as they are more likely to influence the quality rating. Additionally, the normalization appears to capture the effects of distortions, which may help to ensure the robustness of the model.

### VI.2.3 Model architecture

The architecture of the proposed model is illustrated in Fig. VI.2. In contrast to state-of-the-art quality models for 360-degree images, the proposed model adopts the patch-wise training. Therefore, an architecture with a deep CNN is designed rather than a multichannel one. The latter is found to be highly complex, requiring more computational resources and is difficult to train. The proposed model is composed of four convolutional blocks (Conv Block) with a doubling number of filters, ranging from 64 to 512. This way, the CNN model can learn more discriminative features and be able to achieve a better representation of these features [136]. Each block is composed of two Conv layers with $3 \times 3$ filter size, each followed by a batch normalization (BN) layer [189], and then a ReLU activation function. The structure of the Conv Block is

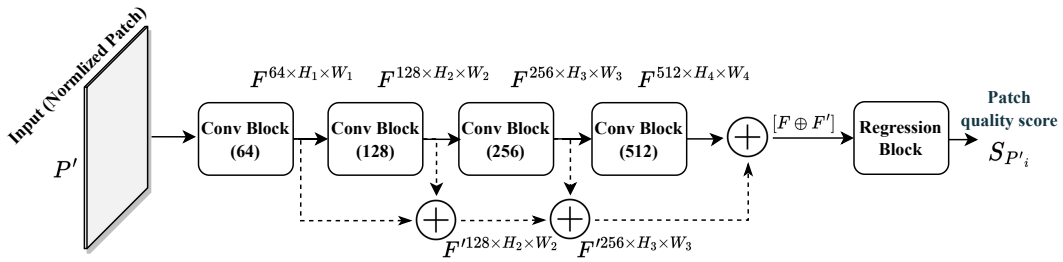



**Figure VI.2:** Architecture of the proposed model. $F$ and $F'$ stand for feature maps, and $S_{P'_i}$ represents the predicted quality score associated with patch $P'$.

depicted in Fig. VI.3. BN is used to make the model faster, stable and more robust to bad initialization [189]. It is recommended to place BN right after the Conv layers and before the activation function, which helps to produce activations with a stable distribution [189]. All convolutions are used with zero-padding to preserve more features and produce an output of the same dimension as the input.

Right after the second activation (ReLU) in each Conv Block, a spatial attention (SPA) module is used. The latter outputs a refined feature maps $\mathbf{F} \in \mathbb{R}^{D_n\times H_n\times W_n}$, where, $D$, $H$, and $W$ stand for the height, width, and dimension in terms of number of channels of the deep features $\mathbf{F}$, respectively, and $n$ corresponds to the number of Conv Blocks. Here, The *spatial* term refers to the domain space encapsulated within each feature map. Therefore, the spatial attention represents the attention mechanism, *a.k.a.* attention mask on the learned feature maps. It conveys "*what*" within each feature map is important to learn and to focus on.

The composition of the SPA module is detailed in Fig. VI.3 (blue rectangle). It comprises a three-fold sequential operation. The first consists of a pooling stage in order to encode and capture highly discriminative features before applying the attention. Commonly, a max- or average-pooling is applied [96] at this stage, and in some cases they are used together as in [190]. Differently, the proposed SPA module applies the generalized mean pooling (GeM) [191], which generalizes the pooling equation for each feature map $X_D$ of $\mathbf{F} \in \mathbb{R}^{D\times H\times W}$ as follows:

$$y = \left(\frac{1}{|X_D^s|}\sum_{x\in X^s} x^{p_D}\right)^{\frac{1}{p_D}}, \tag{VI.1}$$

where $y$ represents the aggregated value, $X_D$ the set of values, $s$ the pooling stride of $2\times2$, and $p_D \in [1,\infty)$ is the hyperparameter that controls the pooling for each feature map $X_D$. When $p_D = 1$, the GeM corresponds to average pooling. When $p_D \to \infty$, the GeM corresponds to max pooling. As the GeM pooling is a differentiable operation as stated in [191, 192], $p_D$ can be considered as a trainable parameter. This helps updating it through backpropagation. Hence, the SPA module exploits the states between average and max pooling by relying on backpropagation to learn the hyperparameter

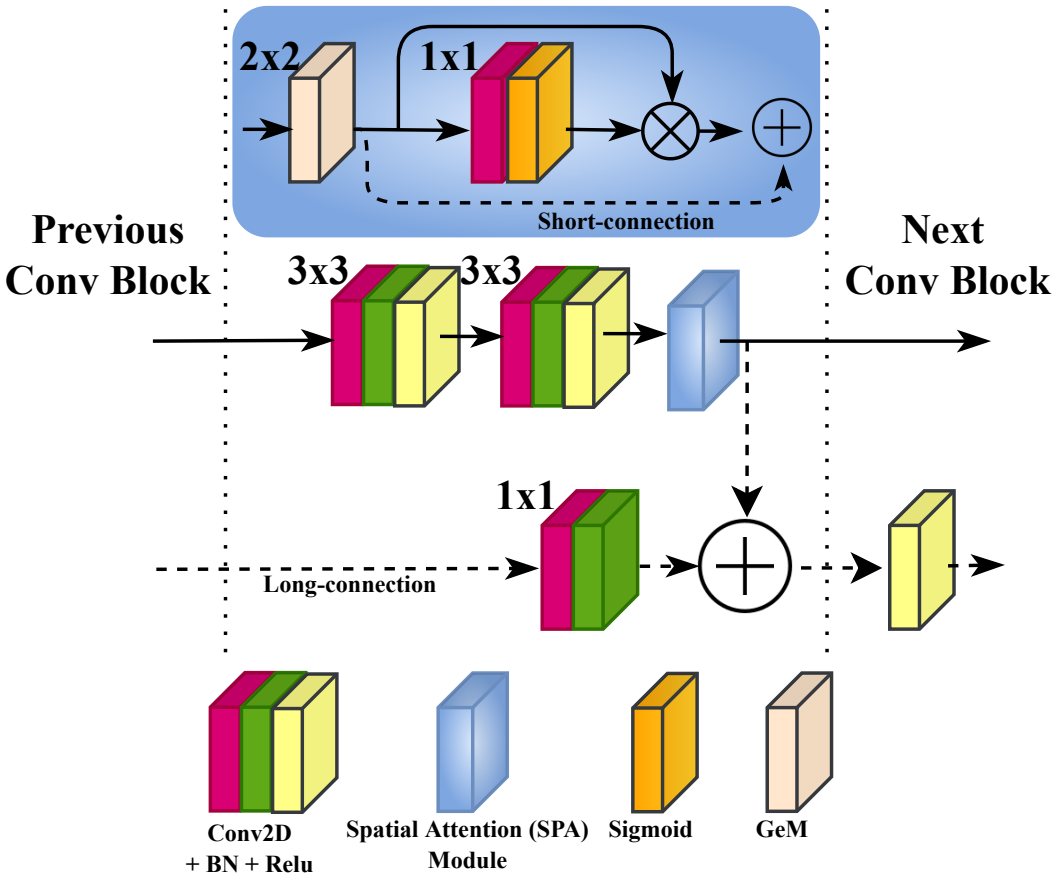



**Figure VI.3:** Architecture of the Conv Block with ⊗ element-wise multiplication and ⊕ element-wise addition. The 3 × 3 and 1 × 1 correspond to the kernel sizes for the convolution layers, and 2 × 2 represents the stride for the pooling layer (GeM).

$p_D$, allowing it to use the adequate pooling method. The second part consists of creating the spatial attention mask $M_{SPA}(\mathbf{F}) \in \mathbb{R}^{H \times W}$ by means of the sigmoid activation function $\sigma(.)$ as follows:

$$M_{SPA}(\mathbf{F}) = \sigma(Conv^{1\times1}([GeM^{2\times2}(\mathbf{F})])) \tag{VI.2}$$

$M_{SPA}(\mathbf{F}) \in \mathbb{R}^{H \times W}$ is generated using the aggregated maps obtained by $GeM(.)$. First, $\mathbf{F} \in \mathbb{R}^{D \times H \times W}$ goes through a Conv layer with $1 \times 1$ kernel size and a single filter, resulting in $\mathbf{F} \in \mathbb{R}^{1 \times H \times W}$. The idea is to reduce the dimensionality of the generated feature maps by creating a single-channel one. Then, the sigmoid activation function $\sigma(.)$ is applied on the output of the Conv layer. As sigmoid is a probabilistic activation function, it will map all the values of the input $\mathbf{F} \in \mathbb{R}^{1 \times H \times W}$ to a range between 0 and 1, generating the spatial attention mask $M_{SPA} \in \mathbb{R}^{H \times W}$. The latter is then applied to all the feature maps in the output of $GeM(.)$ using element-wise product as denoted

in Eq.VI.3

$$\mathbf{F}_{SPA} \in \mathbb{R}^{D\times H\times W} \quad = \quad M_{SPA} \otimes GeM^{2\times 2}(\mathbf{F}) \tag{VI.3}$$

Within each SPA module, a short skip-connection is implemented, connecting the output of $GeM(.)$ with the refined feature map $\mathbf{F}_{SPA} \in \mathbb{R}^{D\times H\times W}$. Here, the connection is accomplished by means of element-wise addition, as illustrated in Fig. VI.3. As a result, the aligned features will be greater compared to non-aligned ones. The output of the SPA module is then obtained as follows:

$$\mathbf{F} \in \mathbb{R}^{D\times H\times W} \quad = \quad ReLU(\mathbf{F}_{SPA} \oplus GeM^{2\times 2}(\mathbf{F})) \tag{VI.4}$$

The output of the SPA module is first fed to the next Conv Block in a feed forward fashion. Additionally, it is acknowledged that, the earliest convolutions in a CNN capture low-level features, whereas latter convolutions focus on high-level semantic features, on the one hand. On the other hand, the HVS is highly sensitive to low-level features, such as spatial frequency, line orientation, texture and contrast [193–195]. The use of such features at later stages in CNN could have the potential to improve the model's performances and stability for various image processing tasks [90, 196]. Therefore, the proposed model implements a long skip-connection by means of hierarchical element-wise additions at each Conv Block. Here, the $\mathbf{F} \in \mathbb{R}^{D\times H\times W}$ are hierarchically added together into $\mathbf{F}' \in \mathbb{R}^{D\times H\times W}$ as follows:

$$\mathbf{F}'_i \quad = \quad ReLU(\mathbf{F}'_i \oplus Conv^{1\times 1}_{filter=filter_i}(\mathbf{F}_{i-1})), \tag{VI.5}$$

where $i$ denotes the Conv Block number. The $Conv^{1\times 1}$ with the same filter number as the block $i$ is required to match the dimension for the addition operation. At the last Conv Block, both $\mathbf{F}$ and $\mathbf{F}'$ are added together. Finally, GAP [124] is used to reduce the spatial dimensions of the encoded feature maps by generating a feature vector $V_{\mathbf{F}} \in \mathbb{R}^{D\times 1\times 1}$. This operation is known to decrease overfitting, and is accomplished as follows:

$$y^c \quad = \quad \frac{1}{N}\sum_{i,j}[\mathbf{F} \oplus \mathbf{F'}]^c_{i,j}, \tag{VI.6}$$

where $y^c$ is the output value of feature map $[\mathbf{F} \oplus \mathbf{F'}]^c$ at channel $c$, and $(i, j)$ is the pixel index of the corresponding feature value. The feature vector $V_{\mathbf{F}}$ obtained by the GAP operation is fed to the quality regression block, where the obtained features are fused to estimate the quality score. The architecture of the regression block is illustrated in Fig. VI.4 and is composed of two FC layers with dimensions of 1024 and 512, respectively, each followed by a ReLU activation function and a dropout layer

for regularization. A final FC layer with a single node and a linear activation is added to deliver the final quality score. Weights initialization in the model is performed according to He *et al.* [127] to start the training with a Gaussian probability distribution initialization. The latter helps the model to avoid numerical difficulties due to unstable initial weights [126].

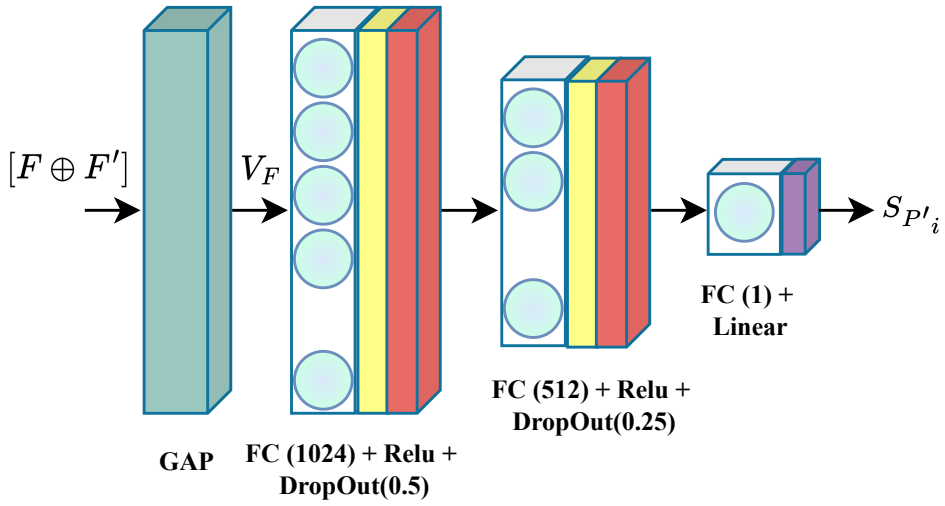



**Figure VI.4:** Architecture of the used regression block. "GAP" corresponds to global average pooling, and "FC" stands for a fully connected layer. $V_F$ is the generated feature vector.

### VI.2.4 Loss function

Generally, IQA is considered as a regression task. Hence, the most commonly used loss functions are the mean square error (MSE) and mean absolute error (MAE). It is known that MAE is less sensitive to outliers in the data, but it is not differentiable at zero. Whereas, MSE is differentiable everywhere, but it is highly sensitive to outliers. The Huber loss [197] is both differentiable everywhere and robust to outliers. It combines the best characteristics of both loss functions. Therefore, the proposed model uses the Huber loss, defined as follows:

$$L_\delta(y, S_{P'_i}) = \begin{cases} \frac{1}{2}(y - S_{P'_i})^2 & \text{for } \left|y - S_{P'_i}\right| \leq \delta, \\ \delta \left|y - S_{P'_i}\right| - \frac{1}{2}\delta^2 & \text{otherwise,} \end{cases} \tag{VI.7}$$

where $y$ is the ground truth score (MOS) and $S_{P'_i}$ is the predicted one. $\delta \in \mathbb{R}^+$ controls the use of either MAE or MSE. Its value is defined as 1.35 based on [197].

### VI.2.5 Quality scores aggregation

As a patch-based strategy, the proposed model predicts a set of quality scores *LQs* corresponding to $N$ sampled patches from each 360-degree image $I$. To drive the quality of the whole 360-degree image, one can apply a simple and straightforward average of *LQs*. However, as demonstrated in Chapter IV, the quality over a scene is non-uniformly distributed, meaning that certain regions are more likely than others to contribute to the global quality. This is even more true for 360-degree images. Moreover, the perceived quality is highly affected by the most distorted regions among the selected ones [162], as the human gaze tends to fall on these salient regions when exploring a scene. A simple arithmetic mean cannot express such aspects, notwithstanding the non-uniformly distribution of the quality. Giving the same importance to all local qualities by averaging *LQs* may not be consistent with (i) the scene exploration and (ii) the quality distribution.




**Figure VI.5:** Illustration of saliency map to highlight the important regions. (a) a 360-degree image, (b) its saliency, and (c) their superposition.

In response to the limitation of simply averaging *LQs*, the proposed model uses visual saliency to weights local qualities in *LQs*. In this regard, it considers the relevance of salient regions over less salient ones by means of visual attention. Hence, for each 360-degree image $I$, a saliency map $Sal_I$ is generated using the 360-degree saliency model proposed by Xia *et al.* [198]. Fig. VI.5 illustrates the obtained saliency map of a 360-degree image. As it is shown, important regions are highlighted and are mostly located around the equator. By taking the influence of saliency into consideration, patches in salient regions contribute significantly to the global quality score with high weight values. $S_I$ is then computed by weighting the predicted local qualities in *LQs* for patches sampled from the image $I$ using the estimated weights $W_i$ with $i \in \{1, 2, ..., N\}$, and $N$ is the number of patches.

$$S_I = \frac{\sum_{i=1}^{N} W_i \times S_{P_i'}}{\sum_{i=1}^{N} W_i}, \tag{VI.8}$$

where the weights are obtained as follows:

$$W_i = \frac{\sum_{R_P} Sal_{P_i}(k)^{\text{ if } Sal_{P_i}(k) \geq th}}{\sum_{R_I} Sal_I(k)}, \tag{VI.9}$$

$Sal_{P_i}$ represents the saliency region corresponding to the extracted patch $P_i$, $R_P$ and $R_I$ represents the resolution of $P$ and $I$, respectively. Differently from standard approaches where the weights are the summation of pixel intensities within a patch [176], we compute the ratio with respect to the overall saliency of the image $I$. By doing so, the weights will reflect the importance of the local quality of $P$ with regard to the global quality of $I$. Besides, only values greater than a threshold $th$ are summed together, taking into consideration only higher saliency values. $th$ is a tunable parameter according to the predicted scores as well as the used databases.

Prior to weighting local qualities by the corresponding visual saliency, outliers rejection (OR) from the $LQs$ is applied. Scores falling far from the median of $LQs$ are discarded, aggregating only those within an agreement interval. This technique is motivated by subjective quality ratings, in which only scores that concur are considered to generate the MOS. The outliers are detected using the standard deviation $\sigma(.)$ of $LQs$. In this case, a score is identified as an outlier only if it falls outside a $\pm\lambda * \sigma(LQs)$ range, where $\lambda \in \mathbb{R}^+$ is a parameter used to determine the appropriate agreement range with respect to the variability among $LQs$. This operation is described in Algorithm 3. Formally, the final score $S_I$ is obtained as follows:

$$S_I = \frac{\sum_{i \in |s|} W_i \times S_{p_i}}{\sum_{i \in |s|} W_i}, \quad \text{with} \quad s = \pm\lambda * \sigma(LQs) \tag{VI.10}$$

## VI.3 Experiments and results

### VI.3.1 Experimental setup

#### VI.3.1.1 Datasets

The training and evaluation of the proposed model is carried out on three publicly available 360-IQA datasets namely OIQA [108], CVIQ [90], and MVAQD [32]. Details regarding each one are provided in Sec. II.2. We conduct the performance comparison with SOTA models on the OIQA and CVIQ databases. As for an in-depth analysis of the proposed model with ablation studies, we include MVAQD along with OIQA and CVIQ.

**Algorithm 3** Outliers rejection

**procedure** DETECTOUTLIER($LQs$, $W$)
  $V_LQs = \emptyset$ ▷ List of valid local qualities.
  $V_W = \emptyset$ ▷ List of corresponding weights.
  $s = \pm\lambda * \sigma(LQs)$ ▷ Agreement range.
  **for** $i$, $S_{P'_i}$ in enumerate($LQs$) **do**
    **if** $S_{P'_i} \in s$ **then**
      $V_LQs \leftarrow V_LQs + S_{P'_i}$
      $V_W \leftarrow V_W + W_i$
    **else**
      Reject $S_{P'_i}$
    **end if**
  **end for**
  Return $V_LQs$ and $V_W$
**end procedure**

#### VI.3.1.2 Implementation Details

The proposed model is implemented using TensorFlow and trained on a server equipped with Intel Xeon Silver 4208 2.1GHz, 192G RAM and a GPU Nvidia Telsa V100S 32G. The batch size was set to 32 and the Adam optimizer [181] is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set initially to $1e-3$ with a learning decay of $^{1e-4}/_{EPOCHS}$ to help the optimization of the model. A five-fold cross validation is adopted for a complete assessment within each of the selected datasets. Each fold was trained for 100 epochs. During training, the datasets are randomly separated into 80% for training (20% from which was dedicated for validation) and 20% for testing. To ensure a complete separation of the training and testing sets, the distorted images linked to the same pristine source are allocated to the same set. This helps it possible to test the model on unseen content, avoiding a very common mistake that is frequently overlooked, in which datasets are split based on distorted images. Therefore, the model's assessment is made on already seen content, resulting in unreliable assessments.

#### VI.3.1.3 Evaluation metrics

The evaluation criteria recommended by the VQEG, including PLCC, SRCC, and RMSE are used for the evaluation in addition to the statistical significance method proposed by Krasula *et al.* [45]. As explained in Sec. I.1.4, the statistical significance analysis requires the SD of the MOS. Unfortunately, we were only able to collect the required data for the OIQA dataset during the study. Thus, we only perform this analysis with OIQA.

As one of the main motivation behind the development of the proposed model is to achieve better accuracy while having less complexity compared to state-of-the-art

models in general and multichannel ones in particular, we performed an inference analysis. Hence, a complexity analysis is provided in terms of model's parameters and the number of floating-point operations (FLOPs). The latter provides insight on the computations required by the model. A large number of FLOPs implies a higher complexity, suggesting a longer calculation time. Since the inference analysis is independent of the training, we used a different hardware configuration. A computer equipped with an Intel® Core™ i9-9880H @ 2.30GHz, 32GB of RAM, and an Nvidia Quadro T2000 MAX-Q 4GB GPU is used to measure the computational complexity.

### VI.3.2 Performance comparison with SOTA models

With the aim to illustrate the effectiveness of the proposed model, a comparison with sixteen state-of-the-art IQA models, including 2D-IQA and 360-IQA models is performed. For each category, traditional and deep-learning (DL) based models are selected. Traditional models consist of those based on pixel-wise fidelity or NSS. The selected models include PSNR, SSIM [72], MS-SSIM [199], FSIM [195], BRISQUE [159], and BMPRI [200], DB-CNN [101], and DipIQ [188], representing 2D-IQA models. S-PSNR [201], WS-PSNR [201], SSP-BOIQA [32], Yun *et al.* [97], and MFILGN [98], MC360IQA [90], Zhou *et al.* [92], VGCN [100], and AHGCN [94], representing 360-IQA models. The MC360IQA, Zhou *et al.*, VGCN, and AHGCN are all deep learning based solutions using the multichannel paradigm with a varying number of channels, from six to twenty channels. Hence, they are considered as highly complex models.

The overall and per individual distortion performances in terms of PLCC, SRCC, and RMSE are summarized in Table VI.1 for OIQA and Table VI.2 for CVIQ, where the best and second-best performances are highlighted in bold and underlined, respectively. The performances of the proposed model are reported as the median over five-folds evaluation.

From the performances obtained on OIQA (Table VI.1), we observe that the proposed model achieves the best performance on the overall database compared to both 360-IQA and 2D-IQA models. This observation is valid regardless of the used local qualities' aggregation method, demonstrating its superiority over SOTA models. In particular, multichannel based models, where the proposed model outperformed the MC360IQA in terms of PLCC (resp. SRCC) by approx. 5% (resp. 4.7%), Zhou *et al.* by approx. 8% (resp. 4.2%), VGCN by approx. 1.4% (resp. 1.2%), and AHGCN by approx. 0.7% (resp. 0.4%). A similar behavior can be observed with the prediction error in terms of RMSE. This actively illustrates the accuracy of our model and its ability to evaluate the quality of 360-degree images close to human judgment. Regarding the performances across different type of distortions, the SOTA models still lags behind the performances of the proposed model for JPEG and GB. For JP2K and WGN, the proposed model achieved competitive performances compared to VGCN, AHGCN, and Yun *et al.* in terms of PLCC and SRCC. In terms of RMSE, however, the proposed model achieved the lowest error across all distortions as well as the overall database. The superiority of the proposed model is highlighted once more. When

**Table VI.1:** Performance comparison with SOTA models on the OIQA database. The Best performance is highlighted in **bold** and the second-best is underlined. $\text{Ours}_{\text{Avg}}$, $\text{Our}_{\text{OR}}$, and $\text{Our}_{\text{OR+SAL}}$ stand for the proposed model with average, OR-based, and OR + saliency based aggregation of local qualities, respectively.

| | | | Overall | | | JPEG | | | JP2K | | | WGN | | | GB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ref- | Model | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ |
| 2D-IQA | FR | PSNR | 0.6910 | 0.6802 | 10.388 | 0.8658 | 0.8291 | 7.8570 | 0.8492 | 0.8421 | 7.9357 | 0.9317 | 0.9008 | 4.6392 | 0.6357 | 0.6374 | 10.250 |
| | FR | SSIM | 0.8892 | 0.8798 | 6.5814 | 0.9409 | 0.9346 | 5.3193 | 0.9336 | 0.9357 | 5.3829 | 0.9026 | 0.8846 | 5.4965 | 0.9188 | 0.9238 | 5.2404 |
| | FR | MS-SSIM | 0.8427 | 0.8332 | 7.7442 | 0.9312 | 0.9188 | 5.7214 | 0.9265 | 0.9267 | 5.6560 | 0.9672 | 0.9484 | 3.2460 | 0.8623 | 0.8624 | 6.7250 |
| | FR | FSIM | 0.9274 | 0.9225 | 8.2501 | 0.9478 | 0.9351 | 7.3215 | 0.9545 | 0.9573 | 6.5924 | 0.9466 | 0.9176 | 6.0670 | 0.9444 | 0.9478 | 6.4372 |
| | NR | BRISQUE | 0.8424 | 0.8331 | 11.261 | 0.9160 | 0.9392 | 8.9920 | 0.7397 | 0.6750 | 15.082 | 0.9553 | 0.9372 | 3.4270 | 0.8663 | 0.8508 | 9.6970 |
| | NR | BMPRI | 0.6503 | 0.6238 | 15.874 | 0.9160 | 0.8954 | 7.8861 | 0.8322 | 0.8214 | 12.280 | 0.9611 | 0.9490 | 3.5340 | 0.5199 | 0.3807 | 12.248 |
| | NR | DB-CNN | 0.8852 | 0.8653 | 9.7172 | 0.9755 | **0.9607** | 4.9350 | 0.9770 | **0.9786** | 3.8324 | 0.9772 | **0.9786** | 3.8323 | 0.9536 | 0.8865 | 5.8752 |
| | NR | DipIQ | 0.7012 | 0.6917 | 10.259 | 0.8291 | 0.7891 | 8.7833 | 0.9165 | 0.9182 | 6.0300 | 0.9556 | 0.9432 | 3.7742 | 0.9321 | 0.8983 | 4.8161 |
| 360-IQA | FR | S-PSNR | 0.7153 | 0.7115 | 10.052 | 0.8703 | 0.8285 | 7.7319 | 0.8555 | 0.8489 | 7.7811 | 0.9190 | 0.8846 | 5.0329 | 0.6929 | 0.6917 | 9.5736 |
| | FR | WS-PSNR | 0.6985 | 0.6932 | 10.294 | 0.8607 | 0.8278 | 7.9919 | 0.8435 | 0.8322 | 8.0719 | 0.9221 | 0.8853 | 4.9415 | 0.6609 | 0.6583 | 9.9652 |
| | NR | SSP-BOIQA | 0.8600 | 0.8650 | 7.3131 | 0.8772 | 0.8345 | 7.6201 | 0.8532 | 0.8522 | 7.5013 | 0.9054 | 0.8434 | 5.4510 | 0.8544 | 0.8623 | 6.8342 |
| | NR | Yun *et al.* | 0.9437 | 0.9369 | 7.1911 | 0.9612 | 0.9536 | 6.3330 | 0.9697 | 0.9676 | 5.3941 | 0.9789 | 0.9737 | 3.8453 | 0.9645 | 0.9558 | 5.1582 |
| | NR | MC360IQA | 0.9247 | 0.9187 | 4.6247 | 0.9279 | 0.9190 | 4.5058 | 0.9324 | 0.9252 | 4.5825 | 0.9344 | 0.9345 | 3.7908 | 0.9220 | 0.9353 | 4.5256 |
| | NR | Zhou *et al.* | 0.8991 | 0.9232 | 6.3963 | 0.9363 | 0.9405 | 5.6911 | 0.9200 | 0.9343 | 5.8862 | 0.9682 | 0.9570 | 3.3304 | 0.9252 | 0.9200 | 4.9721 |
| | NR | VGCN | 0.9584 | 0.9515 | 5.9670 | 0.9540 | 0.9294 | 6.7201 | 0.9771 | 0.9464 | 4.7721 | **0.9811** | 0.9750 | 3.4932 | 0.9852 | 0.9651 | 3.3270 |
| | NR | AHGCN | 0.9649 | 0.9590 | 5.4871 | 0.9649 | 0.9276 | 5.8860 | **0.9820** | 0.9643 | 4.2360 | 0.9706 | **0.9786** | 4.3410 | 0.9756 | 0.9759 | 4.2640 |
| | NR | $\text{Ours}_{\text{Avg}}$ | 0.9668 | 0.9585 | 3.5707 | 0.9741 | 0.9429 | 3.5956 | 0.9731 | 0.9607 | **3.4249** | 0.9755 | 0.9643 | 2.8710 | **0.9880** | **0.9786** | **2.1828** |
| | NR | $\text{Ours}_{\text{OR}}$ | 0.9709 | 0.9626 | 3.3538 | 0.9763 | 0.9571 | 3.4221 | 0.9737 | 0.9536 | 3.3843 | 0.9761 | 0.9607 | 2.8373 | 0.9865 | **0.9786** | 2.2843 |
| | | $\text{Ours}_{\text{OR+SAL}}$ | **0.9721** | **0.9629** | **3.3317** | **0.9779** | 0.9500 | **3.3125** | 0.9741 | 0.9536 | 3.3587 | 0.9764 | 0.9714 | **2.7005** | 0.9866 | **0.9786** | 2.2388 |

looking at the performance of 2D-IQA models, it is clear that DB-CNN outperforms 2D traditional models, including FR models like SSIM and MS-SSIM. This demonstrates the advantages of a deep learning-based model for quality assessment over traditional approaches.

Among the performances of the proposed model, the best is achieved when using OR and saliency together as the aggregation strategy. It appears that, discarding predicted local qualities that are outside the agreement range $\pm\lambda * \sigma(LQs)$ before averaging the scores is improving the correlation performances. This is demonstrated by the acquired performances on the overall database and across distortions compared to considering all local qualities, except for GB. By simply averaging all local qualities resulted in slightly better correlation performances for GB. Such a behavior can be explained by the fact that the model agrees more on the quality of blurred patches compared to other distortions. When using saliency to weight predicted local qualities according to the importance of their corresponding patches after discarding outliers, a similar behavior is observed. It is known that, saliency is affected by the distortion in general and blur in particular, since the content is smoothed. Hence, using saliency as a weighting strategy is not contributing significantly with the GB distortion.

Table VI.2 further summarizes the performances results of the proposed model as well as SOTA models on the overall and on each distortion type of CVIQ. The first observation that emerges is that on overall, the best performances are obtained by deep learning based 360-IQA models. Considering all content on CVIQ, the proposed model achieved slightly worse in terms of accuracy compared to VGCN while outperforming MC360IQA, Zhou *et al.*, and AHGCN. In terms of monotonicity, our model achieved the best correlation. As for the prediction errors, the MC360IQA scored the best, with a slight difference. With regard to the performances obtained on OIQA (Table VI.1), those ones on CVIQ seems to be slightly worse when compared to SOTA models. Here, the diversity of the training dataset is affecting the accuracy of the model. It is worth noting that CVIQ is less diverse compared to OIQA in terms of (i) distortion types and (ii) content, which are paramount to better train the model. Regarding the performances across the different distortion types, the proposed model reached competitive performances with multichannel ones. Despite less diversity on CVIQ, the proposed model still perform well. For the JPEG distortion, some 2D models appear to perform well, such as SSIM and FSIM. The latter obtained the best accuracy with images compressed using H.265/HEVC. Such an achievement is due to the fact that these models have access to pristine images.

The proposed model's correlation performance appears to be influenced by the aggregation strategy. Except for H.265/HEVC distortion, the average solution resulted in the best outcomes, making adaptive pooling less effective on CVIQ. This is closely tied to the nature of the content available on CVIQ. As the OR-based aggregation decreased the performances, it implies that the predicted local qualities are mostly close to the median. Therefore, less diversity exist among *LQs*.

**Table VI.2:** Performance comparison with SOTA models on the CVIQ database. The Best performance is highlighted in **bold** and the second-best is underlined. $\text{Ours}_{\text{Avg}}$, $\text{Our}_{\text{OR}}$, and $\text{Our}_{\text{OR+SAL}}$ stand for the proposed model with average, OR-based, and OR + saliency based aggregation of local qualities, respectively.

| | | | Overall | | | JPEG | | | H.264/AVC | | | H.265/HEVC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ref- | Model | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ |
| 2D-IQA | FR | PSNR | 0.7320 | 0.7662 | 9.0397 | 0.7342 | 0.8643 | 8.5866 | 0.7572 | 0.7592 | 8.0448 | 0.7169 | 0.7215 | 8.3279 |
| | FR | SSIM | 0.8857 | 0.8972 | 6.2140 | 0.9334 | 0.9749 | 3.7986 | 0.9451 | 0.9457 | 4.0165 | 0.9220 | 0.9232 | 4.6219 |
| | FR | MS-SSIM | 0.8762 | 0.8875 | 6.4836 | 0.9140 | 0.9628 | 4.6101 | 0.8794 | 0.8805 | 5.8583 | 0.8604 | 0.8610 | 6.1165 |
| | FR | FSIM | 0.934 | 0.9152 | 4.8964 | 0.9839 | 0.6939 | 2.8928 | 0.9534 | 0.9439 | 4.0327 | **0.9617** | 0.9532 | 3.9239 |
| | NR | BRISQUE | 0.7448 | 0.7641 | 9.0751 | 0.8489 | 0.9091 | 7.1137 | 0.7193 | 0.7294 | 8.4558 | 0.7151 | 0.7104 | 8.4646 |
| | NR | BMPRI | 0.7919 | 0.7470 | 8.5258 | 0.9874 | 0.9562 | 2.5597 | 0.7161 | 0.6731 | 9.3318 | 0.6154 | 0.6715 | 9.3071 |
| | NR | DB-CNN | 0.9356 | 0.9308 | 4.9311 | 0.9775 | 0.9576 | 3.3862 | 0.9564 | 0.9545 | 3.9063 | 8.6460 | 0.8693 | 5.3350 |
| | NR | DipIQ | 0.7060 | 0.6232 | 9.9601 | 0.9285 | 0.7930 | 6.3530 | 0.6203 | 0.6353 | 9.6954 | 0.3611 | 0.6366 | 11.216 |
| 360-IQA | FR | S-PSNR | 0.7467 | 0.7741 | 8.9066 | 0.7520 | 0.8772 | 8.1974 | 0.7690 | 0.7748 | 7.8743 | 0.7389 | 0.7428 | 8.0515 |
| | FR | WS-PSNR | 0.7498 | 0.7755 | 8.8816 | 0.7604 | 0.8802 | 8.1019 | 0.7726 | 0.7748 | 7.8143 | 0.7430 | 0.7469 | 7.9974 |
| | NR | SSP-BOIQA | 0.8900 | 0.8561 | 6.9414 | 0.9155 | 0.8533 | 6.8471 | 0.8850 | 0.8611 | 7.0422 | 0.8544 | 0.8410 | 6.3020 |
| | NR | Yun *et al.* | 0.9481 | 0.9344 | 4.4964 | 0.9844 | 0.9650 | 2.9910 | 0.9576 | 0.9545 | 3.5623 | 0.9289 | 0.9278 | 4.4534 |
| | NR | MC360IQA | 0.9506 | 0.9139 | **3.0935** | 0.9746 | 0.9316 | 2.6388 | 0.9461 | 0.9244 | **2.6983** | 0.9126 | 0.8985 | **3.2935** |
| | NR | Zhou *et al.* | 0.9020 | 0.9112 | 6.1170 | 0.9572 | 0.9611 | 5.6014 | 0.9533 | 0.9495 | 3.8730 | 0.9291 | 0.9141 | 4.5252 |
| | NR | VGCN | **0.9651** | 0.9639 | 3.6573 | **0.9894** | **0.9759** | **2.3590** | 0.9719 | 0.9659 | 3.1490 | 0.9401 | 0.9432 | 4.0257 |
| | NR | AHGCN | 0.9643 | 0.9623 | 3.6990 | 0.9869 | 0.9686 | 2.6162 | **0.9793** | **0.9753** | 2.7084 | 0.9419 | 0.9412 | 3.9657 |
| | NR | $\text{Ours}_{\text{Avg}}$ | 0.9645 | **0.9642** | 3.3398 | 0.9835 | 0.9475 | 3.0334 | 0.9736 | 0.9716 | 2.8318 | 0.9419 | 0.9535 | 3.9149 |
| | NR | $\text{Ours}_{\text{OR}}$ | 0.9629 | 0.9563 | 3.8672 | 0.9835 | 0.9535 | 3.0337 | 0.9719 | 0.9733 | 2.9205 | 0.9389 | 0.9525 | 3.9872 |
| | NR | $\text{Ours}_{\text{OR+SAL}}$ | 0.9630 | 0.9611 | 3.8789 | 0.9843 | 0.9578 | 3.1435 | 0.9735 | 0.9686 | 2.8048 | 0.9502 | **0.9566** | **3.6326** |

### VI.3.3 Cross-datasets evaluation

With the intent to demonstrate the generalization ability of the proposed model to new content built using different conditions, we carried out a cross-dataset evaluation. Hence, we trained our model on the OIQA datasets and tested its performance on CVIQ and vice versa. Table VI.3 reports the obtained performances in terms of accuracy (PLCC) and monotonicity (SRCC). As shown, our model outperformed the others while requiring the lowest computational complexity, demonstrating its ability to generalize to new content and distortions. For instance, by training on OIQA and testing on CVIQ our model achieved a PLCC (resp. SRCC) score of 0.9145 (resp. 0.9020), outperforming MC360IQA by (approx. 10% PLCC and 7% SRCC), VGCN by (approx. 3% PLCC and 4% SRCC), and Zhou *et al.* by (approx. 8% PLCC and 9% SRCC). By training on CVIQ and testing on OIQA, a similar behavior can be observed. Here, the proposed model outperformed even an NSS-based model (MFILGN), illustrating its superiority over standard approaches. Comparing among the training datasets, we observe that training on OIQA resulted in significantly better performances. Here, the diversity in terms of content as well as distortion types appears to significantly contribute to the generalization capability. Whereas, less diversity as in the case of CVIQ, relatively poor performances are attained. Besides, OIQA comprises GB and WGN, which are not available on CVIQ. This affects the ability of the models to adapt to new distortions. Nevertheless, the proposed model attained satisfactory performances compared to SOTA.

**Table VI.3:** Cross-database performances comparison of the proposed model with SOTA with respect to their complexity. The best performance is highlighted in **bold**.

| | | Trained / Tested on | | Complexity | |
|---|---|---|---|---|---|
| | | OIQA / CVIQ | CVIQ / OIQA | # Params | # FLOPs |
| MC360IQA | PLCC↑ | 0.8249 | 0.7443 | 22.4 M | 22.7 G |
| | SRCC↑ | 0.8442 | 0.6981 | | |
| VGCN | PLCC↑ | 0.8886 | 0.7911 | 26.7 M | 220 G |
| | SRCC↑ | 0.8629 | 0.7832 | | |
| MFILGN | PLCC↑ | - | 0.7885 | - | - |
| | SRCC↑ | - | 0.7589 | | |
| Zhou et al. | PLCC↑ | 0.8470 | 0.7350 | 29.3 M | 6.45 G |
| | SRCC↑ | 0.8250 | 0.7410 | | |
| Ours | PLCC↑ | **0.9145** | **0.8029** | **6.19 M** | **3.38 G** |
| | SRCC↑ | **0.9020** | **0.7926** | | |

In terms of complexity, the proposed model with 6.19 million of trainable parameters and 3.38 G of FLOPs has drastically fewer parameters and requires the least operations compared to the other models. In particular, VGCN (26.7M, 220G) requires much computational complexity. This is due to its architecture, involving twenty ResNet-

18 in parallel with a graph CNN and a subnetwork composed of the DB-CNN. In the case of MC360IQA (22.4M, 22.7G) and Zhou *et al.* (29.3M, 6.45G), a higher complexity is also illustrated. However, for the latter it is significantly lower in terms of number of FLOPs compared to the other multichannel models. The reason lies in its weights sharing among the CNN channels.

We further monitored the evolution of PLCC and RMSE performances of the proposed model during training on the whole OIQA and CVIQ databases to analyze its behavior. From Fig. VI.6, we first observe that training on OIQA converges faster than CVIQ. Training on CVIQ required 407 epochs, whereas training on OIQA lasted 336 epochs. Such a behavior is due to the size of the training sets. With CVIQ, the training set consists of 528 360-degree images, whereas OIQA has 320. As a result, the latter is about 40% smaller than the former. Moreover, PLCC and RMSE seem to improve at the same time in both databases, until reaching a tie by the 150$^{th}$ epoch for PLCC and 170$^{th}$ for RMSE. Overall, the results obtained on both databases are fairly competitive, with OIQA marginally outperforming. This actively exhibits the proposed model's ability to represent the training data as well as generalize to new ones.

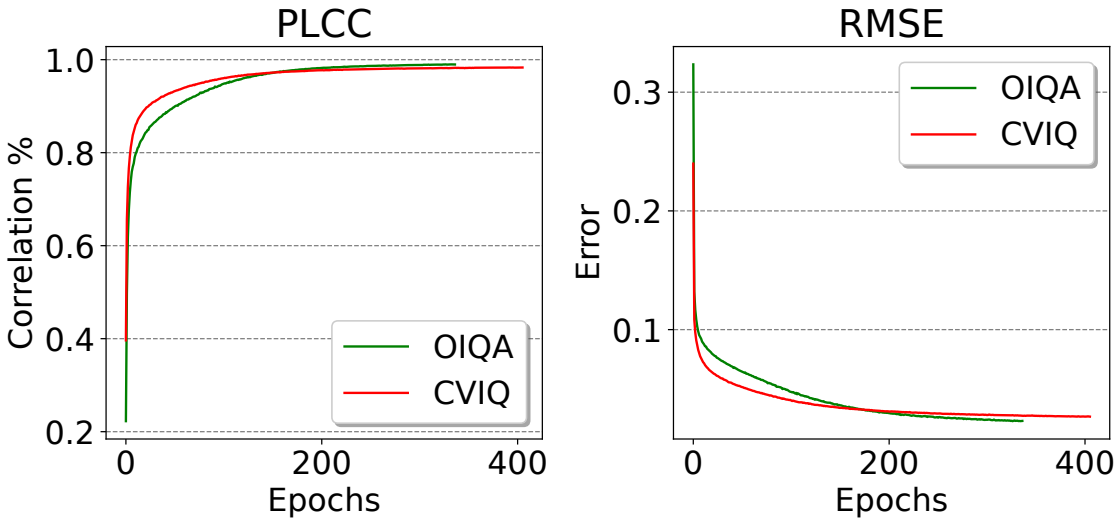



**Figure VI.6:** Evolution of PLCC and RMSE during training the proposed model on OIQA and CVIQ.

### VI.3.4 Ablation study

We conduct several ablation studies to analyze the impact of various components composing our model, including (i) the use of input normalization prior to training the model, (ii) the use of the SPA module, short and long skip-connections, and (iii) loss functions by comparing the use of the Huber loss to MSE and MAE.

### VI.3.4.1 Input normalization

By using local contrast normalization (LCN) on sampled patches prior to training the model, an improvement up to 4% is achieved, as illustrated in Fig. VI.7. The performances in terms PLCC (resp. SRCC) are increased by approx. 1.38% (resp. 1.16%) on OIQA, and by 4.31% (resp. 0.4%) on MVAQD. On CVIQ, the accuracy dropped slightly by 0.09% whereas the monotocicity improved by 1.5%. In terms of prediction errors, the adoption of LCN improves the prediction on both OIQA and MVAQD. Based on the overall performance of the LCN, one may conclude that normalizing input patches could be advantageous for using RGB content for IQA tasks. This lends credence to the idea of applying IQA adaptive representations on the input data prior to training.

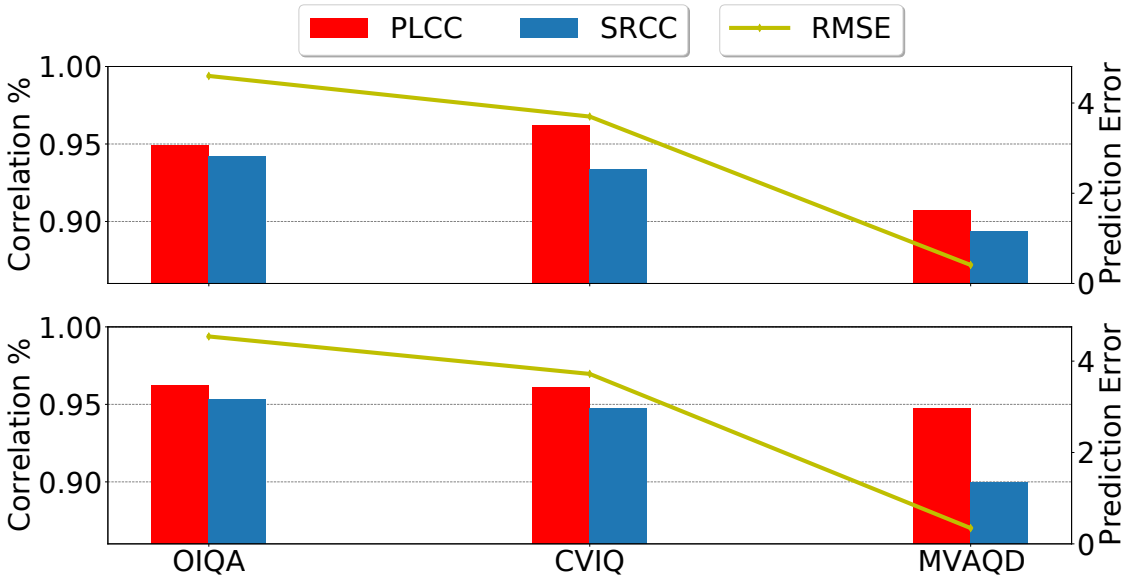



**Figure VI.7:** LCN vs. RGB patches. (top) RGB and (bottom) LCN.

### VI.3.4.2 Sampling methods

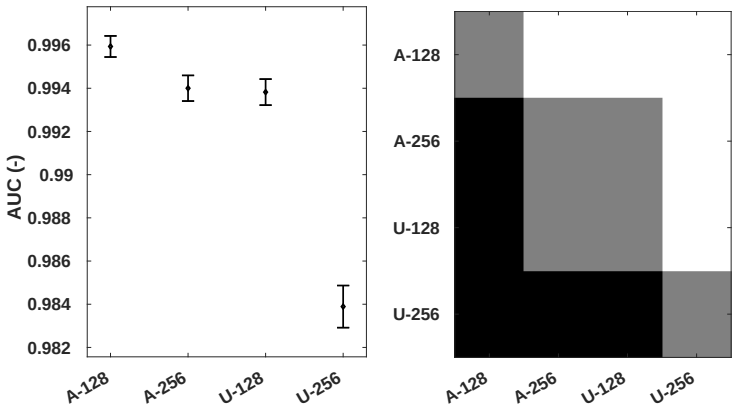
With the intent to evaluate the efficacy of using an adaptive sampling strategy and demonstrate the advantage of considering 360-IQA specific characteristics, we compare its performances with a uniform sampling strategy. The latter samples patches uniformly from the sphere without considering their importance. Here, all patches are sampled with the same size $\alpha_0$ (described in SEC. VI.2.1) along the latitude and longitude. In addition, we further analyze the influence of the patch size on the performance of the model. We set the resolution corresponding to $\alpha_0$ to 128 and 256 pixels. The correlations performances are provided in Table VI.4 and the statistical significance analysis conducted on OIQA is illustrated with Fig.VI.8.

From Table VI.4, the first observation that emerges is that the overall performances of the uniform sampling lags behind that of adaptive sampling. Incorporating 360-IQA specific properties to sample patches on the sphere appears to improve the

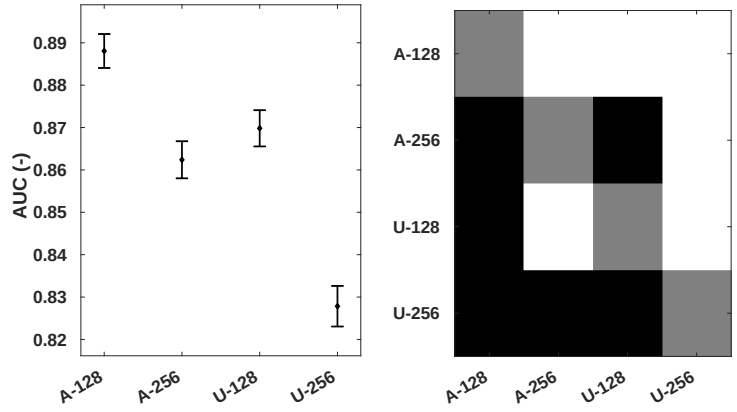
**Table VI.4:** Ablating the sampling methods with regard to the patches' size, *i.e.* $\alpha_0$.

| | Sampling | Uniform | | Adaptive | |
|---|---|---|---|---|---|
| | $\alpha_0$ size | 128 px | 256 px | 128 px | 256 px |
| OIQA | PLCC↑ | 0.9600 | 0.9572 | **0.9612** | 0.9607 |
| | SRCC↑ | 0.9524 | 0.9491 | **0.9563** | 0.9553 |
| | RMSE↓ | **3.8535** | 4.2070 | 3.9344 | 3.9890 |
| CVIQ | PLCC↑ | 0.9592 | 0.9587 | **0.9618** | 0.9603 |
| | SRCC↑ | 0.9390 | 0.9401 | 0.9553 | **0.9570** |
| | RMSE↓ | 3.7900 | 3.8417 | 3.9227 | **3.7665** |
| MVAQD | PLCC↑ | 0.9450 | 0.8940 | **0.9512** | 0.9409 |
| | SRCC↑ | 0.9022 | 0.8371 | **0.9402** | 0.9253 |
| | RMSE↓ | **0.2809** | 0.9399 | 0.3296 | 0.3899 |

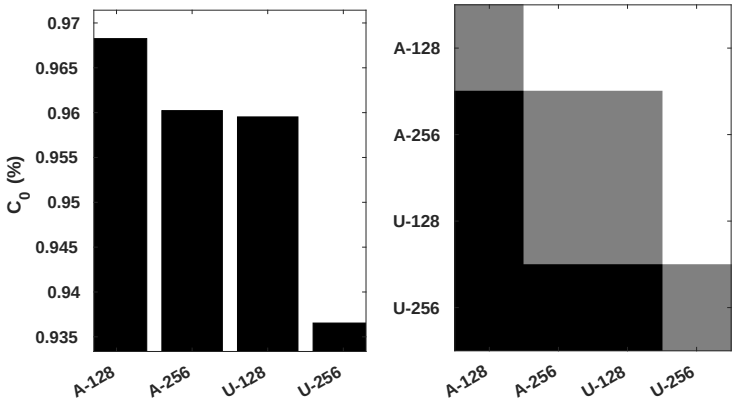
performances of the model. As it can be seen, a performance gain is achieved on all datasets by using the adaptive sampling, with important margins up to 10% in terms of SRCC when $\alpha_0 = 256$ on MVAQD. Regarding the influence of the size of $\alpha_0$, one can observe that setting $\alpha_0 = 128$ outperformed $\alpha_0 = 256$ despite the used sampling method, except on CVIQ. On the latter, using $\alpha_0 = 256$ resulted in a slight improvement in terms of correlation monotonicity (SRCC) and prediction errors (RMSE). By setting $\alpha_0 = 128$, more patches can be obtained, creating a rich training set with more examples compared to setting $\alpha_0 = 256$. The amount of the training data certainly matters because it affects the accuracy of the model. Looking closely into the case of CVIQ with $\alpha_0 = 256$, one might assume that fewer patches introduce less redundancy as all patches from the same image are labeled with the same MOS. This is also depicted by the small margin in performances on OIQA that are within 1% of improvement. However, by analyzing the statistical significance on OIQA in terms of the ability to distinguish and classify the stimulus with better quality among image pairs, setting $\alpha_0 = 128$ is found to be significantly superior as illustrated with the plots in Fig VI.8. From these plots, several observations can be drawn. First, the scores obtained by the both Better vs. Worse (Fig VI.8 a and c) analysis is significantly higher compared to the different vs. similar (Fig VI.8 b) one. A difference up to approx. 10% can be seen with the error bars, which represents 95% confidence intervals. This indicates that distinguishing between different/similar pairs remains a difficult challenge when compared to better/worse pairs, which correspond to the visual quality judgment by the HVS. Second, the use of adaptive sampling with $\alpha_0 = 128$ significantly outperforms the other three settings, proving the superiority of (i) incorporating 360-IQA characteristics and (ii) generating larger training sets. Based on the overall assessment and statistical significance analysis, we chose $\alpha_0 = 128$ with the adaptive sampling for training the proposed model.




(a) Better vs. Worse (AUC)




(b) Different vs. Similar (AUC)




(c) Better vs. Worse ($C_0$)

**Figure VI.8:** Statistical significance analysis on OIQA among the sampling methods with respect to the size of $\alpha_0$ using the Krasula *et al.* method. "A-128/A-256" stand for the adaptive sampling with $\alpha_0 = 128/256$ respectively. "U-128/U-256" stand for the uniform sampling. For the significance plots, a white/black square: row model is statistically better/worse than the column one; gray square: statistically indistinguishable.

### VI.3.4.3 SPA module and Skip-connections

Each convolutional block in the proposed model is augmented with a spatial attention (SPA) module, as outlined in Sec. VI.2.3. The latter is required to assist the model in focusing on significant features. This is mostly accomplished by the learnable weights via backpropagation, which optimizes the attention masks $M_{SPA}(\mathbf{F}) \in \mathbb{R}^{H \times W}$ at each step. The results of the ablation study are provided in Table VI.5 to evaluate the efficacy of the SPA module and the utilization of short- and long-skip connections. In addition, a computational complexity analysis is performed, and the results are provided in the same table. The computational time for prediction is obtained as the average of prediction times across 100 images. The comparison is conducted with the SAL-360IQA [202], a previous version of the proposed model, where a pooling step using min-, max-, and average-pooling is employed at each Conv Block instead of the SPA module.

**Table VI.5:** Ablating the use of the SPA module and skip-connections. The Best performance is highlighted in **bold** and the second-best is underlined.

| | | OIQA | | | CVIQ | | | MVAQD | | | Computational complexity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Version | Skip-conx | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | PLCC↑ | SRCC↑ | RMSE↓ | #Params↓ | #FLOPs↓ | Time↓ |
| SAL-360IQA | ✗ | 0.9611 | 0.9547 | 3.8950 | 0.9589 | 0.9531 | 3.8308 | 0.9507 | 0.9428 | 0.3311 | 9.89 M | 5.15 G | 0.52s |
| Ours | ✗ | 0.9618 | 0.9537 | 3.9118 | **0.9627** | 0.9555 | **3.6519** | 0.9428 | 0.9296 | 0.3839 | **5.74 M** | **3.34 G** | 0.38s |
| | Short | 0.9664 | 0.9583 | 3.6214 | 0.9615 | 0.9534 | 3.7132 | **0.9546** | **0.9468** | 0.3430 | **5.74 M** | **3.34 G** | **0.35s** |
| | Long | 0.9647 | 0.9567 | 3.6818 | 0.9565 | 0.9587 | 3.9390 | 0.9474 | 0.8863 | 0.3437 | 6.19 M | 3.38 G | 0.36s |
| | Short + Long | **0.9668** | **0.9585** | **3.5707** | 0.9607 | **0.9634** | 3.7475 | 0.9512 | 0.9402 | **0.3296** | 6.19 M | 3.38 G | 0.36s |

The performances in Table VI.5 suggest that the proposed model is robust. An accuracy up to 0.97, 0.96, and 0.95 expressed by the PLCC is obtained on OIQA, CVIQ, and MVAQD, respectively. The remaining evaluation metrics exhibit similar behavior, supporting its robustness. This demonstrates the effectiveness of the adopted training strategy, including the adaptive patch sampling, normalization, and model design. When comparing among the proposed model and its version without the SPA module and skip-connections, the latter reaches the best performance overall, while exhibiting less complexity. The complexity is decreased by approx. 50%, 42%, and 39% in terms of #Params, #FLOPs, and prediction time, respectively, as illustrated by the complexity analysis.

The ablation study on the use of skip-connections revealed that it can improve significantly the performances as depicted by Fig. VI.9. In particular, the short-connection within the SPA module. Aligning each feature map (before applying attention) with its refined version (after the attention), makes the aligned weights greater compared to non-aligned ones. This helps to highlight even more the important features within the spatial dimension of each feature map. In the case of long-connection, it appears that a significant improvement is attained. Here, reusing the earliest features at the

last stage through hierarchical element-wise summation brought additional information that could be lost between the first Conv Block and the last one. When compared to short-connection, the latter is adding more values to the overall performance of the model, as it can be seen with Fig. VI.9. Using only short- vs long-connection, the model is able to classify image pairs into better/worse significantly better when using only long-connection. However, combining both skip-connections yielded no statistically significant difference when compared to using only short-connections. Besides, when combining short and long skip-connections with the SPA module, the model achieves the best overall performances, as highlighted in Table VI.5. As a result, the latter is adopted as our model's final architecture.

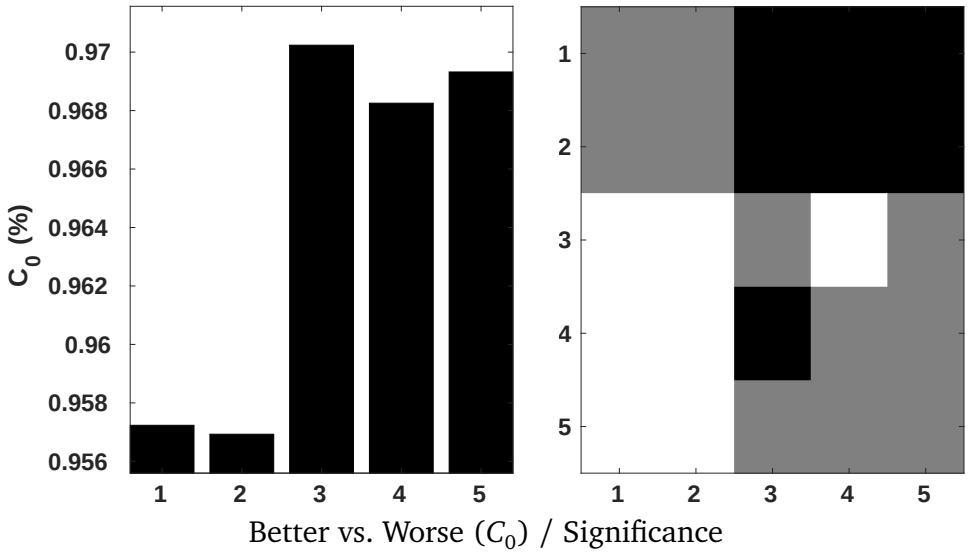
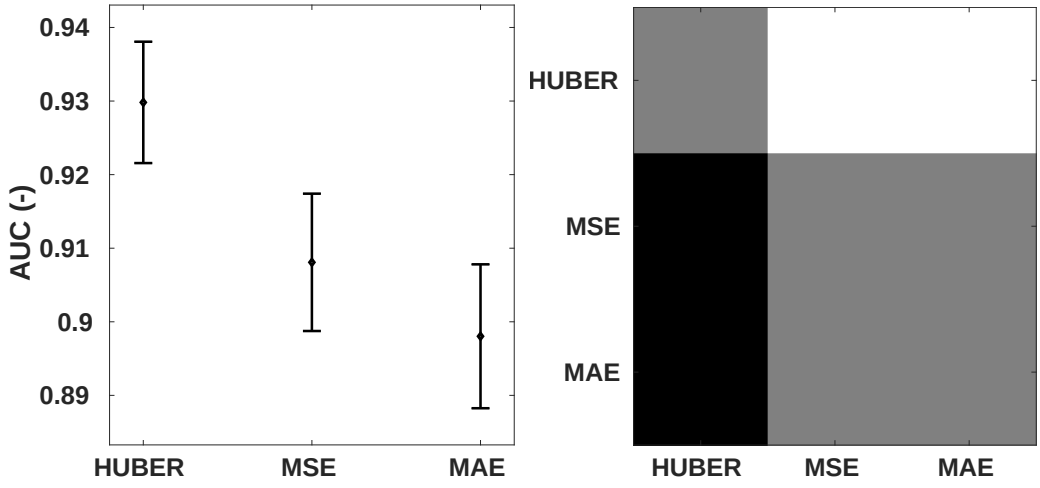


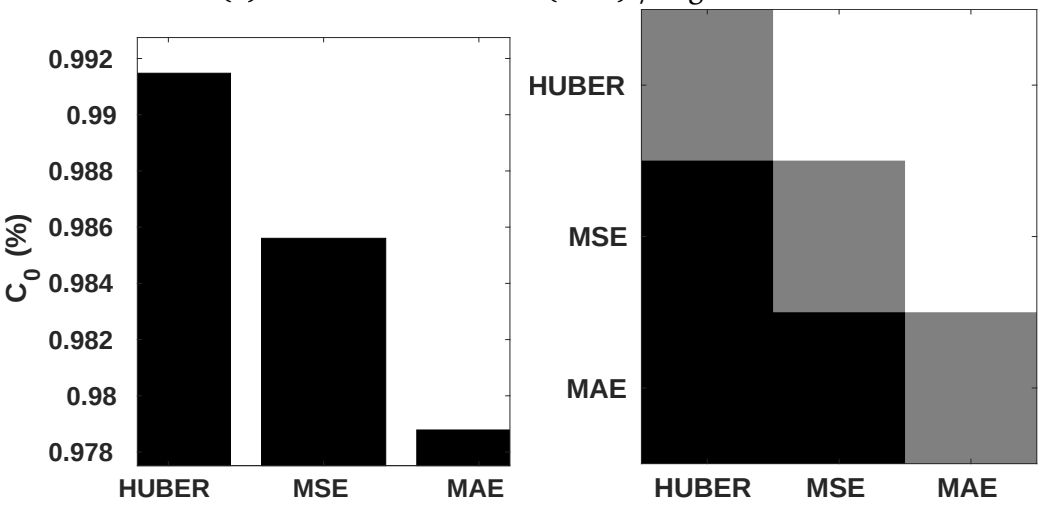

**Figure VI.9:** Statistical significance analysis on OIQA for the SPA module/skip-connections using the Krasula *et al.* method. 1: No SPA/No skip-connections, 2: SPA/No skip-connections, 3: SPA/Short-connection, 4: SPA/Long-connection, 5: SPA/Short+Long connections. For the significance plot, a white/black square: row model is statistically better/worse than the column one; gray square: statistically indistinguishable.

#### VI.3.4.4 Loss functions

The Huber loss, which combines the properties of MSE and MAE, was applied to train our model. To ascertain it effectiveness, we conduct a comparison analysis with the use of MSE and MAE by the Krasula *et al.* methods on OIQA. The results are illustrated with Fig. VI.10 in terms of (a) capacity to classify into different/similar pairs using the AUC and (b) the percentage of correct classifications into better/worse pairs denoted as $C_0$. According to the provided plots, the Huber loss appears to be significantly superior to MSE and MAE with both AUC different/similar and $C_0$ better/-worse analysis. This actively demonstrate its interest regarding the contribution to




(a) Different vs. Similar (AUC) / Significance




(b) Better vs. Worse ($C_0$) / Significance

**Figure VI.10:** Statistical significance analysis on OIQA among the Huber, MSE, and MAE loss functions using the Krasula *et al.* method. For the significance plots, a white/black square: row model is statistically better/worse than the column one; gray square: statistically indistinguishable.

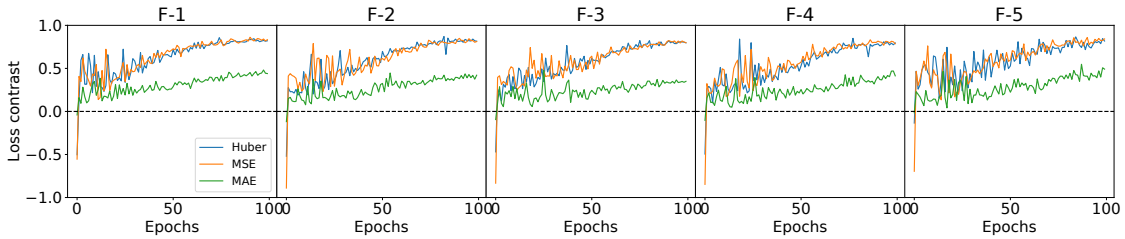



**Figure VI.11:** Contrast (val_loss - loss)/(val_loss + loss) between training and validation losses for the five folds, F-1 to F-5, (0 → equal loss between training and validation.

the accuracy of the model. It is proven that better/wore classification is less challenging than different/similar classification. Yet, an accuracy of approx. 0.93 is achieved with the Huber loss compared to 0.91/0.90 with MSE/MAE. Besides, MSE performed significantly better than MAE for the AUC different/similar while being statistically indistinguishable for $C_0$ better/worse.

In addition to the performance comparison, we provide in Fig. VI.11 the contrast between the training and validation losses to analyze the evolution of training losses. The contrast is computed as $(val_loss - loss)/(val_loss + loss)$. The latter provides insights on the training behavior, including 1) under-fitting happening when it can neither model the training data nor generalize to new one, 2) over-fitting when a model learns the detail and noise in the training data to the extent that it negatively impacts its performance on new data, and 3) good-fit when reaching a stable learning. A contrast equal to 0 depicts an equal loss between training and validation, whereas a contrast equal or close to 1 suggests an important gap between both losses, with $val_loss$ being higher and the opposite if equal or close to −1. From the curves, there appear to be no sign of under-fitting regardless of the used loss function, depicting the efficacy of the adopted training strategy. The MAE appears to have less gap between the training/validation losses, compared to Huber or MSE. The behavior of the latter can be explained by the fact that the training loss is improving faster than the validation one. However, the achieved losses by the Huber loss are smaller than MSE and MAE, with an important margin compared to MAE. For instance, the Huber loss attained a training/validation losses of 0.0003/0.003 compared to MSE : 0.0005/0.006 and MAE : 0.024/0.071 for the same fold. This supports the observations drawn above from the statistical significance analysis.

#### VI.3.4.5 Local qualities aggregation

A patch-based CNN model is basically trained on individual patches extracted from the input images. This means that the model is trained only on these patches, without having access to the whole 360-degree images. Therefore, $N$ local qualities corresponding to $N$ patches are predicted. The mapping of these individual scores to a single quality score could be challenging. This operation must be performed by adaptive ag-

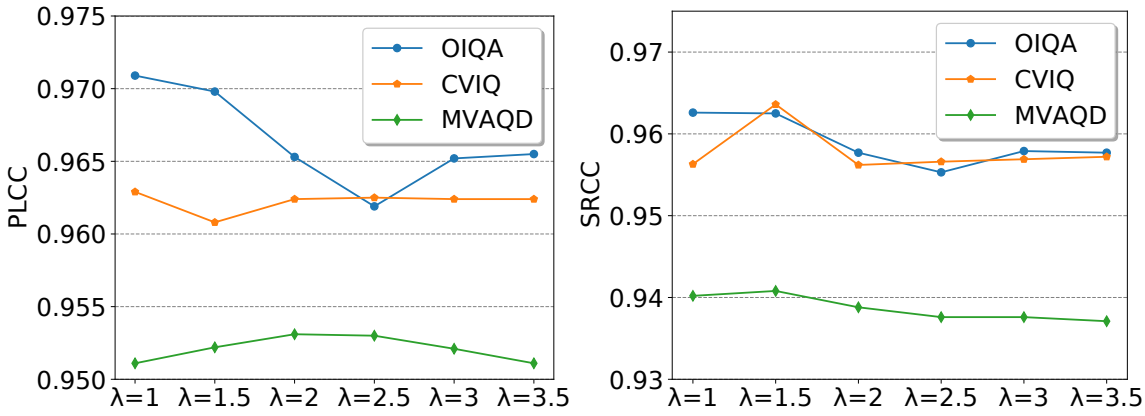



**Figure VI.12:** Evaluating the performances of the OR-based local qualities' aggregation by varying the value of $\lambda$.

gregation to improve the correlation with the human judgement of perceived quality. As described in Sec. VI.2.5, the proposed model uses an aggregation strategy based on OR. To investigate the effect of varying the $\lambda$ parameter on the performances of the correlation, we conducted a comparison analysis in terms of PLCC and SRCC on OIQA, CVIQ, and MVAQD. The performances are shown in Fig. VI.12. As it can be seen, the PLCC is strongly affected by the value of $\lambda$, in particular on OIQA. The PLCC decreases with bigger $\lambda$ values until $\lambda = 2.5$ and then slightly increases again. The same can be observed with SRCC. As for CVIQ, the PLCC appears to be stable and less affected by $\lambda$. This supports the observation drawn in Sec. VI.3.2 where the predicted local qualities seemed to be within an agreement. However, the SRCC increases with $\lambda = 1.5$ and then decreases before it stabilizes. Here, the $\lambda$ parameter is affecting the monotonicity over the accuracy. For MVAQD, both PLCC and SRCC slightly improve at different $\lambda$ values before decreasing. These behaviors suggest that the variation among local qualities is variable with respect to the used datasets.

## VI.4 Conclusion

In this chapter, an attention aware patch-based CNN model for blind 360-IQA was presented. Spatial attention is used to help the model focus on spatially meaningful features. Skip-connections within the spatial attention module were also integrated to align the preserved features via spatial attention. The exploration behavior, as well as a latitude-based selection, were used to sample patches appropriately on the sphere for training the model. The latter has shown a significant improvement over standard sampling, not accounting for 360-IQA peculiarities. Patch quality aggregation was accomplished adaptively using saliency and outlier reduction, which resulted in better correlation performances with the ground truth. The proposed model demonstrated a good performance across different databases and distortions. In comparison to SOTA

models in general, and multichannel ones in particular, the proposed model attained competitive or much better performances while maintaining the lowest complexity. This demonstrates the value of the adopted (i) appropriate patch sampling, (ii) data representation, (iii) model architecture, and (iv) aggregation strategy. Furthermore, the proposed model's generalization capacity demonstrated its superiority in adapting to new content and distortions through cross-database evaluations.

# General Conclusion

With the increase of immersive media demands and users' expectations, providing a good QoE became the focus of content providers, such as Meta, Netflix, Youtube, etc. In addition, as the technologies allowing users to use immersive applications became accessible to the general public, the need for appropriate tools contributing to the evaluation and improvement of QoE becomes urgent. Proper examination of immersive media in terms of their immersive capacity and quality requires the accessibility of measurement instruments and test protocols for the assessment of different aspects. IQA among other tools can be used to attain this goals.

The objectives of the presented thesis are to first investigate and understand factors influencing 360-IQA from the subjective and objective perspectives. The main contributions pertain to both aspects of IQA, with an emphasized focus on the objective one. Due to the Covid-19 pandemic, conducting more subjective experiments was quite challenging.

First, part of the research work conducted in this thesis was dedicated to explore the influence of HMDs on subjective quality ratings. The hypothesis that the HMD has a significant impact on observers' quality ratings was confirmed overall and per-distortion levels. The obtained results allowed to determine to what extent the HMD may influence the final ratings as well as the generation of cyber-sickness. In addition, the observations drawn from the conducted subjective experiments allowed to question the reliability of existing studies, in particular the construction of 360-degree databases.

Second, we deeply investigated the use of CNNs for 360-IQA at various scales and from different perspectives. This was achieved by an extensive benchmark of widely adopted CNNs by including different architectures, image representations, and training strategies. Various configurations were considered, comprising the use of projected and radial content, multichannel paradigm and patch-wise training, and retraining on well-known 2D IQA databases. The obtained results showed that recommendations coming from other image processing communities may not hold for IQA in general, and 360-IQA in particular. The conclusion of this study served for the design of accurate and predictive 360-IQA models based on CNN either with transfer-learning or designed from scratch.

Next, various contributions intended to improve the 360-IQA processing chain were described. The main focus of these contributions were the pre- and post-processing. Pre-processing, in addition to data representation and preparation are crucial for

achieving predictability and robustness. The obtained results supported these assumptions, and allowed to draw solid conclusions on the usefulness of adaptive (i) data-augmentation, (ii) patches sampling, and (ii) data representation. As for the post-processing, adaptive aggregation was found useful compared to simple methods, and that the correlation with subjective scores can be significantly improved.

Finally, based on observations from the various studies conducted in this thesis, we designed two NR 360-IQA models based on CNNs. The first one named SP36IQA is a viewport-based multichannel CNN model that incorporates visual scanpaths and JND probability maps, in addition to visual features. The important feature of this model is the adaptive weights estimation of each considered viewport. For this, features based HVS properties are used. The second model named SAL-360IQA, is a patch-based CNN model featuring spatial attention and skip connections. The spatial attention is used to help the model focus on spatially meaningful features. The skip-connections are used at two stages. The first is within the spatial attention module to align the preserved features via spatial attention. The second stage consists of a long skip-connection inter convolutional blocks so as to reuse features from earliest layers at later ones. In comparison to SOTA models in general, and multichannel ones in particular, the proposed model reached competitive or much better performances while maintaining lower complexity.

# List of publication

- **International journal papers**:
  - **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh. "Convolutional neural networks for omnidirectional image quality assessment: A benchmark". IEEE Transactions on Circuits and Systems for Video Technology (2022). doi:10.1109/tcsvt.2022.3181235. (**Impact Factor: 5.859**). [203]
  - **Abderrezzaq Sendjasni** and Mohamed-Chaker Larabi. "Attention-aware patch-based CNN for blind 360-degree image quality assessment". ***Under review***, 2022
- **International conference papers**:
  - **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh. "On the Improvement of 2D Quality Assessment Metrics for Omnidirectional Images". In Electronic Imaging, Image Quality and System Performance XVII, pages 287–1, 2020. [109]
  - **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh. "Convolutional neural networks for omnidirectional image quality assessment: Pre-trained or re-trained?" In IEEE International Conference on Image Processing (ICIP), pages 3413–3417, 2021. [152]
  - **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh.

"On the Influence of Head-Mounted Displays on Quality Rating of Omnidirectional images". In Electronic Imaging, Image Quality and System Performance XVIII, pages 296-1, 2021. [204]

- **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh. "Perceptually-weighted CNN for 360-degree image quality assessment using visual scan-path and JND". In IEEE International Conference on Image Processing (ICIP), pages 1439–1443, 2021. [151]
- **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh. "Visual Scan-Path based Data-Augmentation for CNN-based 360-degree Image Quality Assessment". In London Imaging Meeting, volume 2021, pages 21–26. Society for Imaging Science and Technology, 2021. [205]
- **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi, and Faouzi Alaya Cheikh. "Patch-based CNN Model for 360 Image Quality Assessment with Adaptive Pooling Strategies". In Electronic Imaging, Image Quality and System Performance XIX. 2022. (***Best paper award***)
- **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi. "SAL-360IQA: A Saliency Weighted Patch-based CNN Model for 360-degree Images Quality Assessment". In IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pages 1-6, 2022. [202]
- **Abderrezzaq Sendjasni**, David Traparic, Mohamed-Chaker Larabi. "Investigating normalization methods for CNN-based image quality assessment". In IEEE International Conference on Image Processing (ICIP), pages 4113-4117, 2022. [206]
- **Abderrezzaq Sendjasni**, Mohamed-Chaker Larabi. "Transfer learning from vision transformers or ConvNets for 360-degree images quality assessment?". In IEEE International Conference on Image Processing (ICIP), pages 4133-4137, 2022. [207]

## Limitations and future work

This thesis has several contributions to 360-IQA. We proposed CNN based models to blindly predict the quality of 360-degree images. In addition, several recommendations on the use of CNNs for this purpose were stated.

Even though this thesis covers the solution to several issues related to SQA and OQA of 360-degree images, the link between subjective understanding and objective evaluation is a promising direction that is still uncovered. In our opinion, the promising direction to trace could be the incorporation of HMD-induced distortion into the process of objectively evaluating 360-degree images. As the observers rates the quality of 360-degree images using HMD, the rendered content is only seen by the observers. The rendering effects are not considered in OQA frameworks, which could be seen as inconsistent with subjective evaluation.

Regarding the usage of CNNs for 360-IQA, we concluded that patch-based ones are

more appropriate if a suitable training strategy is used. However, labeling individual patches with the same MOS introduces redundancy. Since not all patches provide the same details or are of equal quality, this might lead to unstable training, on the one hand. On the other hand, the high resolution of 360-degree images makes it more inappropriate to assign the quality of the whole image to small portions of it. Accordingly, solving the labeling issue would benefits to training CNNs for IQA in general, and 360-IQA in particular.

Regardless of the degree of immersion 360-degree content can deliver to viewers, stereoscopic 360-degree content can provide even more immersion. It allows users to have both stereoscopic and 360-degree experiences. This may help to improve the users' QoE. In the meanwhile, quality assessment may prove more challenging since it must deal with stereoscopic and 360-degree content at the same time. Only a few efforts have been made to evaluate the quality of stereoscopic 360-IQA. To the best of our knowledge, only the works reported in [208–213] focused on stereoscopic 360-IQA. Therefore, the latter is a promising direction to follow. All the observations and conclusions drawn from this thesis may pave the way toward this goal.

# Bibliography


[1] *Optimizing 360 photos at scale*, https://engineering.fb.com/2017/08/31/ml-applications/optimizing-360-photos-at-scale/, [online], Accessed: 2022-09-14.

[2] *VR video formats explained*, https://360labs.net/blog/vr-video-formats-explained, [online], Accessed: 2022-09-14.

[3] Z. Wang, A. Bovik and L. Lu, 'Why is image quality assessment so difficult?' In *IEEE ICASSP*, vol. 4, Orlando, FL, USA, 2002, pp. IV–3313.

[4] B. Keelan, *Handbook of image quality: characterization and prediction*. CRC Press, 2002.

[5] ITU, 'Influencing factors on quality of experience (QoE) for virtual reality services,' ITU-T VCEG (Q13/12), Tech. Rep. ITU-T G.1035, 202.

[6] F. Biocca and B. Delaney, 'Immersive virtual reality technology,' *Communication in the age of virtual reality*, vol. 15, no. 32, pp. 10–5555, 1995.

[7] V. Gisbergen and M. Sebastiaan, 'Contextual connected media: How rearranging a media puzzle, brings virtual reality into being,' 2016.

[8] A. Perkis, C. Timmerer, S. Baraković, J. Husić and *et al.*, 'Qualinet white paper on definitions of immersive media experience (imex),' *European Network on Quality of Experience in Multimedia Systems and Services, 14th QUALINET meeting (online), May 25, 2020. Online: https://arxiv.org/abs/2007.07032*,

[9] B. Witmer and M. Singer, 'Measuring presence in virtual environments: A presence questionnaire,' *Presence*, vol. 7, no. 3, pp. 225–240, 1998.

[10] M. Schuemie and P. S. *et al.*, 'Research on presence in virtual reality: A survey,' *CyberPsychology & Behavior*, vol. 4, no. 2, pp. 183–201, 2001.

[11] S. Huiberts, 'Captivating sound the role of audio for immersion in computer games,' Ph.D. dissertation, University of Portsmouth, 2010.

[12] R. Kaplan-Rakowski and K. Meseberg, 'Immersive media and their future,' in *Educational media and technology yearbook*, Springer, 2019, pp. 143–153.

[13] G. Burdea and P. Coiffet, *Virtual reality technology*. John Wiley & Sons, 2003.

[14] J. Zheng, K. Chan and I. Gibson, 'Virtual reality,' *Potentials*, vol. 17, no. 2, pp. 20–23, 1998.

[15] J. Carmigniani and B. Furht, 'Augmented reality: An overview,' *Handbook of augmented reality*, pp. 3–46, 2011.

[16] P. Milgram and F. Kishino, 'A taxonomy of mixed reality visual displays,' *IEICE TRANSACTIONS on Information and Systems*, vol. 77, no. 12, pp. 1321–1329, 1994.

[17] M. Speicher, B. Hall and M. Nebeling, 'What is mixed reality?' In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[18] G. Koulieris, B. Bui, M. Banks and G. Drettakis, 'Accommodation and comfort in head-mounted displays,' *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[19] R. Kennedy, E. Lane, K. Berbaum and M. Lilienthal, 'Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness,' *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[20] H. Kim, J. Park, Y. Choi and M. Choe, 'Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment,' *Applied ergonomics*, vol. 69, pp. 66–73, 2018.

[21] A. Singla, S. Fremerey, W. Robitza and A. Raake, 'Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays,' in *9th QoMEX*, Erfurt, Germany, 2017, pp. 1–6.

[22] A. Kopyt and J. Narkiewicz, 'Technical factors influencing simulator sickness,' *Zeszyty Naukowe Politechniki Rzeszowskiej. Mechanika*, vol. 85, no. 2888, pp. 455–467, 2013.

[23] H. Allum and F. Honegger, 'Interactions between vestibular and proprioceptive inputs triggering and modulating human balance-correcting responses differ across muscles,' *Experimental brain research*, vol. 121, no. 4, pp. 478–494, 1998.

[24] E. A. Yan Ye and J. Boyce, 'Algorithm descriptions of projection format conversion and video quality metrics in 360lib,' *JVET of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11*, 2017.

[25] C. Brown, *Bringing pixels front and center in vr video*, https://blog.google/products/google-ar-vr/bringing-pixels-front-and-center-vr-video/, Online; accessed October 2022.

[26] C. Fu, L. Wan, T. Wong and C. Leung, 'The rhombic dodecahedron map: An efficient scheme for encoding panoramic video,' *IEEE TMM*, vol. 11, no. 4, pp. 634–644, 2009.

[27] H. Lin, C. Huang, C. Li, Y. Lee and J. L. *et al.*, 'AHG8: An improvement on the compact ohp layout,' *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO / IEC JTC1 / SC29 / WG11, JVET-E0056*, 2017.

[28] S. A. *et al.*, 'AHG8: Efficient frame packing for icosahedral projection, document jvet-e0029, joint video exploration team of itu-t sg16 wp3 and iso,' IEC JTC1 / SC29 / WG11, Geneva, Switzerland, Tech. Rep., 2017.

[29] M. Coban, D. Van and M. Karczewicz, 'AHG8: Adjusted cubemap projection for 360-degree video,' *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO / IEC JTC1 / SC29 / WG11, JVET-F0025*, 2017.

[30] J. Li, Z. Wen, S. Li, Y. Zhao, B. Guo and J. Wen, 'Novel tile segmentation scheme for omnidirectional video,' in *IEEE ICIP*, Phoenix, AZ, USA, 2016, pp. 370–374.

[31] M. Yu, H. Lakshman and B. Girod, 'Content adaptive representations of omnidirectional videos for cinematic virtual reality,' in *3rd International Workshop on Immersive Media Experiences*, 2015, pp. 1–6.

[32] X. Zheng, G. Jiang, M. Yu and H. Jiang, 'Segmented spherical projection-based blind omnidirectional image quality assessment,' *IEEE Access*, vol. 8, pp. 31 647–31 659, 2020.

[33] A. Abbas and D. Newman, 'AHG8: Rotated sphere projection for 360 video,' *JVET-F0036, Hobart, AU*, vol. 31, 2017.

[34] Z. Chen, Y. Li and Y. Zhang, 'Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation,' *Signal Processing*, vol. 146, pp. 66–78, 2018.

[35] V. Zakharchenko, E. Alshina, K. Choi, C. Pujara and A. Dsouza, 'AHG8: Coding performance impact of omnidirectional projection rotation,' *Joint Video Exploration Team of ITU-T SG*, vol. 16, 2017.

[36] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou and X. Cao, 'Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression,' *IEEE TIP*, vol. 27, no. 12, pp. 6039–6050, 2018.

[37] S. board VR, *Base stations  VR arcades*, https://springboardvr.com/blog/base-stations-vr-arcades, Online; accessed October 2021.

[38] S. Daly, 'Visible differences predictor: An algorithm for the assessment of image fidelity,' in *Human Vision, Visual Processing, and Digital Display III*, SPIE, vol. 1666, 1992, pp. 2–15.

[39] D. Silverstein and J. Farrell, 'The relationship between image fidelity and image quality,' in *IEEE ICIP*, vol. 1, 1996, 881–884 vol.1. DOI: 10.1109/ICIP.1996.559640.

[40] N. Burningham, Z. Pizlo and J. Allebach, 'Image quality metrics,' *Encyclopedia of imaging science and technology*, vol. 1, pp. 598–616, 2002.

[41] H. Sheikh and A. Bovik, 'Information theoretic approaches to image quality assessment,' in *Handbook of image and video processing*, Elsevier, 2005, pp. 975–989.

[42] Z. Wang and A. Bovik, 'Modern image quality assessment,' *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[43] M. Shahid, A. Rossholm, B. Lövström and H. Zepernick, 'No-reference image and video quality assessment: A classification and review of recent approaches,' *EURASIP Journal on image and Video Processing*, vol. 2014, no. 1, pp. 1–32, 2014.

[44] ITU-R, *Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service*. ITU, 2012, vol. 13.

[45] L. Krasula, K. Fliegel, P. L. Callet and M. Klíma, 'On the accuracy of objective image and video quality models: New methodology for performance evaluation,' in *QoMEX*, Lisbon, Portugal, 2016, pp. 1–6.

[46] T. Fawcett, 'An introduction to roc analysis,' *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[47] J. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press, 2014.

[48] J. Hanley and B. McNeil, 'A method of comparing the area under two ROC curves derived from the same cases,' *Radiology. v148*, pp. 839–843,

[49] J. Hanley and B. McNeil, 'The meaning and use of the area under a receiver operating characteristic (ROC) curve.,' *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[50] I. Recommendation, 'ITU-T P. 910,' *Subjective video quality assessment methods for multimedia applications*, 2008.

[51] A. Singla, W. Robitza and A. Raake, 'Comparison of subjective quality test methods for omnidirectional video quality evaluation,' in *IEEE 21$^{st}$ MMSP*, Kuala Lumpur, Malaysia, 2019, pp. 1–6.

[52] Z. Bo, Z. Junzhe, Y. Shu, Z. Yang, W. Jing and F. Zesong, 'Subjective and objective quality assessment of panoramic videos in virtual reality environments,' in *IEEE ICMEW*, Hong Kong, China, 2017, pp. 163–168.

[53] I.-T. R. BT, 'Methodology for the subjective assessment of video quality in multimedia applications (2007),' 1788.

[54] Y. Nehmé, J. Farrugia, F. Dupont, P. L. Callet and G. Lavoué, 'Comparison of subjective methods for quality assessment of 3d graphics in virtual reality,' *ACM TAP*, vol. 18, no. 1, pp. 1–23, 2020.

[55] F. Hofmeyer, S. Fremerey, T. Cohrs and A. Raake, 'Impacts of internal hmd playback processing on subjective quality perception,' *EI*, no. 12, pp. 219–1, 2019.

[56] P. Perez, N. Oyaga, J. Ruiz and A. Villegas, 'Towards systematic analysis of cybersickness in high motion omnidirectional video,' in *10th QoMEX*, Cagliari, Italy, 2018, pp. 1–3.

[57] W. Zhang, W. Zou, F. Yang, L. Lévêque and H. Liu, 'The effect of spatio-temporal inconsistency on the subjective quality evaluation of omnidirectional videos,' in *IEEE ICASSP*, Brighton, UK, 2019, pp. 4055–4059.

[58] W. Zou, L. Yang, F. Yang, Z. Ma and Q. Zhao, 'The impact of screen resolution of hmd on perceptual quality of immersive videos,' in *IEEE ICMEW*, London, United Kingdom, 2020, pp. 1–6.

[59] M. Xu, C. Li, Y. Liu, X. Deng and J. Lu, 'A subjective visual quality assessment method of panoramic videos,' in *2017 ICME*, Hong Kong, China, 2017, pp. 517–522.

[60] C. Curcio and K. Allen, 'Topography of ganglion cells in human retina,' *Journal of comparative Neurology*, vol. 300, no. 1, pp. 5–25, 1990.

[61] S. Rossi, C. Ozcinar, A. Smolic and L. Toni, 'Do users behave similarly in VR? investigation of the user influence on the system design,' *ACM TMCCA*, vol. 16, no. 2, pp. 1–26, 2020.

[62] I.-T. R. P.919, 'Subjective test methodologies for 360$^{c}irc$ video on head-mounted displays,' 2020.

[63] J. Gutierrez and et al., 'Subjective evaluation of visual quality and simulator sickness of short 360 videos: ITU-T Rec. P.919 (early access),' *IEEE TMM*, vol. 24, pp. 3087–3100, 2022. DOI: 10.1109/TMM.2021.3093717.

[64] H. Tran, N. Ngoc, C. Pham, Y. Jung and T. Thang, 'A subjective study on qoe of 360 video for vr communication,' in *2017 19$^{th}$ MMSP*, Luton, UK, 2017, pp. 1–6.

[65] F. Kozamernik, V. Steinmann, P. Sunna and E. Wyckens, 'Samviq—a new ebu methodology for video quality evaluations in multimedia,' *SMPTE Motion Imaging Journal*, vol. 114, no. 4, pp. 152–160, 2005. DOI: 10.5594/J11535.

[66] A. Singla, S. Fremerey, W. Robitza, P. Lebreton and A. Raake, 'Comparison of subjective quality evaluation for hevc encoded omnidirectional videos at different bit-rates for uhd and fhd resolution,' in *Proceedings of the on Thematic Workshops of ACM Multimedia*, 2017, pp. 511–519.

[67] M. Xu, C. Li, Z. Chen, Z. Wang and Z. Guan, 'Assessing visual quality of omnidirectional videos,' *IEEE TCSVT*, vol. 29, no. 12, pp. 3516–3530, 2019. DOI: 10.1109/TCSVT.2018.2886277.

[68] S. Fremerey, R. Huang, S. Göring and A. Raake, 'Are people pixel-peeping 360° videos?' *Electronic Imaging*, no. 10, pp. 220–1, 2019.

[69] T. Chu, H. Zepernick and M. Elwardy, 'Rating duration analysis for subjective quality assessment of 360° videos,' in *2020 ICVRV*, Recife, Brazil, 2020, pp. 42–46.

[70] M. Yu, H. Lakshman and B. Girod, 'A framework to evaluate omnidirectional video coding schemes,' in *IEEE International Symposium on Mixed and Augmented Reality*, Fukuoka, Japan, 2015, pp. 31–36. DOI: 10.1109/ISMAR.2015.12.

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

[72] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, 'Image quality assessment: From error visibility to structural similarity,' *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[73] S. Chen, Y. Zhang, Y. Li, Z. Chen and Z. Wang, 'Spherical structural similarity index for objective omnidirectional video quality assessment,' in *IEEE ICME*, San Diego, CA, USA, 2018, pp. 1–6.

[74] V. Zakharchenko, P. Kwang and H. Jeong, 'Quality metric for spherical panoramic video,' in *Optics and Photonics for Information Processing X*, vol. 9970, 2016, pp. 57–65.

[75] G. Luz, J. Ascenso, C. Brites and F. Pereira, 'Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment,' in *IEEE 19$^{th}$ MMSP*, Luton, UK, 2017, pp. 1–6.

[76] E. Upenik and T. Ebrahimi, 'Saliency driven perceptual quality metric for omnidirectional visual content,' in *IEEE ICIP*, Taipei, Taiwan, 2019, pp. 4335–4339.

[77] C. Ozcinar, J. Cabrera and A. Smolic, 'Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality,' *IEEE JETCAS*, vol. 9, no. 1, pp. 217–230, 2019. DOI: 10.1109/JETCAS.2019.2895096.

[78] S. Croci, C. Ozcinar, E. Zerman, J. Cabrera and A. Smolic, 'Voronoi-based objective quality metrics for omnidirectional video,' in *QoMEX*, Berlin, Germany, 2019, pp. 1–6.

[79] Z. Wang, E. Simoncelli and A. Bovik, 'Multiscale structural similarity for image quality assessment,' in *IEEE ACSSC*, vol. 2, Pacific Grove, USA, 2003, pp. 1398–1402.

[80] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy and M. Manohara, 'Toward a practical perceptual video quality metric,' *The Netflix Tech Blog*, vol. 6, no. 2, 2016.

[81] S. Croci, C. Ozcinar, E. Zerman, S. Knorr, J. Cabrera and A. Smolic, 'Visual attention-aware quality estimation framework for omnidirectional video using spherical voronoi diagram,' *Quality and User Experience*, vol. 5, no. 1, pp. 1–17, 2020.

[82] X. Sui, K. Ma, Y. Yao and Y. Fang, 'Perceptual quality assessment of omnidirectional images as moving camera videos,' *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 3022–3034, 2022. DOI: 10.1109/TVCG.2021.3050888.

[83] H. Sheikh, A. Bovik and G. Veciana, 'An information fidelity criterion for image quality assessment using natural scene statistics,' *IEEE TIP*, vol. 14, no. 12, pp. 2117–2128, 2005. DOI: 10.1109/TIP.2005.859389.

[84] L. Valero, B. BAlexander, B. Johannes and P. Simoncelli, 'Perceptually optimized image rendering,' *J. Opt. Soc. Am. A*, vol. 34, no. 9, pp. 1511–1525, Sep. 2017. DOI: 10.1364/JOSAA.34.001511. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-34-9-1511.

[85] K. Ding, K. Ma, S. Wang and P. Simoncelli, 'Image quality assessment: Unifying structure and texture similarity,' *IEEE TPAMI*, vol. 44, no. 5, pp. 2567–2581, 2020. DOI: 10.1109/TPAMI.2020.3045810.

[86] H. Kim, H. Lim and Y. Ro, 'Deep virtual reality image quality assessment with human perception guider for omnidirectional image,' *IEEE TCSVT*, vol. 30, no. 4, pp. 917–928, 2020. DOI: 10.1109/TCSVT.2019.2898732.

[87] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' in *IEEE CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.

[88] T. Truong, T. Tran and T. Thang, 'Non-reference quality assessment model using deep learning for omnidirectional images,' in *IEEE ICAST*, Morioka, Japan, 2019, pp. 1–5.

[89] Y. Xia, Y. Wang and Y. Peng, 'Blind panoramic image quality assessment via the asymmetric mechanism of human brain,' in *IEEE VCIP*, Sydney, NSW, Australia, 2019, pp. 1–4.

[90] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan and S. Ma, 'MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment,' in *IEEE JSTSP*, vol. 14, 2020, pp. 64–77. DOI: 10.1109/JSTSP.2019.2955024.

[91] J. Xu, W. Zhou and Z. Chen, 'Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,' *IEEE TCSVT*, vol. 31, no. 5, pp. 1724–1737, 2021. DOI: 10.1109/TCSVT.2020.3015186.

[92] Y. Zhou, Y. Sun, L. Li, K. Gu and Y. Fang, 'Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network,' *IEEE TCSVT*, pp. 1–1, 2021.

[93] W. Ding, P. An, X. Liu, C. Yang and X. Huang, 'No-reference panoramic image quality assessment based on adjacent pixels correlation,' in *2021 BMSB*, Chengdu, China, 2021, pp. 1–5.

[94] J. Fu, C. Hou, W. Zhou, J. Xu and Z. Chen, 'Adaptive hypergraph convolutional network for no-reference 360-degree image quality assessment,' *arXiv preprint arXiv:2105. 09143*, 2021.

[95] Q. Miaomiao and S. Feng, 'Blind 360-degree image quality assessment via saliency-guided convolution neural network,' *Optik*, vol. 240, p. 166 858, 2021, ISSN: 0030-4026.

[96] L. Yang, M. Xu, X. Deng and B. Feng, 'Spatial attention-based non-reference perceptual quality prediction network for omnidirectional images,' in *IEEE ICME*, Shenzhen, China, 2021, pp. 1–6.

[97] Y. Liu, H. Yu, B. Huang, G. Yue and B. Song, 'Blind omnidirectional image quality assessment based on structure and natural features,' *IEEE TIM*, vol. 70, pp. 1–11, 2021. DOI: 10.1109/TIM.2021.3102691.

[98] W. Zhou, J. Xu, Q. Jiang and Z. Chen, 'No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness,' *IEEE TCSVT*, vol. 32, no. 4, pp. 1778–1791, 2021.

[99] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, 'Rethinking the inception architecture for computer vision,' in *IEEE CVPR*, 2016, pp. 2818–2826.

[100] J. Xu, W. Zhou and Z. Chen, 'Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,' *IEEE TCSVT*, vol. 31, no. 5, pp. 1724–1737, 2021.

[101] W. Zhang, K. Ma, J. Yan, D. Deng and Z. Wang, 'Blind image quality assessment using a deep bilinear convolutional neural network,' *IEEE TCSVT*, vol. 30, no. 1, pp. 36–47, 2020.

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

[103] K. Z. Kao and Z. Chen, 'Video saliency prediction based on spatial-temporal two-stream network,' *IEEE TCSVT*, vol. 29, no. 12, pp. 3544–3557, 2019.

[104] C. Cortes and V. Vapnik, 'Support-vector networks,' *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[105] C. Tian, X. Chai and G. C. *et al.*, 'VSOIQE: A novel viewport-based stitched 360° omnidirectional image quality evaluator,' *IEEE TCSVT (Early Access)*, pp. 1–1, 2022. DOI: 10.1109/TCSVT.2022.3172135.

[106] J. Li, K. Yu, Y. Zhao, Y. Zhang and L. Xu, 'Cross-reference stitching quality assessment for 360 omnidirectional images,' in *27th ACM ICM*, 2019, pp. 2360–2368.

[107] K. Seshadrinathan, R. Soundararajan, A. Bovik and L. Cormack, 'Study of subjective and objective quality assessment of video,' *IEEE TIP*, vol. 19, no. 6, pp. 1427–1441, 2010. DOI: 10.1109/TIP.2010.2042111.

[108] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang and X. Yang, 'Perceptual Quality Assessment of Omnidirectional Images,' in *IEEE ISCAS*, Florence, Italy, 2018, pp. 1–5.

[109] A. Sendjasni, M. Larabi and F. Cheikh, 'On the improvement of 2d quality assessment metrics for omnidirectional images,' in *Electronic Imaging*, Burlingame, California USA, 2020, pp. 287–1.

[110] P. Lebreton, *Siti*, https://vqeg.github.io/software-tools/quality%20analysis/siti/, Online; accessed August 2022.

[111] D. Hasler and S. Süsstrunk, 'Measuring colorfulness in natural images,' in *Human vision and electronic imaging VIII*, International Society for Optics and Photonics, vol. 5007, 2003, pp. 87–95.

[112] W. Sun and R. Guo, 'Test sequences for virtual reality video coding from LetinVR,' *Joint Video Exploration Team (JVET) of ITU-T SG*, vol. 16, 2016.

[113] J. Xiao, K. Ehinger, A. Oliva and A. Torralba, 'Recognizing scene viewpoint using panoramic place representation,' in *IEEE CVPR*, 2012, pp. 2695–2702.

[114] I. BT, '500-14. BT. 500: Methodologies for the subjective assessment of the quality of television images,' *International Telecommunications Union: Geneva, Switzerland*, 2019.

[115] T. Hoßfeld, R. Schatz and S. Egger, 'SOS: The MOS is not enough!' In *IEEE QoMEX*, Mechelen, Belgium, 2011, pp. 131–136. DOI: 10.1109/QoMEX.2011.6065690.

[116] E. Girden, *ANOVA: Repeated measures*. Sage, 1992.

[117] J. Filliben, 'The probability plot correlation coefficient test for normality,' *Technometrics*, vol. 17, no. 1, pp. 111–117, 1975.

[118] W. Kruskal and W. Wallis, 'Use of ranks in one-criterion variance analysis,' *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[119] M. Terpilowski, 'Scikit-posthocs: Pairwise multiple comparison tests in python,' *The Journal of Open Source Software*, vol. 4, no. 36, p. 1169, 2019.

[120] G. Huanga, Z. Liu, L. V. D. Maaten and K. Q. Weinberger, 'Densely connected convolutional networks,' in *IEEE CVPR*, Honolulu, HI, USA, 2017, pp. 4700–4708.

[121] O. Russakovsky, J. Deng, H. Su, J. Krause and S. S. et al., 'ImageNet Large Scale Visual Recognition Challenge,' *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[122] A. Chetouani and L. Li, 'On the use of a scanpath predictor and convolutional neural network for blind image quality assessment,' *Signal Processing: Image Communication*, vol. 89, p. 115 963, Aug. 2020. DOI: 10.1016/j.image.2020.115963.

[123] N. Ahmed and H. Asif, 'Perceptual quality assessment of digital images using deep features,' *Computing and Informatics*, vol. 39, no. 3, pp. 385–409, 2020.

[124] M. Lin, Q. Chen and S. Yan, 'Network in network,' *arXiv preprint arXiv:1312.4400*, 2013.

[125] V. Nair and G. Hinton, 'Rectified linear units improve restricted boltzmann machines,' in *ICML*, 2010.

[126] J. Brownlee, *Deep learning for computer vision: image classification, object detection, and face recognition in python*. Machine Learning Mastery, 2019.

[127] K. He, X. Zhang, S. Ren and J. Sun, 'Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,' in *IEEE ICCV*, Santiago, 2015, pp. 1026–1034.

[128] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse and S. Möller, 'Ndnetgaming development of a no-reference deep CNN for gaming video quality prediction,' *Multimedia Tools and App.*, pp. 1–23, 2020.

[129] L. Bowen, Z. Weixia, T. Meng, Z. Guangtao and W. Xianpei, 'Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception,' *arXiv preprint arXiv:2108.08505*, 2021.

[130] N. Ponomarenko, O. Ieremeiev, V. Lukin and K. E. et al., 'Color image database TID2013: Peculiarities and preliminary results,' in *IEEE EUVIP*, 2013, pp. 106–111.

[131] A. Sendjasni, M. Larabi and F. A. Cheikh, 'Perceptually-weighted CNN for 360-degree Image quality assessment using visual scan-path and JND,' in *IEEE ICIP*, Anchorage, Alaska, 2021, pp. 1439–1443. DOI: 10.1109/ICIP42928.2021.9506044.

[132] L. Li, Z. Li, X. Ma, H. Yang and H. Li, 'Advanced spherical motion model and local padding for 360 video compression,' *IEEE Trans. Image Process*, vol. 28, no. 5, pp. 2342–2356, 2018.

[133] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia and G. Wetzstein, 'Saliency in VR: How do people explore virtual environments?' *IEEE TVCG*, vol. 24, no. 4, pp. 1633–1642, 2018.

[134] C. Li, M. Xu, X. Du and Z. Wang, 'Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model,' in *ACM MM*, Seoul, Republic of Korea, 2018, ISBN: 9781450356657.

[135] L. Kang, P. Ye, Y. Li and D. Doermann, 'Convolutional neural networks for no-reference image quality assessment,' in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1733–1740.

[136] S. Bosse, D. Maniry, K. Müller, T. Wiegand and W. Samek, 'Deep neural networks for no-reference and full-reference image quality assessment,' *IEEE TIP*, vol. 27, no. 1, pp. 206–219, 2018. DOI: 10.1109/TIP.2017.2760518.

[137] T. Truong, H. Tran and T. Thang, 'Non-reference Quality Assessment Model using Deep learning for Omnidirectional Images,' in *IEEE ICAST*, IEEE, Morioka, 2019, pp. 1–5.

[138] H. Sheikh, Z. Wang, L. Cormack and A. Bovik, 'Live image quality assessment database release 2,' *https://live.ece.utexas.edu/research/Quality/subjective.htm*, 2005.

[139] E. Larson and D. Chandler, 'Most apparent distortion: Full-reference image quality assessment and the role of strategy,' *Journal of electronic imaging*, vol. 19, no. 1, p. 011 006, 2010.

[140] Z. Weixia, M. Kede, Z. Guangtao and Y. Xiaokang, 'Uncertainty-aware blind image quality assessment in the laboratory and wild,' *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[141] A. Martın, P. Barham, J. Chen, Z. Chen and A. D. et al., 'Tensorflow: A system for large-scale machine learning,' in *12th {USENIX} Symposium on Operating Systems Design and Implementation*, Savannah, GA, USA, 2016, pp. 265–283.

[142] A. Graves, 'Generating sequences with recurrent neural networks,' *arXiv preprint arXiv:1308. 0850*, 2013.

[143] D. Masters and C. Luschi, 'Revisiting small batch training for deep neural networks,' *arXiv preprint arXiv:1804.07612*, Apr. 2018.

[144] H. G. Kim, H. Lim and Y. M. Ro, 'Deep virtual reality image quality assessment with human perception guider for omnidirectional image,' *IEEE TCSVT*, vol. 30, no. 4, pp. 917–928, 2020. DOI: 10.1109/TCSVT.2019.2898732.

[145] L. Hao, C. Pratik, Y. Hao, L. Michael, R. Avinash, B. Rahul and S. Stefano, 'Rethinking the hyperparameters for fine-tuning,' *arXiv preprint arXiv:2002.11770*, 2020.

[146] C. Zhang, P. Benz, D. Argaw, S. Lee, J. Kim, F. Rameau, J. Bazin and I. Kweon, 'Resnet or densenet? introducing dense shortcuts to resnet,' in *IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, Hawaii, 2021, pp. 3550–3559.

[147] J. Yamanaka, S. Kuwashima and T. Kurita, 'Fast and accurate image super resolution by deep cnn with skip connection and network in network,' in *ICNIP*, Springer, 2017, pp. 217–225.

[148] L. Po, M. Liu, W. Yuen, Y. Li and X. X. et al., 'A novel patch variance biased convolutional neural network for no-reference image quality assessment,' *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1223–1229, 2019.

[149] J. Kim and S. Lee, 'Fully deep blind image quality predictor,' *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2016.

[150] H. Wen and J. Tingting, 'From image quality to patch quality: An image-patch model for no-reference image quality assessment,' in *IEEE ICASSP*, New Orleans, LA, USA, 2017, pp. 1238–1242.

[151] A. Sendjasni, M. Larabi and F. Cheikh, 'Perceptually-weighted CNN for 360-degree image quality assessment using visual scan-path and JND,' in *IEEE ICIP*, Anchorage, AK, USA, 2021, pp. 1439–1443.

[152] A. Sendjasni, M. Larabi and F. Cheikh, 'Convolutional neural networks for omnidirectional image quality assessment: Pre-trained or re-trained?' In *IEEE ICIP*, Anchorage, AK, USA, 2021, pp. 3413–3417.

[153] D. Noton and L. Stark, 'Scanpaths in saccadic eye movements while viewing and recognizing patterns,' *Vision research*, vol. 11, no. 9, 929–IN8, 1971.

[154] W. Sun, Z. Chen and F. Wu, 'Visual scanpath prediction using ior-roi recurrent mixture density network (early access),' *IEEE TPAMI*, pp. 1–1, 2019.

[155] D. Martin, A. Serrano, A. Bergman, G. Wetzstein and B. Masia, 'Scangan360: A generative model of realistic scanpaths for 360° images,' *IEEE TVCG*, vol. 28, no. 5, pp. 2003–2013, 2022. DOI: 10.1109/TVCG.2022.3150502.

[156] F. Chao, C. Ozcinar and A. Smolic, 'Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need,' in *IEEE 23$^{rd}$ MMSP*, Tampere, Finland, 2021, pp. 1–6.

[157] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu and G. Zhai, 'A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison,' in *IEEE MMSP*, Vancouver, BC, Canada, 2018, pp. 1–6.

[158] A. Buslaev and *et al.*, 'Albumentations: Fast and flexible image augmentations,' *Information*, vol. 11, no. 2, 2020, ISSN: 2078-2489. DOI: 10.3390/info11020125. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125.

[159] A. Mittal, A. K. Moorthy and A. C. Bovik, 'No-reference image quality assessment in the spatial domain,' *IEEE TIP*, vol. 21, no. 12, pp. 4695–4708, 2012.

[160] A. Mittal, R. Soundararajan and A. Bovik, 'Making a "completely blind" image quality analyzer,' *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[161] Z. Wang and X. Shang, 'Spatial pooling strategies for perceptual image quality assessment,' in *IEEE ICIP*, Atlanta, GA, USA, 2006, pp. 2945–2948. DOI: 10.1109/ICIP.2006.313136.

[162] A. Moorthy and A. Bovik, 'Visual importance pooling for image quality assessment,' *IEEE journal of selected topics in signal processing*, vol. 3, no. 2, pp. 193–201, 2009.

[163] Z. Wang and Q. Li, 'Information content weighting for perceptual image quality assessment,' *IEEE Trans. on image processing*, vol. 20, no. 5, pp. 1185–1198, 2010.

[164] D. Temel and G. AlRegib, 'A comparative study of quality and content-based spatial pooling strategies in image quality assessment,' in *IEEE GlobalSIP*, FL, USA, 2015, pp. 732–736. DOI: 10.1109/GlobalSIP.2015.7418293.

[165] G. Mingming and P. Marius, 'Spatial pooling for measuring color printing quality attributes,' *Journal of Visual Communication and Image Representation*, vol. 23, no. 5, pp. 685–696, 2012, ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2012.03.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320312000600.

[166] Z. Tu, C. Chen, L. Chen and *et al.*, 'A comparative evaluation of temporal pooling methods for blind video quality assessment,' in *IEEE ICIP*, UAE, 2020, pp. 141–145.

[167] Z. Li, C. Bampis, J. Novak, A. Aaron and et al., 'VMAF: The journey continues,' *Netflix Technology Blog*, vol. 25, 2018.

[168] C. Zewdie, M. Pedersen and Z. Wang, 'A new pooling strategy for image quality metrics: Five number summary,' in *5th EUVIP*, Paris, France, 2014, pp. 1–6.

[169] K. Jongyoo, N. Anh-Duc and L. Sanghoon, 'Deep cnn-based blind image quality predictor,' *IEEE TNNLS*, vol. 30, no. 1, pp. 11–24, 2018.

[170] S. Seo, S. Ki and M. Kim, 'A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions,' *IEEE TCSVT*, vol. 31, no. 7, pp. 2602–2616, 2021. DOI: 10.1109/TCSVT.2020.3030895.

[171] T. Mingxing and L. Quoc, 'Efficientnet: Rethinking model scaling for convolutional neural networks,' in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[172] M. Sonka, V. Hlavac and R. Boyle, 'Image pre-processing,' in *Image Processing, Analysis and Machine Vision*. Boston, MA: Springer US, 1993, pp. 56–111.

[173] D. Heeger, 'Normalization of cell responses in cat striate cortex,' *Visual neuroscience*, vol. 9, no. 2, pp. 181–197, 1992.

[174] R. Li, H. Yang, T. Yu and Z. Pan, 'Cnn model for screen content image quality assessment based on region difference,' in *IEEE 4th International Conference on Signal and Image Processing*, Wuxi, China, 2019, pp. 1010–1014.

[175] J. Kim and S. Lee, 'Deep blind image quality assessment by employing FR-IQA,' in *IEEE ICIP*, Beijing, China, 2017, pp. 3180–3184.

[176] C. Pan, Y. Xu, Y. Yan, K. Gu and X. Yang, 'Exploiting neural models for no-reference image quality assessment,' in *IEEE VCIP*, Chengdu, China, 2016, pp. 1–4.

[177] L. Siwei and E. Simoncelli, 'Nonlinear image representation using divisive normalization,' in *IEEE CVPR*, Anchorage, AK, USA, 2008, pp. 1–8.

[178] M. Rad, P. Roth and V. Lepetit, 'ALCN: Adaptive local contrast normalization,' *Computer Vision and Image Understanding*, vol. 194, p. 102 947, 2020, ISSN: 1077-3142.

[179] S. Pizer, E. Philip, J. Austin and et al., 'Adaptive histogram equalization and its variations,' *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987, ISSN: 0734-189X.

[180] A. Krizhevsky and e. a. G. Hinton, 'Learning multiple layers of features from tiny images,' 2009.

[181] D. Kingma and J. Ba, 'Adam: A method for stochastic optimization,' *arXiv preprint arXiv:1412.6980*, 2014.

[182] A. Sendjasni, M. Larabi and F. Cheikh, 'Visual scan-path based data-augmentation for CNN-based 360-degree image quality assessment,' in *London Imaging Meeting*, Society for Imaging Science and Technology, London, UK, 2021, pp. 21–26.

[183] O. Wiedemann, V. Hosu, H. Lin and D. Saupe, 'Disregarding the big picture: Towards local image quality assessment,' in *$10^{th}$ QoMEX*, Cagliari, Italy, 2018, pp. 1–6.

[184] J. Wu, G. Shi, W. Lin, A. Liu and F. Qi, 'Just noticeable difference estimation for images with free-energy principle,' *IEEE TMM*, vol. 15, no. 7, pp. 1705–1710, 2013. DOI: 10.1109/TMM.2013.2268053.

[185] J. Kim and S. Lee, 'Deep learning of human visual sensitivity in image quality assessment framework,' in *IEEE CVPR*, Honolulu, HI, USA, 2017, pp. 1969–1977.

[186] W. Xue, L. Zhang and X. Mou, 'Learning without human scores for blind image quality assessment,' in *IEEE CVPR*, 2013, pp. 995–1002. DOI: 10.1109/CVPR.2013.133.

[187] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang and C. W. Chen, 'Blind quality assessment based on pseudo-reference image,' *IEEE TMM*, vol. 20, no. 8, pp. 2049–2062, 2018. DOI: 10.1109/TMM.2017.2788206.

[188] K. Ma, W. Liu, T. Liu, Z. Wang and D. Tao, 'DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs,' *IEEE TIP*, vol. 26, no. 8, pp. 3951–3964, 2017. DOI: 10.1109/TIP.2017.2708503.

[189] S. Ioffe and C. Szegedy, 'Batch normalization: Accelerating deep network training by reducing internal covariate shift,' in *ICML*, Lille, France, 2015, pp. 448–456.

[190] S. Woo, J. Park, J. Lee and I. Kweon, 'CBAM: Convolutional block attention module,' in *ECCV*, Munich, Germany, 2018, pp. 3–19.

[191] F. Radenović, G. Tolias and O. Chum, 'Fine-tuning cnn image retrieval with no human annotation,' *IEEE TPAMI*, vol. 41, no. 7, pp. 1655–1668, 2019. DOI: 10.1109/TPAMI.2018.2846566.

[192] Y. Gu, C. Li and J. Xie, 'Attention-aware generalized mean pooling for image retrieval,' *arXiv preprint arXiv:1811.00202*, 2018.

[193] L. Kauffmann, S. Ramanoël, N. Guyader, A. Chauvin and C. Peyrin, 'Spatial frequency processing in scene-selective cortical regions,' *NeuroImage*, vol. 112, pp. 86–95, 2015.

[194] I. . Groen, E. Silson and C. Baker, 'Contributions of low-and high-level properties to neural processing of visual scenes in the human brain,' *Philosophical Trans. of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20 160 102, 2017.

[195] L. Zhang, L. Zhang, X. Mou and D. Zhang, 'FSIM: A feature similarity index for image quality assessment,' *IEEE TIP*, vol. 20, no. 8, pp. 2378–2386, 2011.

[196] R. Ranjan, V. Patel and R. Chellappa, 'Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,' *IEEE TPAMI*, vol. 41, no. 1, pp. 121–135, 2017.

[197] P. Huber, *Robust statistics*. John Wiley & Sons, 2004, vol. 523.

[198] C. Xia, F. Qi and G. Shi, 'Bottom-up visual saliency estimation with deep autoencoder-based sparse reconstruction,' *IEEE TNNLS*, vol. 27, no. 6, pp. 1227–1240, 2016.

[199] Z. Wang, E. Simoncelli and A. Bovik, 'Multiscale structural similarity for image quality assessment,' in *The 37$^{th}$ ACSSC, 2003*, IEEE, vol. 2, 2003, pp. 1398–1402.

[200] X. Min, G. Zhai, K. Gu, Y. Liu and X. Yang, 'Blind image quality estimation via distortion aggravation,' *IEEE TB*, vol. 64, no. 2, pp. 508–517, 2018.

[201] JVET, 'Algorithm description of joint exploration test model 6 (JEM6),' ITU-T VCEG (Q6/16) and ISO/IEC MPEG (JTC 1/SC 29/WG 11), Tech. Rep. JVET-F1001, 2017.

[202] A. Sendjasni and M. Larabi, 'SAL-360IQA: A saliency weighted patch-based cnn model for 360-degree images quality assessment,' in *IEEE ICMEW*, Taipei City, Taiwan, 2022, pp. 1–6.

[203] A. Sendjasni, M. Larabi and F. Cheikh, 'Convolutional neural networks for omnidirectional image quality assessment: A benchmark,' *IEEE TCSVT (Early Access)*, pp. 1–1, 2022. DOI: 10.1109/TCSVT.2022.3181235.

[204] A. Sendjasni, M. Larabi and F. Cheikh, 'On the Influence of Head-Mounted Displays on Quality Rating of Omnidirectional images,' in *Electronic Imaging, Image Quality and System Performance XVIII*, Burlingame, California USA, 2021, pp. 296–1.

[205] A. Sendjasni, M. Larabi and F. Cheikh, 'Visual Scan-Path based Data-Augmentation for CNN-based 360-degree Image Quality Assessment,' in *London Imaging Meeting*, vol. 2021, 2021, pp. 21–26.

[206] A. Sendjasni, D. Traparic and M. Larabi, 'Investigating normalization methods for cnn-based image quality assessment,' in *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022, pp. 4113–4117.

[207] A. Sendjasni and M. Larabi, 'Transfer learning from vision transformers or convnets for 360-degree images quality assessmentf,' in *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022, pp. 4133–4137.

[208] Z. Chen, J. Xu, C. Lin and W. Zhou, 'Stereoscopic omnidirectional image quality assessment based on predictive coding theory,' *IEEE JSTSP*, vol. 14, no. 1, pp. 103–117, 2020. DOI: 10.1109/JSTSP.2020.2968182.

[209] J. Xu, C. Lin, Z. Zhou and Z. Chen, 'Subjective quality assessment of stereoscopic omnidirectional image,' in *Pacific Rim Conference on Multimedia*, Springer, 2018, pp. 589–599.

[210] X. Chai, F. Shao, Q. Jiang, X. Meng and Y. Ho, 'Monocular and binocular interactions oriented deformable convolutional networks for blind quality assessment of stereoscopic omnidirectional images,' *IEEE TCSVT*, vol. 32, no. 6, pp. 3407–3421, 2022. DOI: 10.1109/TCSVT.2021.3112120.

[211] Y. Qi, G. Jiang, M. Yu, Y. Zhang and Y. Ho, 'Viewport perception based blind stereoscopic omnidirectional image quality assessment,' *IEEE TCSVT*, vol. 31, no. 10, pp. 3926–3941, 2021. DOI: 10.1109/TCSVT.2020.3043349.

[212] X. Zhou, Y. Zhang, N. Li, X. Wang, Y. Zhou and Y. Ho, 'Projection invariant feature and visual saliency-based stereoscopic omnidirectional image quality assessment,' *IEEE Transactions on Broadcasting*, vol. 67, no. 2, pp. 512–523, 2021. DOI: 10.1109/TBC.2021.3056231.

[213] J. Xu, Z. Luo, W. Zhou, W. Zhang and Z. Chen, 'Quality assessment of stereoscopic 360-degree images from multi-viewports,' in *Picture Coding Symposium (PCS)*, 2019, pp. 1–5. DOI: 10.1109/PCS48520.2019.8954555.