EPFL

Thèse n° 9936

Image Compression and Quality Assessment using Convolutional Neural Networks

Présentée le 23 octobre 2020

à la Faculté des sciences et techniques de l'ingénieur
Groupe Ebrahimi
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Pinar AKYAZI

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury
Prof. T. Ebrahimi, directeur de thèse
Prof. F. Pereira, rapporteur
Dr M. Rerabek, rapporteur
Prof. J.-Ph. Thiran, rapporteur

École polytechnique fédérale de Lausanne

2020

Everybody is a genius.
But if you judge a fish by its ability to climb a tree,
it will live its whole life believing that it is stupid.
— *Anonymous*

Acknowledgements

My journey in EPFL lasted for almost six years and spanned two labs, while I was accompanied by many individuals who carried me further each day with invaluable support and encouragement. First I would like to thank Prof. Touradj Ebrahimi for giving me the chance to work with him in MMSPG during the last two years. Thank you for making this thesis possible by taking a chance on me, for having faith and trust, for guiding me towards novelties and growth. This journey, although shorter than most others', has changed my perspectives in so many ways and shaped me into the person I am today. Thank you for all the opportunities you've provided that encouraged this change.

I would also like to thank Prof. Pascal Frossard, Prof. Fernando Pereira, Prof. Jean-Philippe Thiran and Dr. Martin Rerabek for accepting my invitation and taking part in the evaluation of my work. Thank you for your time, patience and efforts.

Pascal, this wouldn't be possible without you and my LTS4 years full of drama. Thank you for putting up with my teenage years. Here I also need to extend my gratitude to Prof. Bülent Sankur, Prof. Kadri Özçaldıran, Prof. Murat Saraçlar and Prof. Burak Acar, for their invaluable efforts in my upbringing as a student.

I always reminisced my elementary school years as the best years of my life. My time in EPFL is the first period in my life that managed to be as precious, thanks to the friends who have joined me during this journey. Irenemu, I love you so so much. It's been a privilege working close to you, giggling (or piggling) all the time and sharing all those incredible memories with you. Without your support and kisses on my hair, I would be lost long ago. Vaggelis, Anne-Florimu, Tolis and Eugene, thank you for the laughter and witty jokes, and being there for me through all. The corridor girls, I love each and every one of you. Hermina (my little Zec) and Helena, I'll never forget the laughs we shared, which I believe are the loudest on the planet. I hope to share many more with you (but Hermina doesn't return calls a lot so I'll have to call a million times). My first officemates Laura and Renata, and my corridor neighbor Dorina whom I sometimes shocked with my loud cries, thank you for brightening my days at work.

Beril, you are one of the most incredible people I've met in my life, a priceless gem that is very difficult to describe with mere words, especially here within a limited space. I cannot thank you enough for being my friend, companion and confidant. I wouldn't be the same person as I am today without your care and support. You are the kindest, warmest part of this whole experience and I'm so lucky to have you.

The rest of the unforgettable Turkish gang, a huge bunch of thanks to Alp, Atmed, Aliş, Fatit, Cey and Okan for making me laugh like nobody ever could, for watching the stupidest YouTube

videos and learning their scripts by heart with me. The trend started with Alp and Semih as Ağababa and me taking out my earphones at least 3 times within any conversation in Satellite or Great Escape. Thank you guys for participating in all those wonderful moments. I hope to share many more with you all.
Nicolas, how could I ever graduate without you? Thank you for being my rock and kick start in most difficult tasks, and being a dear friend to me. Thank you for showing me the way and the incredible kindness you bear.
Thank you Adrien, for also being my rock during the toughest times, encouraging me to be strong and trust myself. Also for making me friends with mountains, and bringing peace to my life.
My friends in the other side of the corridor; Vlad, Kostas, Andreas, Benjamin and Nikos, thank you for lighting my days up. A special thanks to my dear Rodi (or Bilù) for making our house the warmest home, putting up with my craziness and being the best flatmate. And of course Tilly, who is also the best and by far the coolest flatmate, thank you for all the memorable times we shared in our little crib.
My girls all the way back from elementary school; Burçin, Sera and Sara Pınar, you have been a part of this journey as always. In moments of despair, I always turn to you. Thank you for your constant support and love, and never losing faith in me.
The other heroes in this journey are Christine, Anne and Rosie, our valuable administrative assistants whom I cannot thank enough for all their support and encouragement. A smile in the morning makes a huge difference, thank you all for never withholding that. And coffee, of course.
Two other special heroes that changed my life, Dr. Georges Gabris and Dr. Nicolas Othenin-Girard, thank you for being there literally every step of the way, for helping me grow up and be the person I am today.
Cutie pie, I don't think this work could ever be complete without you. Nor would *I* be complete without you. Thank you for sharing your heart and your life with me.
And above all, my parents. For bearing with me during this long journey, for sharing every fear and joy, for never giving up on me and demonstrating unconditional love day after day. Sizi yol boyu kadar çok seviyorum.

Lochristi, December 24, 2019

Abstract

The rapid development of digital imaging and video has placed visual contents in the heart of our lives. Digital multimedia span a vast number of areas from business to leisure, including but not limited to education, medicine, accessibility, training, advertisement, entertainment and social networks. The dominance of visual multimedia has created an increasing need for broadcasters and service providers to present contents of superior visual quality while keeping the storage and transmission costs as low as possible. Before finally being presented to users, all contents are processed for transmission, which reduces the quality depending on the characteristics of the processes involved. Besides enhancement methods applied as preprocessing and post-processing, compression is the key step of content delivery. Image and video processing communities have been proposing improved solutions to the multimedia compression problem for decades, using mathematical transforms, augmenting human visual system responses, and finally, incorporating deep neural networks.

What distinguishes the proposed solutions from each other is two fold: one characteristic is the solution architecture, whereas the other aspect is how the solution performs. The performance of image and video compression models can be measured objectively and subjectively, with the latter emphasizing the quality of the content perceived by users. Both when developing and employing compression technologies, providers need to assess the end quality of their product. How this quality is estimated and measured is of key importance.

Standardized psychophysical experiments measure the subjective quality of images and video, with the requirement of the participation of many human subjects. Objective quality assessment methods seek to provide a better alternative by accommodating no human costs at computation time, yet still predicting quality with high accuracy when compared to viewers' opinion. An efficient compression method ideally needs to employ a strong objective metric to measure the impact of degradations effectively, thereby maximize algorithm performance by achieving an optimal rate-distortion trade-off.

In this work, the problem of constructing an end-to-end image compression system using an objective metric with high correlation to subjective ratings is addressed. First, the challenges of building an effective objective metric are discussed and multiple learning-based solutions using convolutional neural networks are proposed. For that means, the construction of a comprehensive database is presented, which involves mean opinion scores of compressed high resolution images, obtained via subjective quality assessment experiments. Afterwards, traditional transform-based codecs are investigated along with recent improvements as well as their learning-based counterparts, leading to the construction of novel end-to-end com-

pression models using convolutional neural networks. The proposed autoencoders initially employ state-of-the-art objective metrics in their cost function. As a final step, overall loss of the compression model is modified to include the aforementioned learning-based objective metric, combining the compression and quality assessment solutions proposed in this work. The presented approaches provide improvements and novel insights to the state of the art both in the domains of image quality assessment and learning-based image compression.

Keywords: Full reference image quality assessment, subjective quality assessment, objective quality assessment, objective metric, learning-based image compression, convolutional neural networks.

Résumé

Les nombreux progrès dont nous sommes témoins dans le domaine numérique placent les contenus audiovisuels (images et vidéos) au centre de nos vies. Nous les retrouvons actuellement dans une multitude de domaines différents, qui relèvent de notre vie privée comme professionnelle. Pour en citer quelques uns, les supports numériques sont très largement répandus dans l'éducation, la médecine, la publicité, les divertissements et les réseaux sociaux. Cette surexposition en constante augmentation force les professionnels de l'audiovisuel, fournisseurs d'accès et de services principalement, à proposer des contenus de toujours meilleur qualité tout en maintenant des coûts de stockage et de transmission à moindre frais. Cependant, préalablement à la visualisation, chaque contenu est traité et manipulé pour satisfaire les contraintes de transmissions. En conséquences, en fonction des caractéristiques propres au contenu, des artefacts diminuant la qualité perceptuelle sont introduits. Malgré de nombreuses solutions palliant à ces artefacts, souvent du pré ou post-traitement, la compression est toujours l'opération déterminante lors de la transmission de contenus. La communauté scientifique du traitement du signal audiovisuel propose depuis des décennies maintenant des solutions au problème de compression. Ont été envisagés par exemple différentes transformées mathématiques, certains algorithmes sont conçu sur le modèle de la perception visuelle humaine, et récemment, l'introduction de réseaux d'apprentissage profonds a montré des résultats bien plus qu'encourageant.

Nous pouvons distinguer les solutions de compression à deux niveaux. Le premier est la structure même de l'architecture du modèle, le second est sa performance. Plus précisément, la performance des systèmes de compression d'image fixes et de vidéos peut être évaluée objectivement et subjectivement, cette dernière mettant l'accent sur la qualité telle qu'elle est perçue par l'utilisateur final. Ces mesures sont capitales car elles sont impliquées tant lors du développement que de l'utilisation d'outils de compression. Il est donc très important de savoir et d'étudier précisément comment la qualité est mesurée et prédite.

Des études psychophysiques, satisfaisant les normes établies, sont menées sur de nombreux sujets humains pour mesurer subjectivement la qualité d'images et vidéos. Les mesures objectives, elles, tendent à s'affranchir des coûts humains, tant sur le plan de l'intervention de sujets que sur le temps de réalisation des expériences. Elles représentent toutefois une alternative efficace pour prédire de manière précise l'opinion de l'utilisateur final. Une méthode de compression performante doit idéalement utiliser une métrique objective de pointe pour mesurer l'impact des dégradations, maximisant ainsi les performances de l'algorithme en réalisant un compromis entre distorsion et débit.

Résumé

Ces travaux abordent les problèmes rencontrés lors de la construction de l'entière pipeline de compression quand une métrique objective hautement corrélée avec les scores subjectifs est utilisée. Dans un premier temps, les multiples challenges relatifs à la conception d'une métrique objective fiable sont examinés, puis surmontés par nos propositions, solutions basées sur l'apprentissage profond de réseaux de neurones convolutionnels. Pour ce faire, une nouvelle base de donnée, comprenant les scores d'opinion moyens relatifs à des images compressées de haute résolution, est constituée lors d'un test subjectif sur la qualité. Ensuite, l'état de l'art, qui inclue les traditionnels codecs basés sur la transformation fréquentielle, leurs récentes améliorations, ainsi que leurs alternatives basées apprentissages, est établit et ouvre la voie à de nouveaux modèles de bout en bout utilisant des réseaux de neurones convolutionnels. Les auto-encodeurs initialement proposées emploient des mesures objectives de l'état de l'art dans leur fonction de coût. En dernier lieu, la fonction de coût du modèle de compression est changée au profit de la métrique objective basée apprentissage profond susmentionnée. Cette solution combine les solutions développées au cours de nos travaux en terme de compression et d'évaluation de qualité. Les approches présentées apportent des améliorations substantielles à l'état de l'art et apportent de nouvelles perspectives dans les domaines de l'évaluation de la qualité d'image et de la compression d'images basée apprentissage profond.

Mots-clés : Évaluation de qualité d'image avec référence complète, évaluation de qualité subjective, évaluation de qualité objective, compression d'image basée apprentissage, réseaux de neurones convolutionnels

Contents

List of Figures

List of Tables

List of Acronyms

2D two-dimensional. 11, 23, 30, 38, 40, 80, 90, 96, 108, 135, 136

3D three-dimensional. 22

ACR Absolute Category Rating. 15

ADMM Adaptive Direction Method of Multipliers. 22

AOM Alliance for Open Media. 18, 19

AVC/H.264 Advanced Video Coding. 18

AVIF AV1 Image File Format. 18, 19

BPG Better Portable Graphics. 18, 21, 23

BSD Berkeley Software Distribution. 17, 18

CABAC Context-Adaptive Binary Arithmetic Encoder. 91

CAE Convolutional Autoencoder. 21, 90, 91, 137

CAE-LM Convolutional Autoencoder with Learning-Based Objective Metric. vii, xii, xiv, 124–132, 137, 138

CI confidence interval. 10, 14, 51, 66

CIBR Content-Based Image Retrieval. 61

CIE International Commission on Illumination. 11

CLIC Workshop and Challenge on Learned Image Compression. 22, 90, 93, 94, 96, 112, 125, 126, 136

CNN Convolutional Neural Network. 19–23, 51, 80, 89, 96

CSF Contrast Sensitivity Function. 12

CSIQ Laboratory of Computational and Subjective Image Quality. x, xiii, 15, 24, 33, 34, 36–40

CTC Common Test Conditions. 60, 64, 75

CTU Coding Tree Unit. 18

CVPR Conference on Computer Vision and Pattern Recognition. 22, 61, 90

CW-SSIM Complex Wavelet Structural Similarity Index. 11

DCT Discrete Cosine Transform. 12, 17, 19, 99

DeepSim Deep Similarity. 13

DIQaM-FR Deep Image Quality Measure. 13, 27, 28

DLM Detail Loss Metric. 12

DMOS Difference Mean Opinion Score. 14, 15, 33, 34, 37

DN Divisive Normalization. 12

DOG Difference of Gaussians. 13, 23

DSCQS Double Stimulus Continuous Quality Scale. 10

DSIS Double-Stimulus Impairment Scale. xiii, 9, 10, 48, 50, 65, 75, 136

DSSLIC deep semantic segmentation-based layered image compression. 23

DVD Design Viewing Distance. 9

DWT Discrete Wavelet Transform. 11, 17, 23, 30, 38, 40, 80, 96

EBCOT Embedded Block Coding with Optimized Truncation. 17

EPFL Ecole Polytechnique Fédérale de Lausanne. x, xiii, 43, 45, 51, 58, 59

FactMS-SSIM Factorized Entropy Model with Multi-scale Structural Similarity Index Loss. xi, 63, 64, 66, 72, 73, 123

FactMSE Factorized Entropy Model with Mean Squared Error Loss. xi, xii, xiv, 63, 64, 66, 72–74, 113, 116, 117, 119, 120, 137

FR Full Reference. 8, 10, 12–14, 138

FRICwRNN Full Resolution Image Compression with Recurrent Neural Networks. xi, 62, 64, 66, 71–74

FSIM Feature Similarity Index. ix, xiii, 11–13, 23, 34–36, 78, 82

GAN Generative Adversarial Network. 19–21, 23, 89

GDN Generalized Divisive Normalization. 21, 90–93

GM Gradient Magnitude. 12

GRDN Grouped Residual Dense Network. 23

GRU Gated Recurrent Unit. 21, 62

HD High Definition. 18, 22, 24, 40, 41, 45, 61, 75, 85, 93, 136

HDR High Dynamic Range. x, 17, 44, 45, 47–51, 53, 56–59, 62, 74, 75, 136

HEIF High Efficiency Image File Format. 18

HEVC/H.265 High Efficiency Video Coding. 18, 23, 66, 120, 123

HVS Human Visual System. 8, 11–13, 17, 19, 21, 23, 27, 29, 74, 123

HyperMS-SSIM Entropy Model with Scale Hyperprior using Multi-scale Structural Similarity Index Loss. xi, 64, 66, 72, 73, 123

HyperMSE Entropy Model with Scale Hyperprior using Mean Squared Error Loss. xi, 63, 64, 66, 73, 74, 123

IFC Image Fidelity Criterion. 11

IGDN Inverse Generalized Divisive Normalization. 92

IQA Image Quality Assessment. 8, 12, 14, 15, 23, 24, 33, 40, 41, 44, 61, 62, 74, 75, 77, 80, 84, 112, 136, 138, 139

IQM Image Quality Metric. 8, 24, 35, 40, 41, 77, 78, 136, 139

ITU International Telecommunication Union. 9, 18

IW-MSE Information Content Weighted MSE. 12

IW-PSNR Information Content Weighted PSNR. 12

IW-SSIM Information Content Weighted SSIM. 12

JND Just-Noticeable Difference. 15

JPEG Joint Photographic Experts Group. 16, 17, 24, 44, 71

JPEG AI JPEG Ad Hoc Group on Learning-based Image Coding. 60

JVET Joint Video Experts Team. 18

LBT Lapped Biorthogonal Transform. 17

Leaky ReLU Leaky Rectified Linear Unit. 30, 32

LIVE Laboratory for Image & Video Engineering. xiii, 14, 24, 33, 34, 36–40, 83

LSTM Long Short-Term Memory. 20–22

MAD Most Apparent Distortion. 12

MCL-JCI Media Communications Lab - JND-based Coded Images. 15

MOS Mean Opinion Score. ix–xi, 10, 14, 15, 22, 24, 32–38, 40, 51, 52, 54–58, 66, 69–71, 74, 75, 77–80, 84

MPEG Motion Picture Experts Group. 18

MS-SSIM Multi-scale Structural Similarity Index. ix, xi, xiii, xiv, 11, 21–24, 34–36, 48, 49, 63–68, 72, 74, 75, 78, 79, 82, 83, 90, 94–96, 102, 108, 111, 113, 114, 116, 123, 136

MSE Mean Squared Error. 10, 12, 13, 21, 63, 66, 74, 77, 81, 90, 92, 94, 96, 102, 113, 123, 125–129, 131, 136, 137

MTurk Amazon Mechanical Turk. 14

NoWCAE No-Wavelet Convolutional Autoencoder. xiv, 94, 96, 136

NR No Reference. 8, 138

NSS Natural Scene Statistics. 11, 23, 60

PC Phase Congruency. 12

PLCC Pearson Linear Correlation Coefficient. xiii, xiv, 28, 29, 34–37, 39, 40, 58, 59, 78, 79, 82–84, 123

PNG Portable Network Graphics. 21, 62

PQ Perceptual Quantizer. 48

PSNR Peak Signal-to-Noise Ratio. ix, xi–xiv, 10, 12–14, 21–24, 34–36, 48, 49, 64–68, 74, 78, 82, 83, 90, 94–96, 102, 111, 113, 116, 123, 124, 126–129, 131

PVD Preferred Viewing Distance. 9

QILV Quality Index based on Local Variance. 11

R-D Rate-Distortion. 8, 17, 20, 22, 74

ReLU Rectified Linear Unit. 32

ResMixWCAE Residual Mixed Wavelet-Based Convolutional AutoEncoder. vii, xi, xii, 24, 99, 100, 102–109, 111, 123, 137

ResWCAE Residual Wavelet-Based Convolutional AutoEncoder. vii, xi, xii, 24, 99, 100, 102–109, 111, 123, 132, 137

RMSE Root Mean Squared Error. 58, 59

RNAB Residual Non-Local Attention Block. 22

RNN Recurrent Neural Network. 19, 21, 89

RR Reduced Reference. 8

SD Standard Definition. 15, 24, 44, 45, 61, 62, 75, 77, 85, 112, 125, 132, 136

SDR Standard Dynamic Range. x, xiii, 44–55, 58, 59, 62, 66, 74, 75, 136

SDSCE Simultaneous Double Stimulus for Continuous Evaluation. 10

SNR Signal-to-Noise Ratio. 10

SROCC Spearman Rank Order Correlation Coefficient. xiii, xiv, 13, 28, 29, 34–37, 39, 40, 58, 59, 78, 79, 82–84

SS Single Stimulus. xiii, 9, 10, 15, 75

SSCQE Single Stimulus Continuous Quality Evaluation. 10

SSIM Structural Similarity Index. ix, xiii, 11, 13, 20, 22, 23, 34–36, 48, 49, 64, 65, 67, 68, 78, 82–84, 102, 123, 136

SVD Singular Value Decomposition. 13

SVM Support Vector Machine. 12

SVR Support Vector Regression. 13

TID2008 Tampere Image Database 2008. 14

TID2013 Tampere Image Database 2013. ix, xi, xiii, xiv, 14, 23, 24, 33–40, 75, 77–80, 82–84, 125, 136

TPT Télécom ParisTech. x, xiii, 45, 51, 58, 59

UHD Ultra High Definition. 24, 44, 61, 62, 75, 77, 85, 112, 113, 125, 132, 136

UQI Universal Quality Index. 11, 12

VAE Variational Autoencoder. 20

VCEG Video Coding Experts Group. 18

VIF Visual Information Fidelity. 11, 12, 48, 49, 64, 65, 67, 68, 82, 83, 102

VMAF Video Multimethod Assessment Fusion. 12, 48, 49, 64–68, 82, 102, 123

VSNR Visual Signal-to-Noise Ratio. 12

VUB Vrije Universiteit Brussel. x, xiii, 45, 51, 58, 59, 66

VVC Versatile Video Coding. 18, 23, 120

WaDIQaM-FR Weighted Average Deep Image Quality Measure. ix, xiii, xiv, 13, 23, 27, 28, 34–40, 82–84, 132, 136

WCAE Wavelet-Based Convolutional AutoEncoder. vi, xi, xiv, 23, 90–94, 96, 97, 99–102, 108, 123, 136, 137

WCG Wide Color Gamut. 45

WIQM Wavelet-based Image Quality Metric. v, ix, xi, xiii, xiv, 29–41, 75, 77–84, 90, 135, 136

WSNR Weighted Signal-to-Noise Ratio. 12

1 Introduction

Visual data is a prominent element of today's world governed by multimedia. Starting with the seeds of photography in the 19th century, visual data transmission solutions have improved both in terms of variety and quality at a vast speed. Photographic film had the advantage of reaching higher spatial resolutions than any other type of imaging detector and a wider dynamic range than most digital detectors due to its logarithmic response to light. The digitization of imagery, on the other hand, provided many benefits compared to photographic images, including easier calibration for photometry, re-usability, decreased calibration sensitivity to exterior conditions such as humidity and temperature and ease of delivery as laboratory processing was no longer a requirement.

Modern day photography is dominated by digital imagery in many areas from business to leisure, including but not limited to education, medicine, accessibility, training, advertisement, entertainment and social networks. A key aspect in the delivery of digital imagery is the trade-off between visual quality and costs of storage and transmission. The term visual quality in the context of imagery signifies the degree of excellence of the transmitted contents to the end user. With improving digital image sensors it is possible to capture images of higher quality, however, storing or transmitting images without any form of compression reduces the quality of experience in other ways. Without optimizing storage, it becomes impossible to store vast amounts of data as needed, and the delivery of digital contents may present long delays in the absence of optimization of transmission costs.

Image compression is a field of research with the aim of decreasing storage and transmission costs by exploiting the redundancies in images, while ensuring minimum visual quality impairment. In order to achieve this goal, the questions that need to be answered are as follows:

1. What is visual quality?

2. How is visual quality measured?
3. How does an image compression solution optimize the trade-off between visual quality and costs?

As the goal of image compression is to maximize the visual quality, a first step is to define visual quality and design a robust way of measuring this quality. The definition of visual quality can change depending on its reference point, i.e. visual quality can be defined as a fidelity measure between the original image and the compressed image in terms of pixel values, pixel statistics, contrast or frequency content. The distributions of these attributes may vary locally and globally. Depending on the characteristics involved in the definition of visual quality, objective metrics may be designed to assess the degree of preservation of the selected characteristics. Image compression solutions then seek to optimize the trade-off between visual quality and costs in terms of the preferred quality measure.

When designing an objective metric, another important aspect is to ensure sufficient correspondence between the objective assessment and subjective opinion of human viewers. Ensuring fidelity between a reference image and a compressed image in terms of average pixel values, for example, does not necessarily translate into a compressed image of high visual quality in the eyes of an observer. This nonlinear relationship between subjective and objective image quality assessment presents another challenge in the design of objective quality metrics, and in turn, image compression solutions. An ideal quality metric is then defined as not only a local or global measure of the correspondence of one or multiple image attributes between reference and distorted contents, but also as a measure of this correspondence in terms of subjective opinion.

Recent advances in machine learning, particularly the use of deep neural networks, have presented results reaching human accuracy in image related tasks concerning object recognition and classification. Recently, a top 5 classification accuracy of 98% was reached [Touvron et al. (2019)] on the ImageNet dataset [Deng et al. (2009)]. Such developments exemplify that machines can be trained to perform related tasks such as image quality assessment and image compression.

Learning-based deep models extract low level features from images using linear and nonlinear operators, and transform these features to yield an objective such as a compressed code of an image or the objective quality score between two images. An end-to-end learning-based image compression solution needs to transform the image into a representation much smaller in size, and then revert this operation as to maximize the quality between the input and the decoded image. Similarly, a learning-based metric needs to analyze the local features of the image and transform them into quality ratings. These two models can also work together, i.e. a compression architecture may employ a learning-based quality metric as a measure of distortion between the reference and decoded images.

In this work, the challenges of objective quality assessment and compression of images are

tackled using machine learning, in particular, convolutional neural networks. First, current image quality assessment and compression methodologies are analyzed. Afterwards, novel solutions are designed to improve the state-of-the-art performances. The presented work on objective quality assessment focuses more on the evaluation of the quality of compressed images at various resolutions. A comprehensive database for training such a metric is prepared and presented within this thesis. Following, novel compression solutions are introduced that particularly investigate the influence of preprocessing steps and network architectures on the compression performance in terms of various objective metrics. Finally, a convolutional autoencoder that employs a learning-based metric is implemented, and the results of the proposed solutions are discussed together with benefits and shortcomings.

1.1 Contributions

The thesis is separated into three main parts, where each part focuses on a different challenge and presents related contributions. Prior to the presentation of novel solutions related to image quality assessment and compression, a broad survey on the state of the art is delivered concerning image quality assessment methodologies, traditional objective metrics and more recent learning-based models, state-of-the-art image quality assessment databases, and transform-based, hybrid and learning-based compression solutions. The succeeding contributions concerning the main parts of the thesis are analyzed within following subsections.

1.1.1 Learning-based image quality assessment

A novel full-reference objective quality assessment metric is proposed using convolutional neural networks, which employs two dimensional wavelet decomposition as a preprocessing step. Involving a three scale wavelet decomposition explicitly provides spatial and frequency contents of the images as input to the network. The features of the reference and distorted images are extracted from their wavelet coefficients instead of pixel values. The proposed metric is able to perform better than the state-of-the-art metrics on the test set of the training database, with limited generalization ability on cross-database evaluations. The favorable effect of wavelet decomposition as a preprocessing step for learning-based image quality assessment is demonstrated.

Analysis of the cross-database performances highlight the absence of a comprehensive database for image quality assessment on high resolution data. Considering the main focus of this work and the costs of constructing such database with various types and levels of distortion, a novel database is introduced that is composed of standard to ultra-high resolution references compressed using different codecs at several bitrates. Subjective ratings of the compressed images were gathered through subjective quality assessment tests. The performance of state-of-the-art image coding solutions are analyzed objectively and subjectively as part of the construction of this database, and results are reported simultaneously.

Following the construction of a new database, novel image quality assessment models are trained. The final model is optimized to evaluate the quality of compressed images of varying resolutions, within a full-reference framework. The proposed model outperforms numerous state-of-the-art image quality assessment metrics on the test set separated from the training database, as well as on cross-database evaluations concerning compressed images.

1.1.2 Learning-based image compression

A novel convolutional autoencoder is designed that employs wavelet decomposition as a preprocessing step. This preprocessing is motivated both by the use of wavelet decomposition in image compression literature, and by the contribution of wavelet decomposition to the performance of the image quality assessment metrics, showing that an effective representation of the image can be achieved via this step. The benefits of the proposed method at low rates are demonstrated through the analysis of the results in comparison to a similar architecture that does not employ such preprocessing. It is highlighted that the proposed model not only enhances the quality of compressed images at low bitrates but also enables achieving quality trade-offs at similar rate points by using different combinations of outputs from separate wavelet scales.

Two novel architectures are presented based on the previous architecture, as to extend the preceding model. The first model employs a wider bottleneck for increasing rate points, as well as an increased number of convolutional layers and residual connections to improve performance. Mixed kernel sizes are involved in a second version of this architecture to investigate the effects of local spread of convolutional filters on the compression performance. It is verified that despite the increased number of layers and residual connections, the model with the wide bottleneck cannot outperform the preceding model at the operating point of the latter. On the other hand, the use of one unique kernel size at all branches are shown to outperform the equivalent model with mixed kernels. It is shown that the multi-scale perspective provided by wavelet transform does not benefit from further extensions such as employing mixed kernels.

While mixed kernels are shown to be ineffective for the preceding model, their influence in the pixel domain is explored through an ablation study. In particular, the effects of varying kernel sizes and their respective contributions to a fixed sized bottleneck is investigated. Through objective and subjective analysis, the advantages and drawbacks of using varying kernel sizes within an autoencoder are analyzed thoroughly. Valuable insight concerning the design of context dependent autoencoders and compression modes is gained through the conducted study.

1.1.3 Towards unified learning-based image compression solutions

Having developed an effective objective metric for assessing the quality of compressed images and presented numerous autoencoder architectures, an analysis on the ultimate goal of creating a unified system that employs learning-based compression and learning-based quality evaluation is presented. Selected models from previously introduced compression and quality evaluation solutions are combined to measure the distortion between reference and compressed image using the learned metric as part of the autoencoder loss. The influence of the learned metric on improving compression performance is analyzed objectively and subjectively. Although the models are able to perform their respective compression and quality evaluation tasks separately, the influence of the learning-based metric on the autoencoder is shown to present adverse effects on the visual quality of the decoded image with respect to the reference, if not guided by a second metric to ensure pixelwise similarity. The analysis of results provide important perspectives on how to design the components of a unified system that is able to address the presented shortcomings.

1.2 Organization of the thesis

The rest of this thesis is organized as follows:

Chapter 2 presents the state of the art related to the topics analyzed in this work. First, the definition and types of quality assessment are explored. After the explanation of preliminary concepts, subjective image quality evaluation methodologies are presented, followed by the literature on full reference objective image quality assessment. The state of the art on full reference image quality assessment databases is introduced in relation to developing and benchmarking objective image quality metrics. Afterwards, image compression literature is visited and a thorough summary on transform-based and learning-based codecs is presented. Chapter 2 also provides perspectives on the motivations of this work in detail.

Part I focuses on developing image quality metrics using convolutional neural networks. In particular, Chapter 3 presents a novel architecture for objective image quality evaluation, which is trained on a database involving various types and levels of distortion. The proposed model is shown to achieve state-of-the-art correlation within the test database, but is not as effective on cross-database evaluations. The reduced effectiveness is partly induced due to the lack of a proper mapping across distortion types and levels of distinct databases. As a response to this shortcoming, a new compression database for image quality evaluation is presented in Chapter 4. The database is used for training and testing a novel image quality model in Chapter 5, which is able to assess the quality of compressed images more efficiently than its state-of-the-art competitors.

Part II analyzes learning-based compression solutions from various perspectives. Specifically, low-rate image compression challenge is tackled in Chapter 6, using wavelet decomposition as a preprocessing step. Extensions to this model are implemented in Chapter 7 so that a higher

number of rate-distortion trade-offs are achieved. In addition, the effect of using increasing kernel sizes to process coarser to finer scales is investigated. Following the shortcomings of compounding wavelet decomposition and mixed kernels, the sole influence of combining varying kernel sizes within autoencoders that operate with pixel valued inputs is explored in Chapter 8.

Part III is directed towards the future of learning-based image compression solutions that employ learning-based image quality metrics as a measure of distortion between original and decoded images. The image quality metric presented in Chapter 5 is used as the distortion measure of an autoencoder presented in Chapter 8. The analysis on the results of such configuration provides insights on how learning-based metrics need to be trained in order to achieve unified learning-based autoencoders with high compression efficiency.

The conclusion is delivered in Chapter 10, which summarizes the achievements of the presented work as well as the shortcomings encountered, and provides suggestions for future directions concerning learning-based image quality evaluation and compression.

2 Relevant work in image quality assessment and compression

2.1 Definition and types of image quality assessment

Digital media has been a governing an important part of our lives for decades, and keeps gathering increasing attention. Receiving high quality contents is one of the top elements in a consumer's requirements list. Broadcasters and service providers try to meet high quality standards that can keep up with the developments in hardware. With the target of providing high quality content comes two important questions: what is the definition of image quality, and how to measure image quality?

Image quality is the degree of excellence of images, determined by the characteristics of visual attributes. Image quality can be defined as a standard of accuracy, fidelity or visual pleasance. Such distinct definitions give way to multiple categories of quantifying image quality. The two main classes of image quality assessment comprise of subjective and objective methods, distinguished by the nature of their scoring procedures. While subjective methodologies rely on the opinions of human viewers, objective evaluations are governed by mathematical models.

Subjective quality assessment methodologies employ human subjects and evaluate the quality of content with respect to viewers' opinion. Procedures for subjective image quality evaluation involve psychophysical experiments, in which a number of viewers are given a set of stimuli to be consumed in either predefined laboratory settings or typical environments with less controllable conditions. The subjects' ratings are recorded and processed to be presented as an indication of the quality of the visual content. Subjective quality assessment methodologies report to what extent the content is perceptually accurate and appealing to human audience, however there are a few drawbacks that introduce impracticalities. The psychophysical experiments are very costly, as they are time consuming in terms of design, preparation and execution. In order to ensure high generalization ability of subjective assessments and reduce content dependency of the results, the tests have to be conducted in a very large scale which makes it even more difficult to gather abundant number of scores [Akyazi and Ebrahimi (2018b)].

Objective assessment methods employ mathematical models to evaluate the degradation and the overall quality of visual content with respect a set of given input parameters. Unlike subjective quality assessment methodologies, the results of these methods are not dependent on opinion, therefore present the advantage of not changing between individuals. Moreover, the complexity of these models are deterministic. On the other hand, assuring the reliability of these objective quality assessment metrics in terms of their correlation with the general opinion of viewers is still a challenging research and application problem.

The general interest of Image Quality Assessment (IQA) methods is to come up with an objective Image Quality Metric (IQM) that is able to determine the quality of an input with high accuracy, that is, with high correlation to the would-be perceived quality, with much reduced cost compared to that of subjective quality assessment. Perceptually accurate objective image quality assessment methods benefit mostly from natural image statistics [Moorthy and Bovik (2011)] or models based on Human Visual System (HVS) [Sheikh and Bovik (2004)], where building such models in a robust fashion is a difficult task given the complexity of the HVS itself [Wang et al. (2002)]. With the recent developments in machine learning, more accurate models on image representation and classification can now be achieved [Simonyan and Zisserman (2014); Szegedy et al. (2015); He et al. (2016)]. Deep neural networks, especially deep convolutional networks have been shown to learn image representations and object classes with high generalization abilities. Such models are data-driven and rely on feature extraction and regression only, and can be used to predict the perceptual quality of the input as well [Bosse et al. (2018); Gao et al. (2017)]. Moreover, these models can be combined with visual saliency models to boost the accuracy, by using local weights to highlight more attractive regions in images.

2.2 Full reference image quality assessment

Subjective and objective image quality assessment methods can be further classified into three categories, depending on whether information from a reference, i.e. pristine quality image, is available fully or partially, or not at all. Full Reference (FR) image quality assessment methods have access to the reference image completely, while Reduced Reference (RR) image quality assessment methods use certain features extracted from the reference but not the image itself. In the case of No Reference (NR) image quality assessment methods, the reference image is not available at all and the quality evaluation is not based on the reference attributes. Since NR image quality assessment methods cannot make use of auxiliary information, in practice they are more challenging compared to RR and FR image quality assessment problems. On the other hand, all three methods have different use cases depending on the availability of the reference in any given application. FR quality estimation is more feasible for media applications such as broadcasting and video-on-demand, where encoder control is necessary. In a NR framework, the Rate-Distortion (R-D) trade-off could easily favor removing any type of noise or artifact. These, however, could have been intentionally placed in the contents for increasing viewers' emotional response, such as film grain and blur. Thus, in this work, we

focus on full reference image quality assessment methods and investigate the related work on this category.

2.2.1 Full reference subjective image quality assessment

Subjective image quality assessments need to be conducted following standardized guidelines. The general viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays is specified in ITU-R Recommendation BT.2022 [ITU-R BT.2022 (2012)]. Subjective experiments can be carried out either in controlled laboratory environments or home environments, between which the conditions of the environment, screen and viewing are specified differently.

The viewing distance of a subject can be either the Preferred Viewing Distance (PVD) or the Design Viewing Distance (DVD), where former is an empirically determined viewing distance based on viewers' preference and the latter is a device specific measure that reflects the distance at which two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye. The maximum observation angle relative to the normal should be constrained so that deviations in reproduced color on the screen are not visible to an observer. Any monitor processing such as image scaling, frame rate conversion or content enhancement should be conducted as to avoid interlacing artifacts and any use of deinterlacing filters should be reported. The minimum and maximum monitor resolutions should also be reported for each experiment. Other monitor specifications such as contrast and brightness should be calibrated according to ITU-R Recommendation BT.814-4 [ITU-R BT.814-4 (2018)] and ITU-R Recommendation BT.815-1 [ITU-R BT.815-1 (1994)].

Methodology for the subjective assessment of the quality of television pictures is specified in ITU-R Recommendation BT.500-13 [ITU-R BT.500-13 (2012)]. Besides the monitor and room specifications for controlled and home environments, ITU-R Rec BT.500-13 lays out the content selection process, observer profiles, assessment instructions, test procedure and presentation of the results.

Full reference subjective image quality assessment is conducted by gathering a number of subjects and stimuli, where each query image is presented along with its reference to each subject. At the beginning of a test session, the viewers are informed about the type of assessment, grading scale, contents and timing. A training session is also run to familiarize subjects with the test conditions and the grading scale examples.

The Double-Stimulus Impairment Scale (DSIS) assessment method involves two variants, where in Variant I and II, the reference and test images are presented only once, or twice consecutively, respectively. Variant II is preferred if the discrimination of very small impairments is desired. Otherwise, the less time consuming alternative Variant I is sufficient. In both settings, the reference and test stimuli are displayed side-by-side, with a 20 pixel width mid-gray level separation in between. When Single Stimulus (SS) assessment method is used,

each stimulus is presented individually. In this case, the reference image may also be included in the test to design a FR assessment. The categorical grading scales for DSIS and SS experiments, which measure the image impairment and image quality are depicted in Table 2.1. Alternative assessment methods such as Double Stimulus Continuous Quality Scale (DSCQS), Single Stimulus Continuous Quality Evaluation (SSCQE) and Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) are used for measuring the quality of stereoscopic image coding, fidelity between two impaired images and comparing different error resilience tools, respectively.

Table 2.1 – Rating scales for DSIS and SS subjective quality assessment methodologies.

	DSIS	**SS**
Scale	5 Imperceptible	5 Excellent
	4 Perceptible, but not annoying	4 Good
	3 Slightly annoying	3 Fair
	2 Annoying	2 Poor
	1 Very annoying	1 Bad

The viewing time for each stimulus does not have to be limited during experiments on images, however experience suggests that extending the viewing period per stimulus beyond 10s does not improve the subjects' ability to grade the quality. To remove bias, contents should be presented in a randomized fashion such that the same content is never presented consecutively. A single test session should last no longer than roughly half an hour to ensure subject concentration and prevent boredom or fatigue that may affect the results. If necessary, the full experiment can be divided into multiple sessions that allow sufficient breaks in between.

When analyzing the results, the overall mean scores, i.e. Mean Opinion Scores (MOSs) are computed along with 95% confidence intervals (CIs) per content. A screening needs to be performed to eliminate outliers among subjects, after which minimum 15 subjects are needed for an experiment to deliver conclusive results. Finally, statistical significance tests can be applied to compare each test condition and determine the relative quality ranking between stimuli.

2.2.2 Full reference objective image quality assessment

Computationally simplest measures to assess the quality of the distorted image in a FR framework are Mean Squared Error (MSE) and its derivatives, Signal-to-Noise Ratio (SNR) and Peak Signal-to-Noise Ratio (PSNR). These metrics are based on pixel color differences between a distorted image and its reference. Despite being widely used, these metrics do not involve any perceptual information, are known to be very content dependent and do not correlate well with subjective ratings [Lin and Kuo (2011)].

Objective color difference metrics measure the color distance between reference and distorted

images in a given color space. CIE76 [CIE (1986)] was the first metric of this kind introduced by the International Commission on Illumination (CIE), and was later refined due to perceptual non-uniformities in the underlying CIELAB color space. CMC [Clarke et al. (1984)], CIE94 [Witt (1995)] and CIEDE2000 [Luo et al. (2001)] superseded the previous formulae, all operating in CIELAB color space. CIEDE2000 is a more complex metric that includes not only weighting factors for lightness, chroma and hue but also takes into account the dependencies between chroma and hue.

Considering multiple factors such as loss of correlation, luminance distortion and contrast distortion, and modeling the quality based on pixel statistics instead of simplistic error averaging yielded to structural similarity metrics such as Universal Quality Index (UQI) [Wang et al. (2002)] and Structural Similarity Index (SSIM) [Wang et al. (2004)]. The structural information (cross-correlation) was defined as the attributes representing the objects in a scene, and evaluated independently from the average luminance (mean) and contrast (variance) of the image. SSIM index is applied locally rather than globally due to the variances of luminance and contrast, and was further extended to Multi-scale Structural Similarity Index (MS-SSIM) [Wang et al. (2003a)]. MS-SSIM takes into account that perceiving image details strongly depends on the sampling density of the image signal and the distance of the image from the viewer. These factors therefore affect the quality of the image and need to be considered when formulating a metric. By iteratively applying a low-pass filter to both reference and distorted images and downsampling the filtered images by a factor of 2, contrast and structure comparisons are carried out at multiple scales. The luminance component is considered only at a single scale and therefore has less emphasis on the overall score. Another variant of SSIM is Complex Wavelet Structural Similarity Index (CW-SSIM), which reduces the sensitivity of SSIM to translation, scaling, and rotation operations that do not cause structural changes in the image [Sampat et al. (2009)]. Such small geometric image distortions lead to consistent phase changes in the local wavelet coefficients, and a consistent phase shift of the coefficients does not change the structural content of the image. CW-SSIM combines magnitude and phase information from two-dimensional (2D) Discrete Wavelet Transform (DWT) coefficients to create a metric resilient to small distortions and provides smaller similarity index values only for large distortions. Quality Index based on Local Variance (QILV) [Aja-Fernandez et al. (2006)] is another metric that focuses on image structure by comparing the local variance distribution of the reference and distorted images.

Shifting the focus of metrics to the structural and contrast changes at multiple scales posed similarities with the HVS, which has high contrast sensitivity as a function of spatial frequency. Numerous metrics such as Image Fidelity Criterion (IFC) [Sheikh et al. (2005)], Visual Information Fidelity (VIF) [Sheikh and Bovik (2004)] and Feature Similarity Index (FSIM) [Zhang et al. (2011)] take into account the HVS responsiveness to changes in image contrast and low-level features. IFC involves Natural Scene Statistics (NSS), proposing that distortions disturb these statistics and make images look unnatural, therefore lower the perceived quality. IFC measures the fidelity between reference and distorted images based on the statistical information they share. VIF is an extension of this metric, which quantifies the information

that could ideally be extracted by the brain from the reference image and measures the quality of the distorted image by computing the loss of this information. Despite the inclusion of HVS responsiveness, these metrics use uniform weights when pooling local quality scores to yield a single image quality score. FSIM and its color extension $FSIM_C$ exploit the fact that visually informative features are concentrated at points of high Phase Congruency (PC), i.e. where the Fourier waves at different frequencies have congruent phases. Considering that PC is contrast invariant, the image Gradient Magnitude (GM) is computed as a complementary feature to encode contrast information. After computing the local similarity map using PC and GM, PC is utilized again as a weighting function to derive the final similarity score.

Extensions of PSNR and UQI considering HVS have been developed, PSNR-HVS and UQI-HVS [Egiazarian et al. (2006)], where the MSE measure in PSNR-HVS is modified according to the visual model weighted cosine transform [Nill (1985)], and UQI-HVS reports a weighted sum of the UQI measures at subbands of one-level discrete wavelet transform applied to reference and distorted images. PSNR-HVS has a further extension, PSNR-HVS-M [Ponomarenko et al. (2015)], which involves a masking model between image blocks based on Discrete Cosine Transform (DCT) coefficients and computation of each coefficient's masking degree using Contrast Sensitivity Function (CSF). Other HVS inspired metrics such as Divisive Normalization (DN) [Laparra et al. (2010)], Visual Signal-to-Noise Ratio (VSNR) and Weighted Signal-to-Noise Ratio (WSNR) [Mannos and Sakrison (1974); Mitsa and Varkur (1993)] try to empirically model the human perception of images from natural scenes. Most Apparent Distortion (MAD) metric proposes that the HVS uses multiple strategies to determine image quality, and models and combines these strategies. The claim is that for images containing near-threshold distortions, the image is most apparent and the HVS looks more for the distortions, whereas for images containing clearly visible distortions, the HVS attempts to seek the image's subject matter. Local luminance and contrast masking are used to estimate perceived distortion in high-quality images, whereas changes in the local statistics of spatial-frequency components are used to estimate perceived distortion in low-quality images [Larson and Chandler (2010)]. As not all image regions contribute to the quality score equally, using a weighted local information model to compute a global image score was considered in information content weighted metrics such as Information Content Weighted MSE (IW-MSE), Information Content Weighted PSNR (IW-PSNR) and Information Content Weighted SSIM (IW-SSIM) [Wang and Li (2011)], where saliency-based multiscale models have been used.

FR-IQA using machine learning

Most of the aforementioned metrics are designed as parametric models that can be tuned depending on content or application types. Learning-based models have the advantage of better generalization capabilities and parameter tuning provided that training involves large datasets with labeled instances.

Designed primarily for video quality assessment, Video Multimethod Assessment Fusion (VMAF) [Li et al. (2016)] fuses VIF and Detail Loss Metric (DLM) [Li et al. (2011)] using a Support

Vector Machine (SVM) regressor, where each metric is weighed in a fashion to preserve the strengths of the two metrics. Labels of the data are provided through subjective experiments on the NFLX Video Dataset. VMAF has proven to be efficient also in image quality assessment, showing substantially high correlation with subjective ratings [Li and Manohara (2019)]. Singular Value Decomposition (SVD) is used for feature extraction in [Narwaria and Lin (2011)], where the features are regressed into a perceptual quality score using Support Vector Regression (SVR).

A nonlinear combination of features extracted from several difference of Gaussian frequency bands are computed for Difference of Gaussians (DOG)-PSNR, DOG-SSIM and DOG-FSIM metrics [Pei and Chen (2015)], where the DOG features mimic the contrast sensitivity function of the HVS. DOG-based metrics have feature extraction and regression steps using random forest implementation. For both SVD-based and DOG-based metrics, four different models were trained using distinct publicly available databases with subjective scores, and cross-database evaluations were presented. The DOG-based models, especially DOG-SSIM, were efficient in terms of predicting scores that correlate well with subjective scores in terms of Spearman Rank Order Correlation Coefficient (SROCC), a correlation measure that checks the rank order between predictions and ground truth.

Deep Similarity (DeepSim) [Gao et al. (2017)] employs a convolutional architecture, i.e. VGGnet [Simonyan and Zisserman (2014)], computes local similarities between the features at each layer and explores various pooling methods to estimate a global quality score. Different preprocessing methods, layer types and depths were also investigated. Mid-level features were found to be the most effective within a 37-layer model, while the high-level features were most robust in terms of representing image quality. Preprocessing in terms of resizing the input image, using rectifying linear units in between convolutional layers, and max-pooling were also found to be efficient. It must be noted that DeepSim uses VGGnet feature maps pre-trained for recognition and shows that features learned for image recognition are also meaningful in the context of perceived quality.

While the aforementioned methods involve feature learning and regression, Deep Image Quality Measure (DIQaM-FR) is an end-to-end trained method for FR image quality assessment. This work again uses the VGGnet architecture [Simonyan and Zisserman (2014)] with 10 convolutional and 5 pooling layers for feature extraction combined with two fully connected layers for regression in a Siamese network. The reference and distorted images are separated into random patches to allow for artificial data augmentation. The features from reference and distorted image patches are fused before regression to increase the accuracy of the model. Another extension of DIQaM-FR is the weighted version of the model, referred to as the Weighted Average Deep Image Quality Measure (WaDIQaM-FR), where two fully connected layers running in parallel to the quality regression layers are added. These layers estimate the weights of local patches with respect to the overall quality score, thereby including a saliency weighted distortion pooling. A similar weight estimation network is used for the metric paPSNR [Bosse et al. (2019)] to estimate localized distortion sensitivity for MSE and

thereafter, PSNR.

2.3 Full reference image quality assessment databases

Just as important as the IQA solutions themselves is the abundance of rich datasets for training the algorithms, when necessary, and benchmarking the results. Objective IQA databases are constructed using a variety of reference images and processing them with different methodologies that introduce characteristic artifacts at various levels. When constructing a database for FR-IQA, ideally the distorted images need to be subjectively compared to their respective references in a controlled or home environment. Crowdsourcing is one of the most practical methods for recruiting many subjects, with services such as Amazon Mechanical Turk (MTurk)[1] or microWorkers[2] that provide scalable workforce for human intelligence tasks. Ratings gathered from subjects are then processed to yield Mean Opinion Score (MOS) values and CIs that can be used for IQA tasks. This section provides a list of existing databases that are widely used for FR-IQA:

- The Tampere Image Database 2008 (TID2008) [Ponomarenko et al. (2009)] contains 25 reference images cropped to 512 × 384 resolution from the Kodak database and 1700 distorted images, where for each reference image there are 17 types and 4 levels of distortion. Reference images were obtained by cropping from Kodak Lossless True Color Image Suite[3]. The distortions include various noise types such as additive Gaussian noise, spatially correlated noise, high frequency noise, impulse noise and quantization noise. JPEG and JPEG 2000 compression artifacts as well as transmission artifacts are presented. MOS measures in the scale $[0,9]$ were obtained from subjective assessments with the participation of 800 subjects.

- The Tampere Image Database 2013 (TID2013) [Ponomarenko et al. (2015)] is an extension to TID2008, which contains the same reference images, where for each reference image there are 24 types and 5 levels of distortion amounting up to a total of 3000 distorted images. A wider spectrum of distortion types have been included in this database, including Gaussian blur, lossy compression of noisy images and sparse sampling and reconstruction. MOS values of the database lie in the range $[0,9]$ with 0 being the lowest quality score and 9 the highest. 971 subjects had been involved in the experiments from five different countries.

- The Laboratory for Image & Video Engineering (LIVE) Database Release 2 [Sheikh (2005)] contains 29 reference images of typically 768 × 512 resolution, and 779 distorted images with 5 different distortion levels for various distortion types such as JPEG, JPEG 2000, Gaussian blur, white noise and bit errors in JPEG2000 bitstream. Difference Mean

[1] https://www.mturk.com
[2] https://www.microworkers.com
[3] http://r0k.us/graphics/kodak/

Opinion Score (DMOS) values are reported in the range [0,100], where 0 indicates the best quality and 100 indicates the worst quality. 20-29 subjects participated in the SS experiments using Absolute Category Rating (ACR) voting scale.

- LIVE Multiply Distorted Image Quality Database (LIVE MD) [Jayaraman et al. (2012)] is comprised of 15 references contaminated by multiple distortions at once, namely, Gaussian blur followed by JPEG compression to simulate image storage scenarios, and Gaussian blur followed by Gaussian noise to simulate camera defocus and sensor noise. Subjective scores for 240 images including references were gathered using ACR and reported as DMOS measures. Natural, people and urban categories are preferred as references, with resolution 1280×720.

- The Laboratory of Computational and Subjective Image Quality (CSIQ) image database [Larson and Chandler (2010)] consists of 30 reference images obtained from public domain resources, each distorted at four to five different levels of distortion. Reference images are chosen to span five categories, namely animals, landscapes, people, plants and urban. Distortion types included in the CSIQ image database are JPEG compression, JPEG 2000 compression, Gaussian blur, Gaussian white noise, Gaussian pink noise and contrast change. As in LIVE database, the ratings are reported in the form of DMOS in the range [0,1]. 35 subjects participated in controlled experiments to gather scores, with each subject viewing a subset of the images. All reference and distorted images are of resolution 512×512.

- The Media Communications Lab - JND-based Coded Images (MCL-JCI) dataset [Jin et al. (2016)] consists of 50 reference images with higher resolution, i.e. 1920×1080. Each reference image was JPEG compressed using quality factors ranging from 1 to 100, to yield 5000 distorted images in total. Ten semantic categories were involved in the reference selection spanning indoor and outdoor natural images. More than 150 volunteers participated in the subjective tests that were run in a controlled environment. The ratings are not reported in the form of MOS values, but rather in Just-Noticeable Difference (JND) which is a statistical measure that accounts for the maximum difference unnoticeable to humans. In such testing, subjects were shown the distorted images sequentially, with decreasing quality from factor 100 to 1. Noticeable differences that were reported with respect to reference were recorded as JND instances.

- Waterloo Exploration Database [Ma et al. (2016)] is a large-scale database that is composed of 4744 pristine references, which is a much higher number compared to the previously introduced IQA databases. The references are of Standard Definition (SD) resolution and have a corresponding set of distorted images of size 94880. Four different types of distortion, namely JPEG compression, JPEG2000 compression, white Gaussian noise contamination and Gaussian blur were involved at 5 levels each. Instead of collecting subjective ratings, three alternative test criteria are proposed to evaluate the performance of IQA models, namely the pristine/distorted image discriminability

test (D-test), the listwise ranking consistency test (L-test), and the pairwise preference consistency test (P-test).

2.4 Lossy image compression

All digital visual data received by consumers is compressed due to storage and transmission constraints. Different levels and types of degradation are introduced in images depending on the characteristics of the applied compression algorithm. State-of-the-art lossy compression and transmission schemes introduce artifacts in images such as blur, blocking, ringing, contrast changes and color bleeding. These degradations are reflected as visual quality impairments to end users. The goal of broadcasters and service providers is to ensure high visual quality while maintaining low storage and transmission costs. In end-to-end systems that are delivering content to the viewers, the degree of annoyance that could be introduced at the output has to be anticipated using an efficient metric, and minimized accordingly.

As the quality of compressed images with respect to references can be measured using metrics that rely on various factors such as human visual system, natural scene statistics and image fidelity, the same factors need to be considered when designing a compression algorithm. The compressed data will have high visual quality, also referred to as visually lossless or near-lossless, provided that key perceptual elements are preserved. Defining and conserving these components are the main challenges in the field of image compression.

For decades, image and video processing communities have been proposing different solutions to improve compression efficiency. A typical image compression architecture, i.e. image encoder, first maps the data to a lower dimensional space, then quantizes and entropy codes this representation to construct the encoded bitstream. The decoder then reverses these compression steps and decompresses the data to reconstruct the original image with high quality. All three major components of image compression, namely, feature extraction, quantization and entropy coding, are active research problems.

While traditional image compression algorithms use hand-crafted features and fixed transforms to represent the image in reduced dimensions, learning-based image compression models extract features from training instances and seek to minimize a cost function that typically measures the distance between reference and decompressed images. The performance of learning-based models present more flexible and generalizable compression solutions, and have started to reach the performance of their state-of-the-art transform-based counterparts.

2.4.1 Transform-based image codecs

One of the most widely used coding standards, JPEG [Wallace (1992)], was developed almost three decades ago by the standardization effort Joint Photographic Experts Group (JPEG), a sub-group of ISO/IEC Joint Technical Committee 1, Subcommittee 29, Working Group

1 (ISO/IEC JTC1/SC29/WG1). JPEG was the first international digital image compression standard for still images in grayscale and color. The goal while developing the standard was to have state-of-the-art rate and accompanying image fidelity over a wide range of R-D trade-offs, create a parametrizable encoder to allow the user to determine a particular trade-off, have tractable computational complexity and be applicable to any kind of digital source image. Additionally, the standard essentially had four modes of operation: sequential encoding, progressive encoding, lossless encoding and hierarchical encoding, to be preferred depending on transmission constraints. The encoding strategy of JPEG relies on DCT, where the input image is first divided into 8×8 blocks and the DCT coefficients are quantized such that the high-frequency information that is potentially discarded by HVS is removed. A backward-compatible extension was released with JPEG XT, with support for integer bit depths between 9 and 16, High Dynamic Range (HDR) imaging and floating-point coding, lossless and near-lossless coding and alpha channel coding [Richter et al. (2016)]. JPEG XT also includes reference software implementation and conformance testing specification.

Prior to JPEG XT, two other formats were also standardized by the committee: JPEG 2000 [Taubman and Marcellin (2012)] and JPEG XR [Dufaux et al. (2009)]. The central concept behind JPEG 2000 is scalability, where the compressed bitstream consists of an embedded collection of smaller streams, representing the image at any of a number of different resolutions, each at any of a number of different qualities. The scalability of JPEG 2000 comes from its use of DWT and Embedded Block Coding with Optimized Truncation (EBCOT), and delivers superior compression ratio at the expense of higher computational complexity. As DCT blocks are discarded with the use of wavelets, the images no longer suffer from blocking artifacts of JPEG, but typical ringing artifacts around the edges of objects still persist. Despite the aforementioned advantages and other features such as lossless and lossy compression modes, HDR imaging and alpha channel coding, JPEG 2000 has not been adopted in consumer digital market to the extent of replacing the legacy JPEG.

JPEG XR standard, with XR standing for extended range, primarily targeted the representation of continuous-tone still images such as photographic images, while reducing the computational complexity compared to that of JPEG 2000. Target applications involved robust and high-fidelity image acquisition technologies and computationally constrained environments such as mobile and embedded applications. JPEG XR provides native support for both RGB and CMYK color spaces as well as YUV, monochrome and arbitrary n-component color formats. A lifting-based reversible hierarchical Lapped Biorthogonal Transform (LBT) is used, which involves a DCT-like transform at the core step followed by an optional overlap filtering to exploit block correlations. Unlike JPEG 2000, JPEG XR is backward-compatible with legacy JPEG standard. Microsoft has the patents on JPEG XR technology and has released an open source JPEG XR library[4] under Berkeley Software Distribution (BSD) license .

The image compression format that followed was WebP[5], announced in 2010 as a new open

[4] https://archive.codeplex.com/?p=jxrlib
[5] https://developers.google.com/speed/webp

standard under BSD license for lossy compressed images on the web, with smaller image sizes and comparable quality to that of JPEG. Lossy WebP compression uses predictive coding to encode an image, which is the same method used to encode key frames in its video compression counterpart VP8. The values in neighboring blocks of pixels are used to predict the values in a block and only the differences are encoded. In addition to prediction coding, WebP employs block adaptive quantization and boolean arithmetic encoding, which constitute its main compression gains over JPEG.

High Efficiency Video Coding (HEVC/H.265) standard [Sullivan et al. (2012)] is a joint video project of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Motion Picture Experts Group (MPEG), released in 2013 with the aim of superseding the previous video coding standard Advanced Video Coding (AVC/H.264) as beyond-High Definition (HD) resolutions were becoming more widespread. HEVC/H.265 targeted increased video resolution and parallel processing architectures. Main strengths of this standard are Coding Tree Units (CTU) which can use larger block structures up to dimensions 64×64 (which was limited to 16×16 in AVC/H.264), quadtree syntax of CTUs providing improved variable-block-size segmentation, 33 directional modes for intra-prediction within the same frame, improved motion vector prediction and region merging as well as motion compensation filtering. The improved intra-prediction is supported by all block sizes supported by HEVC/H.265, and introduces a large coding gain not only for videos but also for image coding. The Better Portable Graphics (BPG) and High Efficiency Image File Format (HEIF) are types of wrappers used around HEVC/H.265-Intra, with HEIF having been standardized by MPEG. The successor to HEVC/H.265 is currently being developed under Joint Video Experts Team (JVET), referred to as Versatile Video Coding (VVC) [Ohm and Sullivan (2018)], with improvements on intra prediction such as 65 directional modes, rectangular block prediction, transform units larger than 32×32, new prediction modes allowing directional interpolation, and chroma prediction.

The performance of HEVC/H.265-Intra has been shown to be superior to JPEG, JPEG 2000, JPEG XR and WebP [Nguyen and Marpe (2012); Lainema et al. (2016)], with average bitrate savings in the range of 17-44%. One major drawback of HEVC/H.265-Intra is royalties. The most recent image coding standard being developed by the Alliance for Open Media (AOM), which supports a royalty-free and open source ecosystem, is called AV1 Image File Format (AVIF)[6]. AVIF is based on the video codec AV1 that is based on Google's VP9, with more coding options that enable better adaptations to different types of inputs. The 64×64 macroblocks of HEVC/H.265 are extended to 128×128 superblocks that can be partitioned to units as small as 4×4. Along with new optimized quantization matrices, a total of 56 intra-prediction angles are used. According to leading developers, AV1 video codec is expected to deliver a 30% improvement over VP9 and HEVC/H.265. The claim has been verified for AV1's compression efficiency against VP9, but a much less average difference was observed between the performances of AV1 and HEVC/H.265 in the study conducted using HD videos [Akyazi and Ebrahimi (2018a)]. Formal studies to evaluate the performance of AVIF are pending, yet the high performance of

[6] https://aomediacodec.github.io/av1-avif/

AV1 suggests that AVIF may bring advantages over previous image coding standards. Despite having wide software support, a major drawback of the reference software of AV1 and hence AVIF is the encoding speed. Faster implementations are currently being developed by several parties simultaneously.

Parallel to AOM's efforts on creating the next generation state of the art video compression format, the JPEG committee is standardizing the next generation royalty free image coding standard, JPEG XL. The goal of JPEG XL is to develop a new image coding standard that provides state-of-the-art image compression performance, and that addresses shortcomings in current standards. The goal of JPEG XL image codec is to achieve significant compression efficiency improvement over coding standards in common use at equivalent subjective quality, e.g. minimum 60% gain over JPEG while remaining backward-compatible with existing legacy JPEG decoders, offer features for web applications, such as support for alpha channel coding and animated image sequences, and offer support of high-quality image compression and professional photography, including higher resolution, higher bit depth, higher dynamic range, very high quality and wider color gamut coding [Akyazi and Ebrahimi (2019a)]. JPEG XL coding tools include a rich combination of elements that consider HVS and coding efficiency such as variable-size DCT, nonlinear Haar transforms, multiresolution encoding, adaptive quantization, adaptive loop filters and context modeling.

With a provisional release date in October 2019, JPEG XL targets a large variety of use cases including image-rich UIs and web pages on bandwidth-constrained connections such as social media applications, media distribution applications, cloud storage applications, media web sites, animated image applications, mobile applications and games, and high quality imaging applications such as rapid photo viewing, HDR/WCG user interfaces, augmented/virtual reality, image bursts, high-end photography, image mosaics, depth images and printing.

2.4.2 Learning-based image codecs

Machine learning approaches have demonstrated advanced solutions to many image processing problems such as object classification and image enhancement, and are continually improving at image related tasks. Recent learning-based image compression models have reached and even surpassed the performance of transform-based state-of-the-art image codecs. Learning-based methods do not employ hand-crafted features but rather extract a latent representation of the input image through neural networks. Recent learning-based image compression methods involve different neural network architectures, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Generative Adversarial Networks (GAN). Convolutional neural networks extract local features of the image at each layer while recurrent neural networks allow processing the images in a sequential manner. Generative models, on the other hand, create new data from training images by learning the statistics. All these models can be efficient for the task of image compression by training a network end-to-end, provided that the loss function measures a meaningful distortion between

the original and the decoded images [Akyazi and Ebrahimi (2019d)].

Learning-based methods have been involved in image compression in components such as learning more efficient frequency transforms, predictive coding, segmentation and quantization [Jiang (1999)]. More recent methods tackle the entire compression problem in end-to-end models using autoencoder architectures. These autoencoders have three main parts: (i) the encoder where the input is mapped to the latent space, (ii) the bottleneck where the latent representation is coded and (iii) the decoder, where the code is transformed back to yield an output close to the input. A typical bottleneck involves quantization of the code followed by the construction of an efficient representation through methods such as entropy or arithmetic coding. In convolutional architectures, the decoder inverts the encoding process usually by replacing the convolution filters with reversed operators, i.e. deconvolution. Pooling and activation functions are also inverted during decoding.

At very low bitrates, i.e. bitrates below 0.15bpp, the generative adversarial model presented in [Agustsson et al. (2018)] achieves good subjective image quality by fully synthesizing selected regions of images. At similar bitrates, [Cheng et al. (2019a)] has shown that residual CNN architectures and deeper networks increase the performance both subjectively and objectively. Constraining the reconstructed images to follow the input distribution to optimize rate-distortion trade-off extended the generative model in [Agustsson et al. (2018)] to span more R-D points, where at zero rate, the network learns a generative model of the data, and perfect reconstruction is achieved at high enough bitrates [Tschannen et al. (2018)]. A layered conceptual image compression scheme combining Variational Autoencoders (VAE)s and GANs is presented in [Chang et al. (2019)] where the textures are encoded into latent vectors by the VAE. The edge map of the image is simultaneously encoded via screen content coding techniques. The compressed image is synthesized using a GAN using the latent vector such that the edge map is preserved at the decoder side. The proposed method was tested at very low bitrates, and on images containing single objects so that edge map extraction is efficient. A more general framework for variable-rate image compression based on convolutional and deconvolutional Long Short-Term Memory (LSTM) recurrent networks is presented in [Toderici et al. (2015)]. The proposed network achieves progressive encoding as it is able to deliver more accurate representations by sending more bits. Several architectures have been tested: feed forward fully-connected residual encoder, LSTM-based residual encoder, feed-forward convolutional/deconvolutional residual encoder and convolutional/deconvolutional LSTM compression. The fully connected residual encoder employs a stack of fully connected layers at the encoder and decoder. The residual at stage t is fed into the next stage $t+1$ for progressive encoding using more bits. The stages are also implemented using LSTM blocks, or convolutional/deconvolutional filters. A combination of LSTM blocks and convolutional/deconvolutional filters is also tested. The LSTM compressor and the combined compressor were shown to outperform JPEG on thumbnail size images, in terms of SSIM.

The resolution of images has been increased in [Theis et al. (2017)] and [Toderici et al. (2017)], as well as an increase in the speed of the algorithm. The former work employs Convolutional

Autoencoders (CAE) where deconvolutions are replaced by sub-pixel convolutions, and residual connections and leaky rectifications between layers of encoder and decoder are added. The latter compares different RNN types (LSTM and associative LSTM) and introduces a hybrid Gated Recurrent Unit (GRU) [Chung et al. (2014)] and ResNet [He et al. (2016)] model. The extensive comparison between architectures show that it is not easy to pick a winning algorithm, since results vary with respect to different quality metrics and at various bitrates. The authors also show that training the models on "hard to compress" data, i.e. selection of training instances that compress the least using the lossless compression standard Portable Network Graphics (PNG), yields better quality in terms of MS-SSIM [Wang et al. (2003b)] and PSNR-HVS metrics. In progressive encoders, predicting the original image from residuals is shown to facilitate learning also in [Baig et al. (2017)]. The performance can be improved further by learning to inpaint from neighboring pixels before compression, thereby reducing the storage rate. A trade-off parameter for training the models at different bitrates is introduced in [Ballé et al. (2016)] and widely used in other works [Cheng et al. (2018); Ballé et al. (2018); Cheng et al. (2019a)]. In [Cheng et al. (2019b,c)], a further constraint is introduced as a second regularizer that compacts the spatial energy for optimal bit-allocation. The Generalized Divisive Normalization (GDN) [Ballé et al. (2015); Ballé (2018)] nonlinearity, which is a parametric nonlinear transformation that is able to Gaussianize data from natural images, is also shown to increase the performance of end-to-end image compression models.

Multiscale models are of interest in recent works [Wang et al. (2019)] where a Laplacian pyramid is applied to the images to construct a multiscale representation. The multiple scale inputs are processed through separate CNN layers to learn a hierarchical compressive representation of high-resolution images. Such a model is also able to restore an image progressively. A pyramidal decomposition is used in WaveOne encoder [Rippel and Bourdev (2017)] to extract image features, which was inspired by wavelet decomposition for multiresolution analysis, followed by an interscale alignment that leverages information shared across different scales. An adversarial training using GANs is employed to ensure realistic reconstructions at the decoder. Results on Kodak and RAISE-1k [Dang-Nguyen et al. (2015)] databases show that the proposed model was able to outperform JPEG, JPEG 2000, WebP and BPG in terms of MS-SSIM over a wide selection of bitrates while operating at real-time.

WaveOne includes an adaptive codelength regularization that penalizes the entropy of the quantized latent representation. Further improvements are implemented in [Ballé et al. (2018)] where entropy coding involves an autoencoder that learns a hyperprior, capturing the spatial dependencies within the latent representation. Models using MSE and MS-SSIM loss as a distortion measure in the cost function are evaluated, where experimental results on the Kodak dataset show that the model outperforms BPG in terms of MS-SSIM. The hyperprior model is extended to exploit two types of contexts to estimate the distribution of the latent representation, thereby increasing compression efficiency in [Lee et al. (2018)]. Again, the hyperprior model is combined with an autoregressive context model that predicts the latents from their causal context in [Minnen et al. (2018a)]. Other works employ a variety of entropy models including factorized entropy models [Cheng et al. (2018); Akyazi and Ebrahimi (2019c); Cheng

et al. (2019a)] and single-iteration and progressive LSTM-based entropy models [Toderici et al. (2017)] during training, and range coder [Theis et al. (2017)] and context-based adaptive binary arithmetic framework [Marpe et al. (2003)] during test. Instead of entropy estimation, the rate optimization problem is rephrased as an Adaptive Direction Method of Multipliers (ADMM) solvable problem in [Zhao and Liao (2019)] that promotes sparsity in the latent representation. The entropy of the latent representation is modeled by a context model in [Mentzer et al. (2018)]. A conditional probability model of the latent distribution is learned using a 3D-CNN. During training, the autoencoder estimates the entropy of its representation the context model, which is in turn updated to learn the dependencies between the symbols in the latent representation. Spatially local image-dependent multinomial dictionaries are proposed to model the entropy to optimize R-D performance in [Minnen et al. (2018b)].

At the bottleneck, quantization of the latent representation can also be modeled differently. A widely preferred approach is performing uniform scalar quantization, i.e. rounding to nearest integer, which effectively implements a parametric form of vector quantization on the original image space [Ballé et al. (2016)]. This operation is not differentiable and therefore a smooth approximation is implemented by adding uniform noise to the latent representation during training. In [Theis et al. (2017)], a stochastic rounding operation is defined where the derivative is replaced with the derivative of the expectation in the backward pass. Similarly, a stochastic form of binarization is used in [Toderici et al. (2015)]. Another soft relaxation is presented in [Agustsson et al. (2017)] where vector quantization is explored in the context of learned image compression and improvements over scalar quantization are demonstrated. Motivated by the variance of local information content throughout an image, a content-aware bit allocation method is created in [Li et al. (2018)], which operates under a content-weighted importance map. The importance map serves as a continuous alternative of discrete entropy estimation to control the rate and a binarizer is preferred for quantization. An iterative approach is presented in [Cai and Zhang (2018)], where quantizer and encoder-decoder networks are optimized alternatively. The encoder-decoder is first fixed without quantization and a non-uniform quantizer is then optimized using the latent representation features. Following this step, the encoder-decoder network is updated again using the optimized quantizer.

Recently, Workshop and Challenge on Learned Image Compression (CLIC) 2019 organized as part of Conference on Computer Vision and Pattern Recognition (CVPR) demonstrated that the performance of learning-based image compression solutions are able to reach 40dB PSNR and 0.9931 MS-SSIM on 330 HD images provided as the CLIC test set, at an average of 0.86bpp [Zhou et al. (2019)]. A variational end-to-end autoencoder architecture was used to build the winning TUcodec, similar to the model in [Zhou et al. (2018)] which involves a pyramidal feature fusion strategy that learns optimal, nonlinear features for each scale among convolutional layers. An attention mechanism referred to as Residual Non-Local Attention Block (RNAB) was added to capture the global dependencies between features. TUcodec attained 31.22dB PSNR and 0.9739 MS-SSIM when the distortion measure was optimized using PSNR and SSIM metrics, respectively, on an average of 0.15bpp on the same test set. The latter optimization also provided the best MOS in the challenge evaluations for the low-rate

track, followed by the codec ETRIDGUlite [Cho et al. (2019)], which improves the compression efficiency of VVC Intra by employing a Grouped Residual Dense Network (GRDN) [Kim et al. (2019)] as a post-processing step. Another example of such hybrid models is proposed in [Ma et al. (2019)], where a CNN-based re-compression of JPEG 2000 bitstreams and decoder-side post-processing were implemented. Semantically-salient regions are highlighted using a CNN in [Prakash et al. (2017)] that allows encoding the salient regions at higher quality with respect to the background. Such preprocessing was used to enhance the performance of JPEG encoder. A deep semantic segmentation-based layered image compression (DSSLIC) framework is presented in [Akbari et al. (2019)] where the semantic segmentation map of the input is extracted using a pre-trained network. Next, the input image and the segmentation map are used together to yield a coarse reconstruction of the input. The segmentation map and the coarse reconstruction are encoded as base and first enhancement layers, respectively. The coarse reconstruction is then processed using GANs. The residual between the input and the coarse reconstruction is also encoded with BPG as a second enhancement layer. DSSLIC is shown to outperform JPEG, JPEG 2000, HEVC/H.265-Intra and WebP on Kodak dataset at various rates.

2.5 Summary and perspectives

Numerous inspiring works have contributed to advances in both objective image quality assessment and image compression problems. The two domains are intertwined in the sense that the quality of compression algorithms is measured and optimized with the objective metrics, which try to match subjective opinion. It must be noted that, with the rate of progression of machine learning, the performance of objective image quality assessment is not far from reaching and even surpassing the reliability of subjective ratings. To this end, a number of issues still remain to ameliorate the work in both image compression and quality evaluation, which are addressed in this thesis as follows:

1. The HVS and NSS are considered within numerous objective image quality assessment metrics, with an emphasis on multiple scales and differencing such as in MS-SSIM and DOG-SSIM, and local information as in FSIM. Although the components of HVS responses are expected to be inferred in learning-based IQA models, explicitly introducing factors to evoke HVS response would enhance the robustness of the evaluation. In Chapter 3, a new objective image quality assessment method is presented which involves a preprocessing step using 2D DWT and performs better than the state-of-the-art metrics such as PSNR, SSIM, MS-SSIM, FSIM and WaDIQaM-FR when trained and tested on the full TID2013 database.

2. As the 2D DWT preprocessing was shown to improve the performance of learning-based IQA, a similar approach is proposed for the compression of images in Chapter 6. The proposed WCAE model shows that 3-scale 2D DWT coefficients given as the input to the autoencoder perform better than a similar model that excludes such preprocessing.

Further analysis is presented in Chapter 7 on different architectures and rate points using models ResWCAE and ResMixWCAE. All models perform superior to the legacy JPEG encoder, and provide insights on the effects of wavelet scales on the construction of the latent representation and the visual attributes of the decoded image.

3. The results of Chapter 7 trigger curiosity related to the effect of kernel sizes in autoencoder architectures. A novel ablation study on the use of mixed convolutional branches is presented in Chapter 8, which exposes the contributions of layers with varying filter sizes to the overall compression performance.

4. One key issue in objective IQA is the generalization ability of the models, which can turn to be an advantage for the learning-based approaches provided that the training data is large and versatile enough. With subjective IQA being costly as it is, the number databases with reference and distorted image pairs and respective subjective ratings are limited. The existing databases presented in Section 2.3 such as TID2013, CSIQ, LIVE and Waterloo Exploration Database are (i) of SD resolution and (ii) include numerous distortion types and levels that do not necessarily correspond to one another in a way to allow a meaningful combination of databases for comprehensive training. To match the standards of present-day media, HD, Ultra High Definition (UHD) or even higher resolutions need to be present in an IQA database. Moreover, an IQM trained using an IQA dataset comprised of images bearing a broad range of compression artifacts would be superior in evaluating the effect of degradation brought by compression algorithms. In Chapter 4, the construction of a new IQA dataset is presented, consisting of HD and UHD reference images, their distorted counterparts after being compressed and decompressed using state of the art compression algorithms and the MOS of each distorted image. Subjective ratings were gathered by conducting subjective quality experiments in controlled environments. Using this novel dataset, it is possible to train an IQM that is able to rate the impact of compression degradations better than the state of the art IQA algorithms.

5. State-of-the-art learning-based image compression networks are trained using conventional metrics such as PSNR and MS-SSIM due to their simple and differentiable nature. Such metrics, however, poorly correlate with subjective ratings. The IQM described in Chapter 5 is (i) differentiable, (ii) superior to the state-of-the-art objective metrics in terms of correlation with subjective scores and (iii) does not rely on hand crafted features to assess quality. A novel learning-based compression solution is implemented in Chapter 9, by coupling this image quality metric with an autoencoder so that the metric is used as the distortion measure in the autoencoder loss function. The influence of the learning-based metric on the autoencoder performance is analyzed in detail in Chapter 9.

Part I

Learning-based image quality assessment

3 A new objective metric to predict image quality using convolutional neural networks

Objective image quality assessment is evolving towards incorporating models of higher complexity that consider elements of HVS, as well as learning-based models that infer HVS responsiveness through training data. The end-to-end image quality assessment models DIQaM-FR and WaDIQaM-FR proposed in [Bosse et al. (2018)] show superior performance with respect to the state-of-the-art methods when trained and tested on the full TID2013 database [Ponomarenko et al. (2015)]. The models also perform well at cross-database evaluations involving subsets of CSIQ [Larson and Chandler (2010)] and LIVE [Sheikh (2005)] databases, however, their generalization ability is still limited when it comes to involving full CSIQ or LIVE databases for evaluation.

The neural network architectures of DIQaM-FR and WaDIQaM-FR are based on the VGGnet structure depicted in [Simonyan and Zisserman (2014)], which comprises of a number of convolutional layers with 3×3 kernels. The number of outputs are increased through consecutive layers, while during a doubling of the output channels, the output dimensions are reduced to half using MaxPool layers. The very small 3×3 receptive fields are convolved with every pixel of the input through a stride of 1 and each hidden layer is equipped with a non-linear rectification unit, which helps the decision function at the end become more discriminative.

The presented model shows that increasing the number of layers and keeping the receptive field at each layer as small as 3×3 increases the image classification performance.

Image classification and recognition tasks are connected to quality evaluation, as both are related to understanding and analyzing the local characteristics of the image to yield a global categorization. A similar architecture is therefore adopted while constructing DIQaM-FR and WaDIQaM-FR. 10 convolutional layers are used with 5 pooling layers in between, followed by 2 fully connected layers at the end for regression to obtain a final score for the distorted image with respect to the reference. A Siamese network is preferred [Bromley et al. (1994); Chopra et al. (2005)], which comprises of an architecture to learn similarity relations between two inputs, to extract features from reference and distorted images. Corresponding patches from reference and distorted images are processed through two identical architectures that learn different weights. Features extracted from reference and distorted images are then fused and passed through two consecutive fully connected layers that output the final score. The network is trained on randomly sampled patches of dimensions 32×32 from each image in a chosen database, which contributes to the diversity of the data. Different from DIQaM-FR, WaDIQaM-FR involves assigning non-uniform weights to the selected patches in determining their contribution to the overall image score, through the addition of two more fully connected layers running parallel to quality regression layers. The performance of WaDIQaM-FR is superior to that of DIQaM-FR in terms of Pearson Linear Correlation Coefficient (PLCC) and SROCC when trained and tested on TID2013 and LIVE databases, as well as on cross-database evaluations. The comprehensive study presented in [Bosse et al. (2018)] investigates the effect of other numerous hyperparameters on the overall network performance, such as:

- Number of patches (N_p): After training on both TID2013 and LIVE databases using a logarithmic range of N_p values from 2^0 to 2^{10}, it was shown that $N_p = 32$ is a saturation point for the performance of WaDIQaM-FR.
- Feature fusion: It was shown that when using a combination of features obtained from reference and distorted images, i.e. f_r and f_d, concatenating f_r, f_d as well as the difference feature $f_d - f_r$ yields the best performance for both DIQaM-FR and WaDIQaM-FR. The other alternatives that were investigated were using only the difference $f_d - f_r$, or concatenating only f_r and f_d.
- Network depth: Comparisons between the original network and a shallower one with several layers removed show that deeper and more complex models lead to a more accurate prediction, however, the gain in terms of PLCC and SROCC is in the order of 10^{-3} to 10^{-2} for a five-fold increase in parameter size. This suggests that when computational complexity is an issue, the 10-layer architecture can be reduced in size at a relatively small expense in accuracy, both for tests in the same dataset as in training and cross-database evaluation.

Although the performance of DIQaM-FR and WaDIQaM-FR are superior to the state-of-the-art

methods when trained and tested on the full TID2013 database, the generalization ability of the models are still limited in cross-database evaluations. This suggests that extracting only spatial features followed by regression could benefit from auxiliary information such as frequency characteristics of the images in multiple scales. While the preferred VGGnet structure is largely inspired by HVS, explicitly introducing multiscale image features could enhance the model performance. Additionally, it has been shown that deep residual network architectures show better generalization abilities compared to plain VGGnet architectures, and ease the optimization by providing faster convergence at earlier stage even for very deep networks [He et al. (2016)].

In this chapter, a novel convolutional neural network based image quality assessment method that is able to objectively predict the quality of distorted images using the reference image and subjective ratings is presented. The proposed network uses the spatial information as well as the frequency content of the reference and distorted images. The frequency content is analyzed by applying a three-scale wavelet decomposition on the grayscale reference-distorted image patch pairs. The overall architecture is inspired by the work in [Bosse et al. (2018)], where the convolutional neural network branches are composed of 3×3 filters and employ a slightly modified residual network structure [He et al. (2016); Nah et al. (2017)]. In order to extract features from both reference and distorted images, a Siamese network structure is adopted and the features of the two images are concatenated as well as the difference of these features. In order to introduce as many inputs as possible to the model, at each epoch random patches of dimensions 32×32 are selected from reference and distorted image pairs. The final image quality is computed as a weighted linear combination of the patch qualities, where the weights are also determined using two fully connected regression layers that estimate the influence of local patches to the global quality. The proposed network was trained and tested using images from the TID2013 database, with cross database evaluations conducted on LIVE and CSIQ image quality databases. The performance of the proposed method was also compared with those of the state-of-the-art methods in termsPLCC and SROCC and shown to achieve competitive results.

3.1 WIQM framework

The proposed framework first extracts features from both the reference and distorted images. These features are then concatenated into a single feature vector that is passed onto the fully connected layers for regression and an objective quality score is assigned at the output. The architectures that have been built for feature extraction, feature concatenation and regression have been inspired by the implementations in [Bosse et al. (2018)]. A Siamese network is used to extract features from both the reference and distorted images, where the design of the convolutional layers and the preferred building blocks have been changed.

3.1.1 Feature extraction

The main novelty of WIQM lies in the use of color channels of images as well as the wavelet decomposition of the grayscale images up to three scales as the input. 2D DWT is known to be effective in image processing tasks such as denoising, interpolation, sharpening and compression by providing information about both the spatial and frequency content of the image in different scales. High frequency components in images, such as edges and textures, can be distinguished from other components such as noise. The distortions in natural image databases contain artifacts such as blur, noise, illumination effects, blocking and more. These artifacts may affect distinct frequency areas of the images in a different fashion. It is therefore important to analyze the changes between both high and low frequency components of the referefence and distorted images as a result of distortion. Hence, 2D DWT of the reference and distorted images up to three scales were computed using Daubechies wavelets and wavelet coefficients were used for feature extraction.

The main motivation in building the convolutional layers was provided by VGGnets. The convolutional layers in VGGnets are the composed of kernel sizes as small as 3×3 and two basic design rules are followed: (i) the layers have the same number of filters given an output size, and (ii) if the output size is halved then the number of filters is doubled in order to preserve the spatial complexity per layer. A similar architecture as explained in [He et al. (2016)] was followed, given that the residual connections result in a model that is easier to optimize and exhibit lower training error when the depth increases.

To make maximum use of the training database, the input images were divided into N_p number of patches that were selected randomly. The dimensions of each patch were determined as 128×128, thereby resulting in wavelet decompositions of size 64×64, 32×32 and 16×16. Images were normalized prior to network processing, using the mean and standard deviation of the training set. The proposed VGGnet inspired residual convolutional layers were comprised of 8 to 10 weight layers with 3 to 4 shortcut connections for wavelet coefficient inputs and color patch inputs, respectively. The features were extracted using a series of 3×3 conv 32, 3×3 conv 32, 3×3 conv 64, 3×3 conv 64, 3×3 conv 128, 3×3 conv 128, 3×3 conv 256, 3×3 conv 256, with an addition of 3×3 conv 512, 3×3 conv 512 for the color input[1]. The shortcut connections for residual architecture were established by 1×1 convolutional filters of size 64, 128 and 256, with an additional filter of size 512 for the color input. Downsampling was performed by using convolutional layers of stride 2 instead of pooling. At the end of each branch, a 1×1 convolutional layer with 16 filters was used to reduce the output size. Further dimensional reduction was performed for the branches with input size greater than 16×16 by using max pooling. All max pooling layers had 2×2 sized kernels. The output of each branch was then concatenated to form the final feature vector, as shown in Figure 3.1. The convolutional layers of same output size were activated through a Leaky Rectified Linear Unit (Leaky ReLU) where $\text{LeakyReLU}(x) = \max(0, x) + 0.01 \times \min(0, x)$. The choice of this activaton function was to allow

[1] Here, the notation $k \times k$ conv N implies the use of $k \times k$ convolutional filters with N output channels in one layer.

a small nonzero gradient when the unit is not active, thereby preventing all outputs from reducing to zero. Instead of random initial weights, the robust He initialization method [He et al. (2015)] that considers the rectifier nonlinearities was preferred for all convolutional and linear filters.

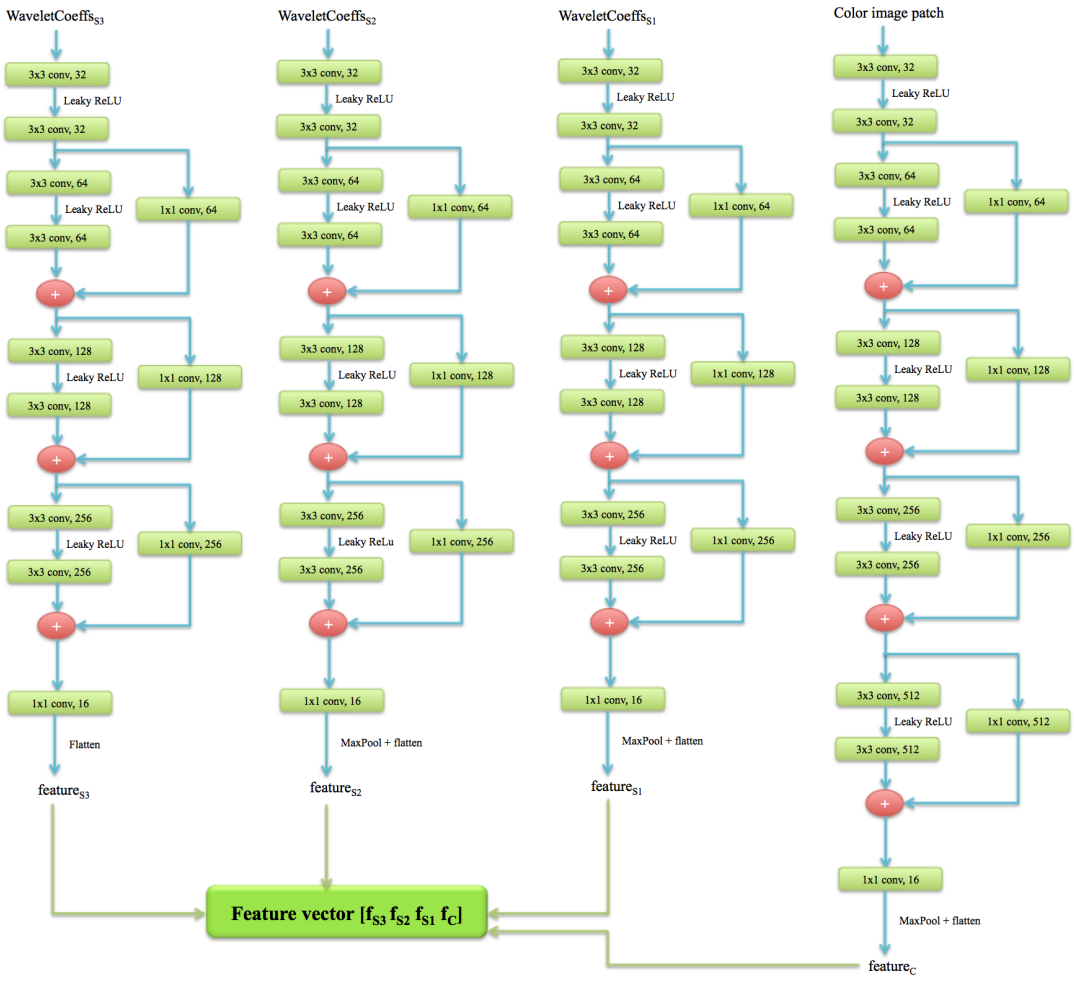


Figure 3.1 – Feature extractor composed of convolutional layers, as previously shown in Figure 3.2 as the CNN block. Inputs of the first three branches from left to right are the wavelet coefficients of the 128 × 128 image patch, where S3 corresponds to the coarsest scale and S1 corresponds to the finest scale. The rightmost branch is the color image patch branch. Features are extracted using a VGGnet inspired architecture involving shortcut connections and 1 × 1 convolution at the end for dimensional reduction. Max pooling is also applied when necessary. Feature vectors of four branches are concatenated into a final feature vector of the input image patch.

3.1.2 Score computation

Following the feature extraction of both reference and distorted image patches, the distorted image features f_D were concatenated with the reference image features f_R. Moreover, the difference vector $f_D - f_R$ was also added as the accuracy was reported to increase by using this configuration in [Bosse et al. (2018)]. The feature vectors were then passed through two fully connected layers for regression, FC 256 and FC 1. Between these layers Leaky ReLU activation was again used prior to dropout regularization with a ratio of 0.5 in order to prevent overfitting [Srivastava et al. (2014)]. The feature vector was separately fed into two fully connected layers for computing local patch weights. The architecture of this block was the same with the output regression layer, FC 256 and FC 1. Between these layers, Rectified Linear Unit (ReLU) activation prior to dropout with a ratio of 0.5 was used. Furthermore, a final ReLU activation was applied before weight computation, in order to ensure the weights to be greater than or equal to zero. Afterwards a small constant $\epsilon = 1e{-}6$ was added to the weights to prevent zero weights. The complete architecture of the proposed network is depicted in Figure 3.2.

For an input patch i, the computed weight was a_i such that $a_i = \max(0, a_i^*) + \epsilon$ where a_i^* is the output prior to ReLU activation. The quality of patch i was computed in the parallel regression branch as y_i. The overall image quality was then computed as a linear combination of patch qualities and patch weights:

$$\hat{q} = \frac{\sum_i^{N_p} a_i y_i}{\sum_i^{N_p} a_i} \tag{3.1}$$

The loss function used for training WIQM was the mean squared error between the computed image quality and the ground truth, i.e., the MOS rating of the image. The proposed network was trained iteratively by back propagation [LeCun et al. (1998, 2012)] over a number of epochs until the error was stabilized. An epoch is defined as the period during which training takes place until the whole data has been processed by the network once. For batchwise optimization, the training data was divided into batches during each epoch. In each batch, $N_p = 32$ patches were extracted from one reference image and a corresponding distorted image. Data augmentation was carried out by flipping each image from left to right and choosing additional $N_p = 32$ patches from each of the flipped images. A total of 128 image patches coming from one reference-distorted image pair were included per batch. The backpropagated loss was the average overall loss between the computed scores and MOS values of the images in each batch. As was done in [Bosse et al. (2018)], patches were randomly sampled every epoch to introduce as many different inputs as possible to the network during training. The Adam method [Kingma and Ba (2014)] was used for batch optimization with the recommended parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a decaying learning rate starting from $\text{lr} = 10^{-4}$ with a decay percentage of 10% every 5 epochs. The validation loss was computed at the end of each epoch, where the validation set had been defined at the beginning of the algorithm instead of choosing random patches at every epoch in order to ensure stability. The final

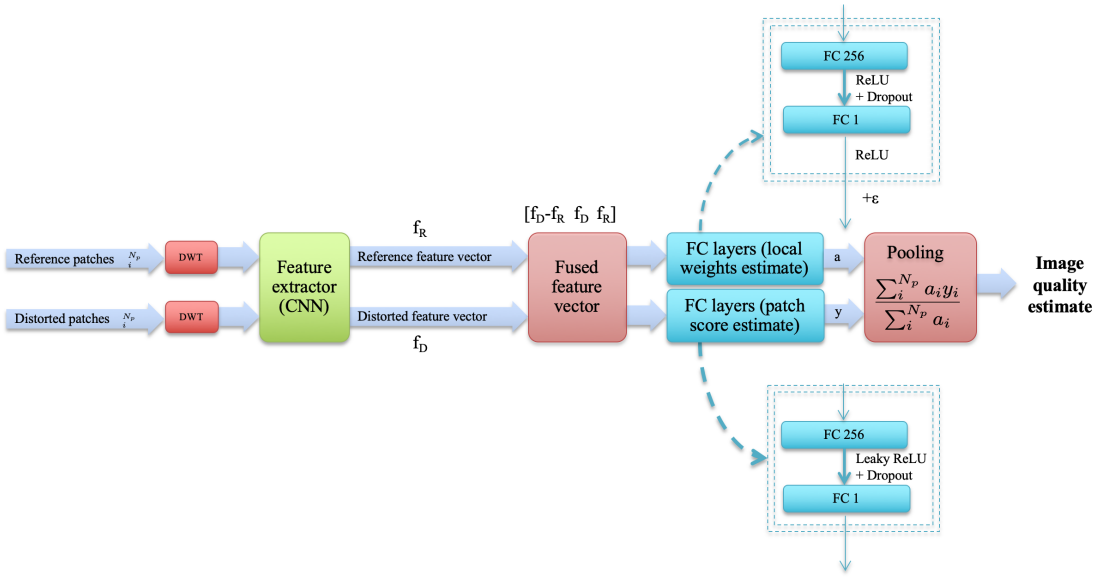


Figure 3.2 – The proposed WIQM framework for training and testing our model. Features are extracted from both reference and distorted image patches, using color information and wavelet decomposition. The reference and distorted feature vectors are concatenated, also with a third difference vector. The final feature vector is passed through parallel fully connected layers for local weight estimation and patch score estimation. Overall score of each image is computed as a linear combination of the weighted patch scores.

model used for accuracy tests was the model with least validation error, which corresponded to using early stopping criterion to stop training [Prechelt (1998)].

3.2 Results and discussion

In this section, experiments and results are presented and discussed. As a first step, the datasets used during training and evaluation are explained in detail. The most comprehensive dataset among the three introduced databases, i.e. TID2013 database, is used for training. Section 3.2.2 depicts and analyzes the results on the test set extracted from the TID2013 database. Section 3.2.3 presents the cross-database evaluation results on the CSIQ and LIVE image databases.

3.2.1 Datasets

IQA databases with MOS or DMOS ratings, i.e. TID2013, LIVE and CSIQ were considered for evaluation. TID2013 database with 25 references and 3000 distorted images, was the most comprehensive among the three and was therefore chosen for training the proposed network.

For running the presented experiments, TID2013 was separated into training, validation and test sets randomly, using 15, 5 and 5 reference images respectively. During cross-database evaluations, the DMOS values of LIVE and CSIQ datasets within ranges $[0,100]$ and $[0,1]$, respectively, were linearly mapped MOS ratings of TID2013, i.e. $[0,9]$.

3.2.2 Results on TID2013 database

Five different models on the TID2013 image database were trained using five random splits, and the average test accuracy was computed over the five test sets. The total number of epochs was 100 for each model, and for each test the model with the lowest validation loss was selected. The averaged training and validation losses are shown in Figure 3.3. The same test images were used to evaluate the PSNR, SSIM, MS-SSIM, FSIM_C and WaDIQaM-FR metrics. WaDIQaM-FR model used during the tests was downloaded from the publicly available set of models in https://github.com/dmaniry/deepIQA, where the model trained on the TID2013 database was chosen for consistency. Table 3.1 presents the performance comparison in terms of the PLCC and SROCC values with respect to the MOS values of each image, averaged over all test images.

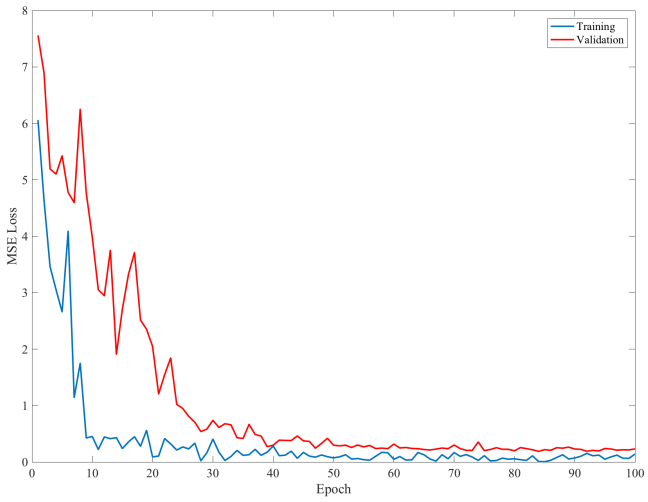


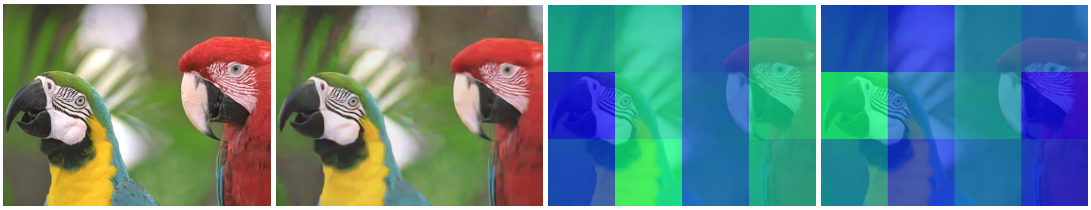
Figure 3.3 – Training and validation losses for 100 epochs on the proposed model, averaged over five runs with random splits for the training and validation sets. Each training set has 15 reference images, whereas validation sets comprise of 5 reference images.

Table 3.1 shows that the performance of WIQM was superior to other tested methods in terms of PLCC and SROCC on the TID2013 test images. To elaborate on how an image score was determined, an example test image from the database was examined in Figure 3.4. The reference image had been distorted due to an image denoising algorithm, where a distortion of level 4 out of 5 had been introduced. Although 32 overlapping random patches were used

Table 3.1 – Performance comparison of the proposed method WIQM and PSNR, SSIM, MS-SSIM, $FSIM_C$, and WaDIQaM-FR in terms of PLCC and SROCC. The reported results have been averaged over five randomly selected tests, each consisting of 5 reference images and the corresponding 600 distorted images in the TID2013 database.

IQM	PLCC	SROCC
PSNR	0.6192	0.7113
SSIM	0.6189	0.5934
MS-SSIM	0.8166	0.8278
$FSIM_C$	0.8512	0.8565
WaDIQaM-FR	0.8566	0.8488
WIQM	**0.8964**	**0.8729**

per image for training, 12 non-overlapping patches were chosen for illustration purposes, to depict local patch scores and weights. Figure 3.4(c) highlights the patch scores computed by WIQM, where the colormap changes from blue (low) to green (high). Figure 3.4(d) depicts the corresponding local weights for the patches 3.4(c). The MOS of the distorted image in Figure 3.4(b) was given as 3.6191 out of 9.000, which indicates a low perceptual quality. Indeed, as can be traced in the figure, the weights assigned to patches with high local quality are low, and weights assigned to patches with lower quality are high, resulting in an overall low score that correlates well with the MOS.

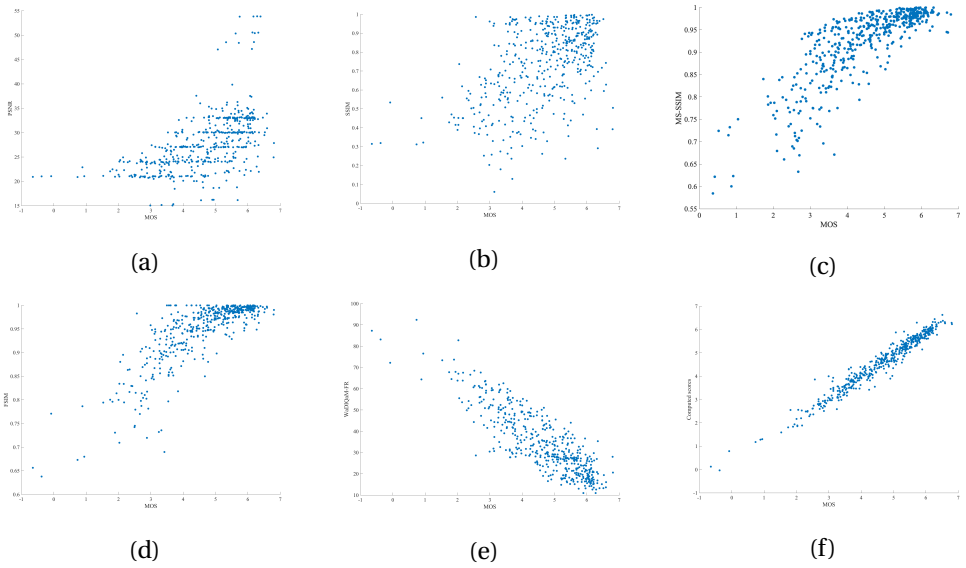


(a) (b) (c) (d)

Figure 3.4 – An example reference image from the TID2013 datased used for testing (a). The distorted image as a result of image denoising, with distortion level 4 out of 5 (b). The local scores computed by the proposed method for each non-overlapping 128×128 patch using the reference and distorted image pair, overlayed on the distorted image (c). The local weights computed by WIQM corresponding to each non-overlapping 128×128 local score patch, overlayed on the distorted image (d). For (c) and (d), the colormap changes from blue to green, where blue indicates low and green indicates high values. Mean squared error loss between the computed score and the given MOS is 4.319×10^{-5}.

Figure 3.5 depicts the computed objective metrics versus MOS values on the TID2013 image database test set. Comparing with the correlation coefficients presented in Table 3.1, it can be verified that below 80% PLCC and SROCC, the distribution of computed scores were very random. WaDIQaM-FR and WIQM were highly correlated with the MOS values, however the

variance of WaDIQaM-FR scores were higher than the proposed method.



(a) (b) (c)

(d) (e) (f)

Figure 3.5 – Computed objective metrics (a) PSNR, (b) SSIM, (c) MS-SSIM, (d) $FSIM_C$, (e) WaDIQaM-FR and (f) the proposed method WIQM on the TID2013 test images versus the MOS values.

3.2.3 Cross-database evaluation

In order to test the generalization ability of the proposed model, objective quality assessment experiments were performed on separate image databases containing different reference and distorted images. The performance of WIQM compared to that of WaDIQaM-FR on the LIVE and CSIQ image quality databases, respectively, is presented in Tables 3.2 and 3.3.

Table 3.2 – Performance comparsion of WIQM and WaDIQaM-FR in terms of PLCC and SROCC on the test images selected from LIVE database.

Distortion type	WaDIQaM-FR		WIQM	
	PLCC	SROCC	PLCC	SROCC
All	**0.8624**	**0.8843**	0.7182	0.8158
JPEG 2000	**0.8653**	**0.8749**	0.8124	0.8404
JPEG	**0.8915**	**0.8915**	0.7691	0.7508
White Noise	0.8814	0.8732	**0.9038**	**0.9338**
Gaussian Blur	**0.9419**	**0.9374**	0.7477	0.9181
Fast Fading Rayleigh	**0.9168**	0.9191	0.8193	**0.9289**

The full LIVE database is comprised of 779 distorted images, however 433 of these share common contents with that of the TID2013 database. In order not to bias the results, the common

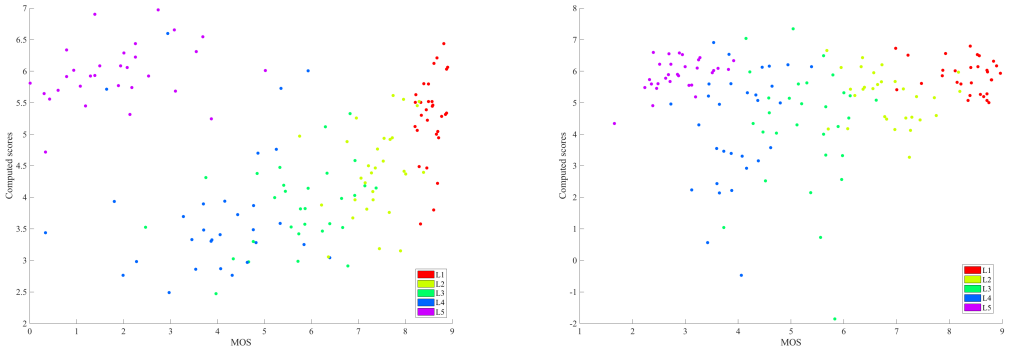
Table 3.3 – Performance comparsion of WIQM and WaDIQaM-FR in terms of PLCC and SROCC on the full CSIQ image quality database.

Distortion type	WaDIQaM-FR		WIQM	
	PLCC	SROCC	PLCC	SROCC
All	**0.6261**	**0.6426**	0.5190	0.5434
Noise	**0.9309**	**0.9161**	0.8769	0.8144
JPEG	**0.9250**	0.9008	0.8769	**0.9311**
JPEG2000	**0.9186**	0.8427	0.8028	**0.8914**
Frequency noise	**0.1859**	**0.0616**	0.1219	0.0424
Gaussian Blur	**0.2970**	**0.3620**	0.0936	0.0089
Contrast	0.6990	0.9167	**0.7325**	0.7496

contents were not included as test material for quality assessment on LIVE database. The DMOS values reported for the LIVE database were also scaled to conform with the range used during training of WIQM. The results depicted in Table 3.2 were evaluated on the remaining 346 distorted images of the LIVE database. Reported correlations concerning the proposed method were averaged on the five different random split models. A closer examination indicates that on the full LIVE database, the overall performance of WaDIQaM-FR was superior to WIQM in terms of both PLCC and SROCC. The same trend is observed for distortions caused by JPEG and JPEG 2000 compression and Gaussian blur. WIQM results are highly correlated with the MOS values in the case of white noise, and fast fading Rayleigh, i.e. bit errors.

None of the 866 distorted images in the CSIQ image database share the same content with TID2013, hence no content was left out for tests on the CSIQ image database. The CSIQ database tests were therefore more comprehensive compared to the tests on LIVE database. For the experiments, the DMOS values of the CSIQ database were again scaled to conform with WIQM training values. Table 3.3 indicates that on the full database, the proposed model has higher correlation with the reported MOS of CSIQ database only for JPEG and JPEG 2000 distortions in terms of SROCC, and contrast distortions in terms of PLCC. The overall correlation of the proposed method with the underlying MOS values was much lower than WaDIQaM-FR, which was mostly brought down by the evaluations on frequency noise and even more by Gaussian blur, where reported scores exhibited a random distribution. WaDIQaM-FR correlations have also dropped down for these two types of distortions, however less drastically. In Figure 3.6(a), the ratings of WIQM are very high for the highest level of blur distortion in the CSIQ database. These erroneous ratings were partly due to the difference between the levels of blur distortion introduced in TID2013 and CSIQ databases, where the highest level of distortion in the CSIQ database is higher compared to the maximum blur distortion in the training data. With such elevated levels of distortion, high frequency features of the image are lost, resulting in a loss of information brought by the wavelet coefficients. In this case the blurred image can be regarded as a very smooth and therefore high quality content. A different effect was observed in the case of additive pink Gaussian noise, as depicted in Figure 3.6(b).

This type of noise had not been introduced in the training set, therefore both WIQM results and WaDIQaM-FR have reduced correlation levels. As the level of frequency noise is increased, sharp frequency features from wavelets elevate the ratings in the proposed model, resulting in uncorrelated scores with the underlying MOS.



(a) (b)

Figure 3.6 – Computed objective metrics versus MOS on CSIQ database for Gaussian blur (a) and additive pink Gaussian noise (b) types of distortion. L1 indicates the lowest level of distortion while L5 indicates the highest level of distortion.

The main differences between WIQM and WaDIQaM-FR is the inclusion of wavelet decomposition as auxiliary information for quality evaluation, and the distinct architectures of feature extraction layers. Focusing on the auxiliary information, it can be concluded that the frequency characteristics of the reference and distorted images helped the model to perform fair objective quality assessment on the TID2013 database. The generalization ability of the model, however, is limited. Nevertheless, competitive performance of the model with the state-of-the-art in cross database results is observed, with superior correlation values with underlying MOS values for white noise and fast fading Rayleigh distortions in LIVE database, and JPEG, JPEG 2000 and contrast distortions in CSIQ database, respectively.

In order to examine the effect of the wavelet decomposition in WIQM, additional experiments were conducted, where the color image patch branch was removed from the architecture in Figure 3.1. In this scenario, only the wavelet coefficients were used for feature extraction and color information was ignored. Table 3.4 depicts the extent of influence brought by the wavelet coefficients to the proposed scheme. Although the color information is undoubtedly useful for objective quality assessment, only the features extracted using the three scale wavelet decomposition are observed to be capable of capturing the quality level of the distorted image compared to the reference with high accuracy.

Examining the residual architecture of feature extraction layers shows that using the 2D DWT as a preprocessing step helps the training error converge much faster compared to the WaDIQaM-FR model. In less than 40 epochs, WIQM model was able to reach the accuracy lev-

Table 3.4 – Influence of wavelet coefficients in objective quality assessment for WIQM. The correlations were averaged over the test sets of TID2013 in five random splits.

Feature extraction	PLCC	SROCC
Wavelet coefficients	0.8572	0.8372
Wavelet coefficients + Color	0.8964	0.8729

els presented in this chapter. On the contrary, preferred WaDIQaM-FR models were expected to converge to a stable loss around 1000 epochs [Bosse et al. (2018)]. In terms of complexity, the proposed model has approximately 9.5M parameters and the network experienced around 14M different training patches in 100 epochs. The complexity of WaDIQaM-FR is much lower with approximately 5.2M parameters, where the number of epochs were reported as 3000 in [Bosse et al. (2018)], resulting in around 178M patches introduced to the system during training. Although WIQM has higher complexity, it is able to learn important features for objective quality assessment relatively faster and with less number of inputs.

In order to increase the generalization ability of WIQM, a thorough exploration of the hyper-parameters is of key importance. This involves testing the effect of using different number of scales, patch sizes, number of patches and learning rates. The relatively low accuracy levels on the LIVE database suggest the need to better incorporate the influence of frequency information in the feature extraction step. Current feature vectors are a concatenation of equally weighted wavelet features in different scales and the color features. Similar to learning the weight of local patches on the overall image quality, the weights of these features can be learned to yield more descriptive feature vectors and hence more accurate results at the expense of increased computational complexity.

3.3 Conclusion

In this chapter, a new full reference objective metric to predict distorted image quality using deep neural networks was presented. The proposed WIQM model employs both the color features of the reference and distorted images, as well as the frequency characteristics extracted from a 3-scale discrete wavelet transform using Daubechies wavelets. The feature extraction module of WIQM was inspired by VGGnets and has shortcut connections that build residual blocks. A Siamese network architecture was used to extract features from the reference and distorted images simultaneously. Regression of patch feature vectors into global image scores was carried out using fully connected layers, computing local scores and their respective weights. The proposed model was trained on the TID2013 image database, and tested on the same database as well as LIVE and CSIQ image quality databases. Results on the TID2013 database showed superior prediction accuracy compared to the state-of-the-art methods. Results on cross-database evaluation indicated that for particular types of distortion the proposed method performs better than the state of the art, however there is room for

improvement in the interpretation of the frequency information for distortion types and levels that have not been included in the training. The contributions of this work can be emphasized as follows:

- The 2D DWT introduced as a preprocessing step before feature extraction constituted the core novelty in this work. The wavelet transform helped boost the performance of the proposed metric on the TID2013 test set and yielded high correlations with the underlying MOS ratings, with results better than the state-of-the-art traditional and learning-based image quality metrics.
- The favorable effect of 2D DWT was also demonstrated through additional experiments showing that high PLCC and SROCC values could be obtained using only wavelet coefficients as input to the model, discarding the color input.
- A thorough analysis on cross-database performance was provided using the results on LIVE and CSIQ databases. The proposed model was superior to WaDIQaM-FR for certain types of distortions in terms of PLCC and/or SROCC, namely, when scoring images contaminated with white noise in LIVE, and JPEG and JPEG 2000 artifacts in CSIQ databases. The performances of WIQM and WaDIQaM-FR were comparable for all other types of distortions, yet the cases where the performance of WIQM was inferior in CSIQ database were investigated. It was concluded that for the types of distortion never seen by the network during training, the prediction performance fell drastically. Moreover, it was observed that during cross-database evaluation, the nonlinear relationship between distortion levels and rating scales of distinct training and test databases result in WIQM scores diverging from the MOS of the test database.
- The presented work highlighted two problems in learning-based IQA:

    1. A good mapping between the scores and distortions across different IQA databases does not exist. The assumption of a linear mapping between the scores is not reliable.
    2. The three databases involved in this work are the major databases used in IQA related research. Neither of the resolutions of the three databases reach HD. Building a comprehensive high resolution database is a requirement for training and testing learning-based image quality metrics in order to present an effective IQM with the data in use today.

In order to improve the cross-database performance of WIQM, a more robust feature extraction strategy could be employed through various steps, including using regression for aggregation of features extracted from the reference and distorted image patches, enhancing the training set that involves more types and levels of distortions, exploring different hyperparameters such as learning rate decay, number of patches, patch dimensions and number of wavelet decomposition scales. The model could also benefit from using a different loss function

instead of the mean squared error, such as the correlation between batch scores and the ground truth, following the same strategy for evaluation. Alternatively, the local weighting of patch scores could be strengthened using robust saliency models based on image quality evaluation in addition to the weighted patch aggregation model presented in this work. Following these modifications, one promising future work would be to adapt the proposed objective image quality metric for video, making use of the temporal correlations between frames and implementing an end-to-end optimization framework for predicting video quality.

Before trying to enhance the network model following the suggestions above, a crucial step is to tackle the problem of the lack of a comprehensive HD database. Moreover, looking at IQA from a streaming and broadcasting point of view, it is important for such a database to comprise of various levels of distortions brought by state-of-the-art compression algorithms. In the next chapter, the construction of a new compression IQA database is presented that is later used to improve WIQM and train other IQM models.

4 Building a high definition image database for compression quality evaluation

Personal contribution: All subjective test material was prepared and all analysis of the collected results was carried out by the author. Tests performed at Ecole Polytechnique Fédérale de Lausanne (EPFL) were conducted under author's supervision.

From TV shows to films in movie theaters, from photos and animations in our smart phones and tablets to the advertisements on billboards, every visual content we receive is compressed to a certain extent. There are numerous different modalities of multimedia, including planar imaging and video, omnidirectional imaging and video, point clouds and light field contents. The goal of service providers and broadcasters is to deliver high quality to end users, where the quality measure is not independent from the context. From a multimedia delivery point of view, quality needs to be calibrated depending on the task at hand and the data involved. For applications such as picture browsing, social media, video on demand and broadcasting, high quality within all multimedia modalities is sought in the presence of compression at varying levels. The performances of compression solutions then need to be optimized to yield an efficient rate-distortion trade-off. In order for such optimization to be carried out effectively, acquiring a complete dataset is crucial.

Cameras, productions and broadcasts are currently able to employ resolutions up to 8K, which

raises the need to deliver high quality content at such high resolutions. Encoders are trying to be optimized so that vast amounts of data can be compressed while retaining the best visual quality. The performance of codecs compressing a range of resolutions of data from SD to UHD needs to be assessed using a metric that is able to capture the characteristics of such data and measure the quality accurately. Thus, when developing image compression standards and objective quality metrics, this reciprocal effect needs to be considered. The answer to how to compress high resolution data concerns the quality measure, and the development of the quality measure depends on many intrinsic data factors including resolution.

In this chapter, the construction of two distinct IQA databases are presented, which are composed of a total of 15 reference images of SD to UHD resolution with 8-bit depth. The first database was constructed as part of the efforts of Joint Photographic Experts Group for creating the next generation image coding standard, referred to as JPEG XL[1]. The second database was created within the scope of activities of the new JPEG initiative, JPEG AI, which aims to investigate current learning-based image compression architectures and their performances. JPEG XL and JPEG AI databases were both evaluated objectively and subjectively, in which subjective tests were conducted in controlled environments with the participation of a total of 111 subjects. Results of the analyses were valuable not only in analyzing and evaluating the compression performance of state of the art codecs but also in terms of the creation of a database that can be used as a baseline for IQA, codec comparison and benchmarking.

4.1 JPEG XL database construction

Joint Photographic Experts Group is a sub-group of ISO/IEC Joint Technical Committee 1, Subcommittee 29, Working Group 1 (ISO/IEC JTC1/SC29/WG1). The group has created the image coding standard JPEG more than two decades ago, as well as many other standards including JPEG 2000, JPEG XS, JPEG XT and JPEG XR. While maintaining the previous standards, JPEG is active in development of novel coding standards in multimedia. In 2017, JPEG published a Call for Proposals for creating the next generation image coding standard, referred to as JPEG XL. The Call targeted development of a standard for image coding that offers substantially better compression efficiency than existing image coding formats (e.g. $>60\%$ over JPEG), along with features desirable for web distribution and efficient compression of high-quality images.

The final Call for Proposals was issued in April 2018 with a deadline for expression of interest and registration in August 2018. Submissions were gathered in September 2018 and the performance of proponents was evaluated via subjective and objective quality assessment tests, following the Recommendations in ITU-R BT.2022 [ITU-R BT.2022 (2012)], ITU-R BT.500-13 [ITU-R BT.500-13 (2012)] and ITU-T P.910 [ITU-T P.910 (2008)]. A total of seven proponents were compared to four anchors at eight different bitrates during objective and four different bitrates during subjective evaluation. All proponents were evaluated using Standard Dynamic Range (SDR) and HDR contents at various resolutions with different characteristics. The

[1] https://jpeg.org/jpegxl/index.html

results of the quality assessment of JPEG XL were first published as an output document of the 81st JPEG meeting, Vancouver, Canada, 13-19 October 2018. The submissions were evaluated by researchers at three different universities, namely, EPFL, Vrije Universiteit Brussel (VUB) and Télécom ParisTech (TPT), using the contents and methodologies defined in the Call.

4.1.1 Dataset

The goal of JPEG XL is to develop a new image coding standard that provides state-of-the-art image compression performance, and that addresses shortcomings in current standards. The activity aims to (i) achieve significant compression efficiency improvement over coding standards in common use at equivalent subjective quality, e.g. > 60% over JPEG, (ii) offer features for web applications, such as support for alpha channel coding and animated image sequences, and (iii) offer support of high-quality image compression, including higher resolution, higher bit depth, higher dynamic range, very high quality and wider color gamut coding. To encourage widespread adoption, an important goal for this standard is to support a royalty-free baseline.

JPEG XL targets a large variety of use cases including image-rich UIs and web pages on bandwidth-constrained connection such as social media applications, media distribution applications, cloud storage applications, media web sites, animated image applications, mobile applications and games, and high quality imaging applications such as rapid photo viewing, HDR/Wide Color Gamut (WCG) user interfaces, augmented/virtual reality, image bursts, high-end photography, image mosaics, depth images, and printing.

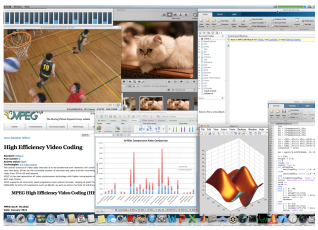
With such a wide range of specifications for the standard, it is important to construct an image dataset that is able to include examples reflecting the target requirements and use cases. Therefore, a total of 67 images were considered for evaluation which were representatives of 5 different categories. The specifications of the full dataset are given in Table 4.1. Resolutions of SDR images varied from SD to UHD whereas HDR images had HD resolution. SDR color images had BT.709 primaries whereas HDR images had BT.2020 primaries. Contents selected for subjective quality assessment experiments for SDR and HDR tests are presented in Figures 4.1 and 4.2, respectively. When selecting the contents for subjective evaluation, expert viewing sessions were conducted to make sure all classes were represented in the experiments and images with distinctive spatial characteristics were involved.

Table 4.1 – Distribution of full set of contents.

Class	Description	Bit depth	Number of contents
A	Natural images (RGB 4:4:4)	8-bit 10-bit	23 10
B	Grayscale images (4:0:0)	8-bit	4
C	Computer generated images (RGB 4:4:4)	8-bit 10-bit 12-bit	1 1 1
D	Screen content images (RGB 4:4:4)	8-bit	3
E	HDR/WCG images (RGB 4:4:4)	12-bit	24
All			**67**

(a) Training

(b) Arri



(c) Apple

(d) Bike

(e) Cafe

(f) Fly

(g) p06

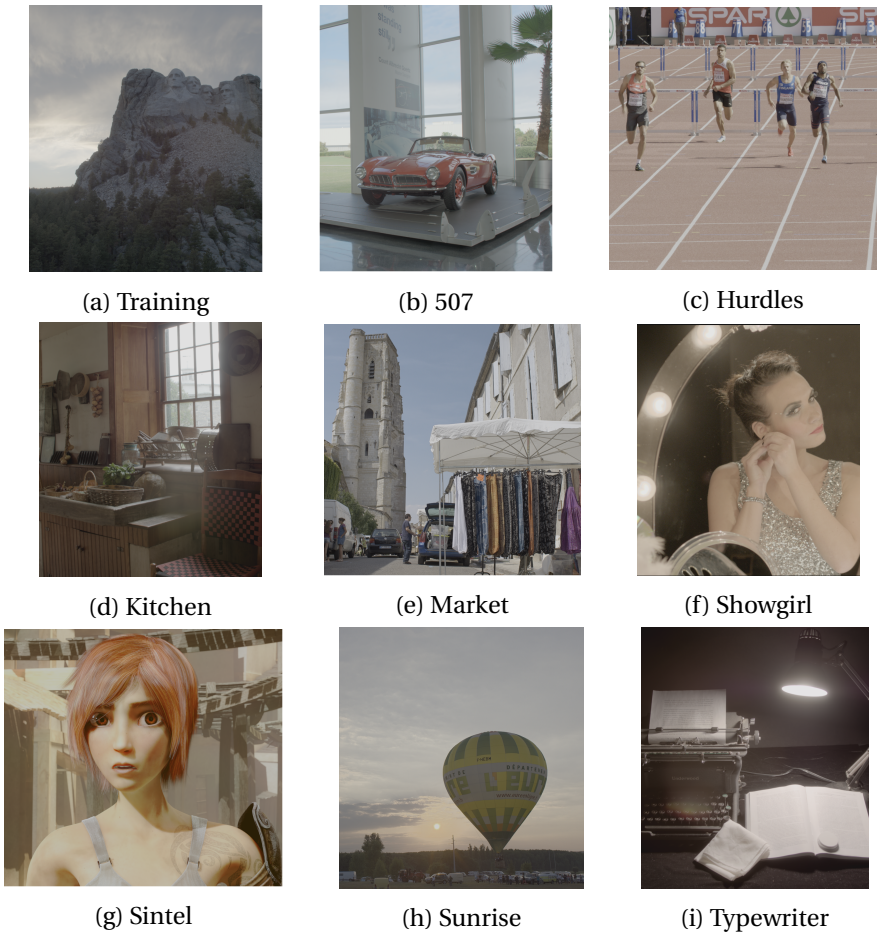
(h) Blender (10-bit)

(i) Woman

Figure 4.1 – Thumbnails of SDR contents selected for subjective quality assessment, after cropping for DSIS experiments. All contents have 8-bit depth except for Blender.

4.1.2 Anchor generation

The proposals were evaluated against four anchors: JPEG, JPEG 2000, HEVC/H.265 and WebP (only for 8-bit SDR contents). At the evaluation time, a specific reference implementation was chosen for each anchor, as follows: JPEG XT reference software (v1.53) for JPEG,



(a) Training (b) 507 (c) Hurdles

(d) Kitchen (e) Market (f) Showgirl

(g) Sintel (h) Sunrise (i) Typewriter

Figure 4.2 – Thumbnails of HDR contents selected for subjective quality assessment, after cropping for DSIS experiments. Linear RGB thumbnails are included here only for demonstration.

Kakadu (v7.10.2) for JPEG 2000, HM-16.18+SCM-8.7 for HEVC/H.265 and cwebp 1.0.0 for WebP. Each selected content was encoded by all anchors using 8 target bitrates in the list $[0.06, 0.12, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00]$bpp. The proponents were required to target the same bitrates over all contents, both for SDR and HDR data. The bitrate range was selected extensive enough to investigate all rate-distortion scenarios from very low rates corresponding to very low image quality to very high bitrates corresponding to transparent image quality. Configurations and command lines for anchor generation are given in Table 4.2. The 12-bit setting was used for HDR images. SDR and HDR command lines were different only for HEVC/H.265.

All input images were in RGB 4:4:4 format. The conversions were handled using HDRConvert[2] with the following command line:

[2]https://gitlab.com/standards/HDRTools/tree/master

HDRConvert -f HDRConvertBT709PPMToYCbCr420fr.cfg -p SourceFile=<RGB444_input>
-p SourceWidth=<width> -p SourceHeight=<height> -p OutputFile=<YCbCr420_input>
-p OutputWidth=<width> -p OutputHeight=<height> –p SourceBitDepthCmp0=<bit_depth>
-p SourceBitDepthCmp1=<bit_depth> -p SourceBitDepthCmp2=<bit_depth>
-p OutputBitDepthCmp0=<bit_depth> -p OutputBitDepthCmp1==<bit_depth>
-p OutputBitDepthCmp2=<bit_depth> -p OutputChromaFormat=1

JPEG XT accepts only 4:4:4 chroma format and the subsampling to 4:2:0 is executed internally. For JPEG XT, the parameter OutputChromaFormat was set to 3 instead of 1. The codebase for anchor generation is available online[3], including the configuration files used during conversions. To ensure reproducibility of results and ease the objective assessment of different proposals, a Docker container was created to automatically perform the objective assessment of a given set of codecs. The container (i) automatically downloads and configures all anchor codecs, metrics and dependencies, (ii) allows easy addition of new (proprietary) codecs by placing binaries and Python encoder/decoder scripts in the designated folder, (iii) allows testing new contents, (iv) includes all running encoding, decoding, and objective evaluation scripts, and (v) automatically generates performance curves of objective results.

4.1.3 Objective quality assessment

Objective quality assessment was carried out over all 8 bitrates for all codecs, in YCbCr color space. The RGB 4:4:4 outputs were converted to YCbCr 4:4:4 using HDRConvert. Selected metrics for objective quality assessment of SDR contents were PSNR, SSIM, MS-SSIM for all contents, with the addition of VIF and VMAF only for 8-bit contents. The first three metrics have been computed using HDRMetrics[4] whereas the VMAF FFmpeg plugin is used for the last two. PSNR is computed on Y channel, and by averaging the PSNR over separate channels.

For HDR contents, PQ-PSNR-Y and PQ-MS-SSIM-Y metrics were computed using HDRMetrics. In order to carry out objective evaluation on HDR images, inverse Perceptual Quantizer (PQ) transfer function was applied first, leading to 12-bit PQ-RGB 4:4:4 images to obtain linear RGB images. Then, following color space conversion from linear RGB to XYZ, PQ transfer function was applied to Y component and PSNR and MS-SSIM metrics were computed on the Y component only. All command lines are provided in Table 4.3 and the configuration files are available online[5].

4.1.4 Subjective quality assessment

DSIS Variant I [ITU-R BT.2022 (2012)] was the test methodology selected for subjective quality assessment. In this test, the stimulus under assessment and the reference are presented

[3] https://github.com/pinarakyazi/codec_compare
[4] https://gitlab.com/standards/HDRTools/tree/master
[5] https://github.com/pinarakyazi/codec_compare

Table 4.2 – Selected parameters and settings for the anchors.

Anchor	Software	Input format	Command line
JPEG	JPEG XT v1.53	RGB 4:4:4 8-bit	jpeg -qt 3 -h -v -oz -q <qp> -s 1x1,1x1,1x1 <input> <output>
		RGB 4:4:4 10-bit	jpeg -qt 3 -g 1 -h -v -oz -q <qp> -R 2 -s 1x1,1x1,1x1 <input> <output>
		RGB 4:4:4 12-bit	jpeg -qt 3 -g 1 -h -v -oz -q <qp> -R 4 -s 1x1,1x1,1x1 <input> <output>
		YCbCr 4:2:0 8-bit	jpeg -qt 3 -h -v -c -oz -q <qp> -s 1x1,2x2,2x2 <input> <output>
		YCbCr 4:2:0 10-bit	jpeg -qt 3 -g 1 -h -v -c -oz -q <qp> -R 2 -s 1x1,2x2,2x2 <input> <output>
		YCbCr 4:2:0 12-bit	jpeg -qt 3 -g 1 -h -v -c -oz -q <qp> -R 4 -s 1x1,2x2,2x2 <input> <output>
JPEG 2000	Kakadu v7.10.2	RGB 4:4:4 8/10/12-bit	kdu_compress -i <input> -o <output> –rate <bpp>
		YCbCr 4:2:0 8/10/12-bit	kdu_v_compress -i <input> -o <output> –rate <bpp> -precise -tolerance 0
HEVC	HM-16.18+SCM-8.7	RGB 4:4:4 8/10-bit YCbCr 4:2:0 8/10-bit	TAppEncoderStatic -c encoder_intra_main_scc.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height> –InputChromaFormat=<chroma_format> –InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth> –OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 -i <input> -b <output> -o /dev/null
		RGB 4:4:4 12-bit YCbCr 4:2:0 12-bit	TAppEncoderStatic -c encoder_intra_main_rext.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height> –InputChromaFormat=<chroma_format>– InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth>– OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 -i <input> -b <output> -o /dev/null
		RGB 4:4:4 12-bit HDR YCbCr 4:2:0 12-bit HDR	TAppEncoderStatic -c encoder_intra_main_rext.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height> –InputChromaFormat=<chroma_format>– InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth>– OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 InputColourSpaceConvert=RGBtoGBR -i <input> -b <output> -o /dev/null
WebP	cwebp 1.0.0	YCbCr 4:2:0 8-bit	cwebp -m 6 -q <qp> -s <width> <height> <depth> <input> -o <output>

Table 4.3 – Command lines for objective metric computations.

DR	Metric	Software	Command line
SDR	PSNR, SSIM, MS-SSIM	HDRMetrics	HDRMetrics -f HDRMetrics.cfg -p Input0File=<reference> -p Input1File=<decoded> -p LogFile=<log_file> -p NumberOfFrames=1 -p Input0Width=<width> -p Input0Height=<height> -p Input1Width=<width> -p Input1Height=<height> -p TFPSNRDistortion=0 -p EnablePSNR=1 -p EnableSSIM=1 -p EnableMSSSIM=1
	VMAF, VIF	FFmpeg	ffmpeg -s:v <width>,<height> -i <decoded> -s:v <width>,<height> -i <reference> -lavfi libvmaf=log_fmt=json:log_path=<log_file> -f null -
HDR	PQ-PSNR-Y PQ-MS-SSIM-Y	HDRMetrics	HDRMetrics -f HDRMetrics.cfg -p Input0File=<reference> -p Input1File=<decoded> -p LogFile=<log_file> -p NumberOfFrames=1 -p Input0Width=<width> -p Input0Height=<height> -p Input1Width=<width> -p Input1Height=<height> -p TFPSNRDistortion=1 -p EnablePSNR=1 -p EnableMSSSIM=1

simultaneously to the subject. The subject is asked to rate the degree of annoyance of the visual distortions in the stimulus under assessment with respect to the reference. The degree of annoyance is divided into five different levels labeled as Very annoying, Annoying, Slightly annoying, Perceptible but not annoying and Imperceptible, corresponding to a quality scale ranging from 1 to 5, respectively.

Content and rate point selection

Subjective tests are costly in terms of time and effort with the need of a minimum of 15 participants per experiment to generalize the results. The duration of each assessment depends on the number of contents to be evaluated. To balance this trade-off, selection of contents and rate points has to be handled meticulously.

The content and rate selection was carried out during expert viewing sessions prior to setting up the experiments, both for SDR and HDR tests. All contents in the dataset were encoded using the anchor software and the decoded images were viewed by experts. In order to obtain meaningful results from the experiments, the selected rate points needed to span a range that covers very low to high bitrates, corresponding to very low to transparent visual quality. The anchor with the best performance, i.e. HEVC/H.265, was used to select such rate points and the selection was verified using other anchors. The contents that were too difficult to examine by naive subjects were excluded from the selection.

Data preparation

Selected contents were then processed according to the DSIS framework. For SDR and HDR tests, a 30 inch Eizo 10bit ColorEdge CG301W monitor with a resolution of 4096 × 2160 and a Sim2 HDR47ES4MB display with 1920 × 1080 resolution were used by all participating labs, respectively. SDR and HDR stimuli were cropped using FFmpeg[6] to fit their respective screen resolutions. The region to be cropped for each stimulus was determined during expert viewing. Each decoded stimulus was placed side by side with its reference, with a 20 pixel mid-gray colored separation in between. The side-by-side stimuli were then displayed in front of the same mid-gray colored background, and were randomized such that the same content was never presented consecutively [ITU-R BT.2022 (2012)]. Two dummy sequences were included in each test, about which the subjects were not informed. A training session was conducted for each subject prior to the experiment, during which three stimuli were presented as examples for the two extremes of the voting scale, i.e. Very annoying (1) and Imperceptible (5), along with an example in the middle, i.e. Slightly annoying (3). For half of the subjects the reference was placed at the right side of the screen, whereas for the other half it was placed on the left to avoid position bias. Each experiment was conducted in two sessions to prevent subject fatigue. The monitors were calibrated using an i1 DisplayPro color calibration device according to the guidelines described in [ITU-R BT.2022 (2012); ITU-R BT.2100 (2018)]. Same guidelines were

[6] http://ffmpeg.org

followed to set up the controlled environment for viewing with a mid gray level background inside the test rooms.

(a) (b)

Figure 4.3 – Consenting subjects during SDR (a) and HDR (b) subjective quality assessment tests conducted at EPFL.

During both SDR and HDR tests, viewing time was not restricted. Subjects, however, were instructed to vote within reasonable time for the experiments to proceed smoothly. No viewing distance or position was specified for the SDR tests. On the other hand, HDR tests were conducted with a fixed distance from the screen as instructed by ITU-R BT.2100. Figure 4.3 depicts the test environment for SDR and HDR experiments.

4.2 Experiments and results on JPEG XL database

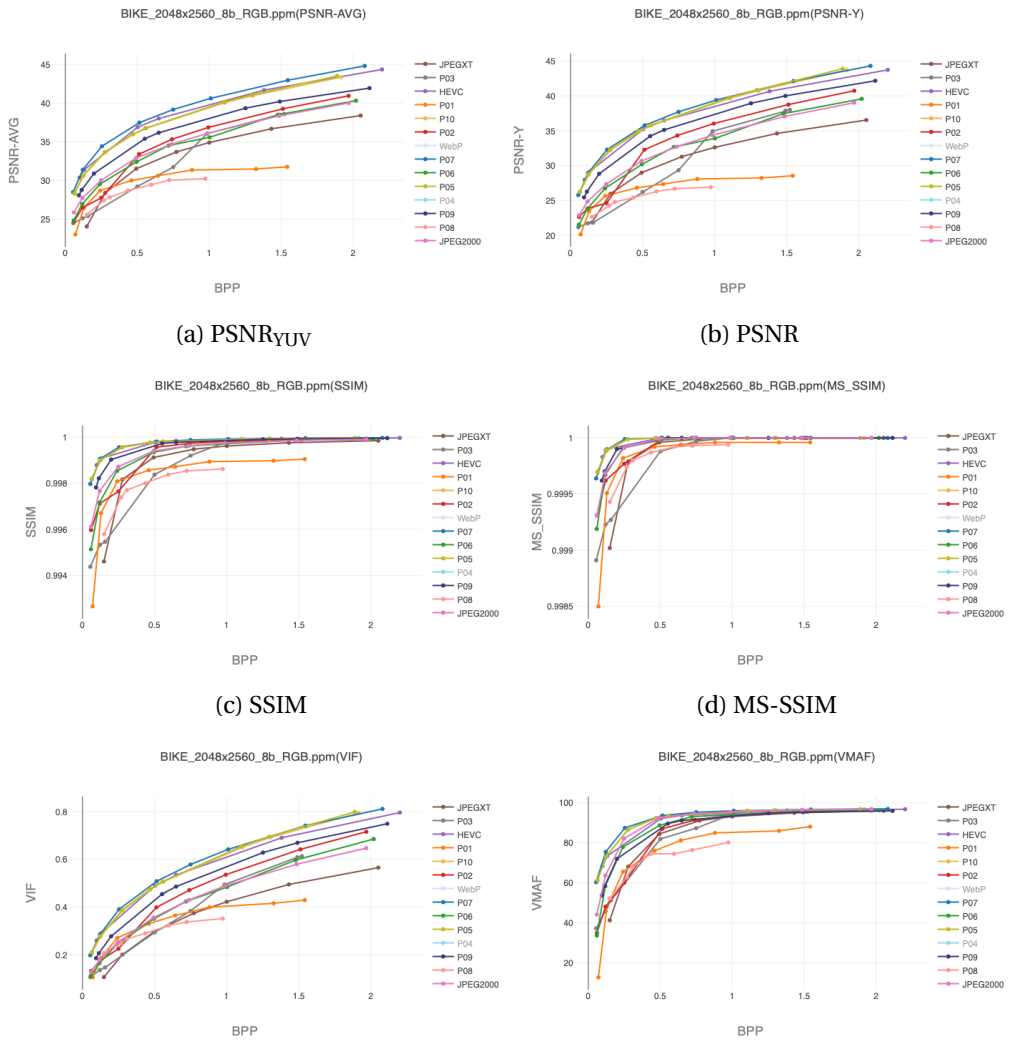
Seven proposals were submitted that fit the requirements of the Call, of which 4 supported bit depths larger than 8, and HDR images. One proponent, P01, was a Convolutional Neural Network (CNN)-based architecture trained end-to-end. 18 and 20 subjects participated to the subjective quality assessment experiments of the SDR contents in EPFL and VUB. 18, 20 and 17 subjects participated to the subjective quality assessment experiments of the HDR contents in EPFL, VUB and TPT, respectively. A standard outlier detection was performed on all sets of raw scores to remove subjects whose ratings deviated strongly from others [ITU-R BT.500-13 (2012)]. None of the subjects were identified as outliers in the experiments.

The MOS and 95% confidence interval (CI)s assuming a Student's t-distribution of the scores were computed for each test condition [De Simone et al. (2011)]. To determine and compare the differences among MOS values obtained for different codecs and bitrates, a one-sided Welch test at 5% significance level was performed on the scores. Bitrates that deviated more than 10% from the target bitrates were excluded from statistical significance tests.

4.2.1 Objective quality assessment results

Objective quality assessment was performed on all 67 contents listed in Table 4.1, at all 8 bitrates, for all proponents and anchors. Interactive plots were generated using the scripts

online. For demonstrative purposes, the objective quality assessment results of the contents Bike in Figure 4.1d and 507 in Figure 4.2b are presented in Figures 4.4 and 4.5. All objective quality assessment results are stored in .json format along with interactive plots. The data is available online and can be accessed by contacting the author.



(a) $PSNR_{YUV}$ (b) PSNR

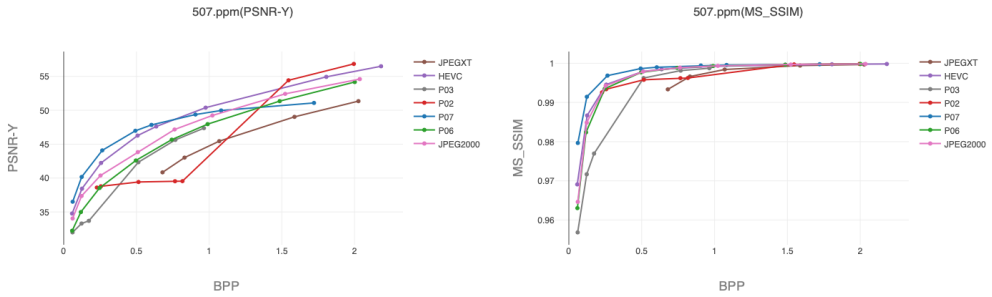
(c) SSIM (d) MS-SSIM

(e) VIF (f) VMAF

Figure 4.4 – Objective results for the SDR content Bike (Figure 4.1d). Results for codecs accepting RGB 4:4:4 as native format are included in the objective comparison.

4.2.2 Subjective quality assessment results

Subjective quality assessment was performed on the selected contents at the screened out bitrates given in Table 4.4, for all proponents and anchors. The MOS vs. bitrate plots and comparisons between pairwise conditions are presented in Figures 4.6-4.9.



(a) PSNR (b) MS-SSIM

Figure 4.5 – Objective results for the HDR content 507 (Figure 4.2b).

Table 4.4 – Original resolutions, classes and selected bitrates for subjective quality assessment of SDR contents.

Name	Class	Resolution	Bitrates
Arri	A	2880 × 1620	[0.06, 0.12, 0.25, 0.50]
Apple	D	2560 × 1440	[0.06, 0.12, 0.25, 0.50]
Bike	A	2048 × 2160	[0.06, 0.12, 0.25, 0.50]
Cafe	A	1280 × 1600	[0.06, 0.12, 1.00, 2.00]
Fly	A	1920 × 1080	[0.06, 0.12, 0.25, 0.50]
p06	A	4064 × 2704	[0.06, 0.12, 0.25, 0.50]
Blender	C	4096 × 1744	[0.06, 0.12, 0.25, 0.50]
Woman	A	2048 × 2560	[0.06, 0.12, 0.25, 0.50]
507	E	944 × 1080	[0.06, 0.12, 0.50, 1.00]
Hurdles	E	1920 × 1080	[0.50, 0.75, 1.00, 2.00]
Kitchen	E	944 × 1080	[0.06, 0.12, 0.25, 0.75]
Market	E	1920 × 1080	[0.75, 1.00, 1.50, 2.00]
Showgirl	E	944 × 1080	[0.75, 1.00, 1.50, 2.00]
Sintel	E	944 × 1080	[0.75, 1.00, 1.50, 2.00]
Sunrise	E	1920 × 1080	[0.50, 0.75, 1.00, 2.00]
Typewriter	E	944 × 1080	[0.75, 1.00, 1.50, 2.00]

Throughout SDR and HDR contents, the proponents P03, P06 and P07 were performing as good as, and even better, than the state-of-the-art codecs. The performance of P01 and P05 were also ample, however, these codecs did not support images with bit depths higher than 8. P07 reached transparent quality at the highest bitrate tested for all contents. P03 also reached transparent quality at the highest bitrate except for the screen content Apple. P06, on the other hand, performed below transparent quality for content Arri and Blender. P03 was the best performer at the highest bitrate of 10-bit computer generated image Blender. It must be noted that the confidence intervals of the competing codecs at selected bitrates were usually overlapping. Examining statistically significant differences between codec performances per bitrate per codec on the left side of Figures 4.6-4.9 show that P03 was indeed performing better than all other proponents for Blender image at the highest bitrate. P03 also performed better than all other proponents except P01 on the Woman image at the highest bitrate, followed by P03 and P01. At intermediate bitrates, performances of P06 and P03 decreased especially for complex contents such as Bike. Statistically significant differences were not observed at the lowest bitrate in general.

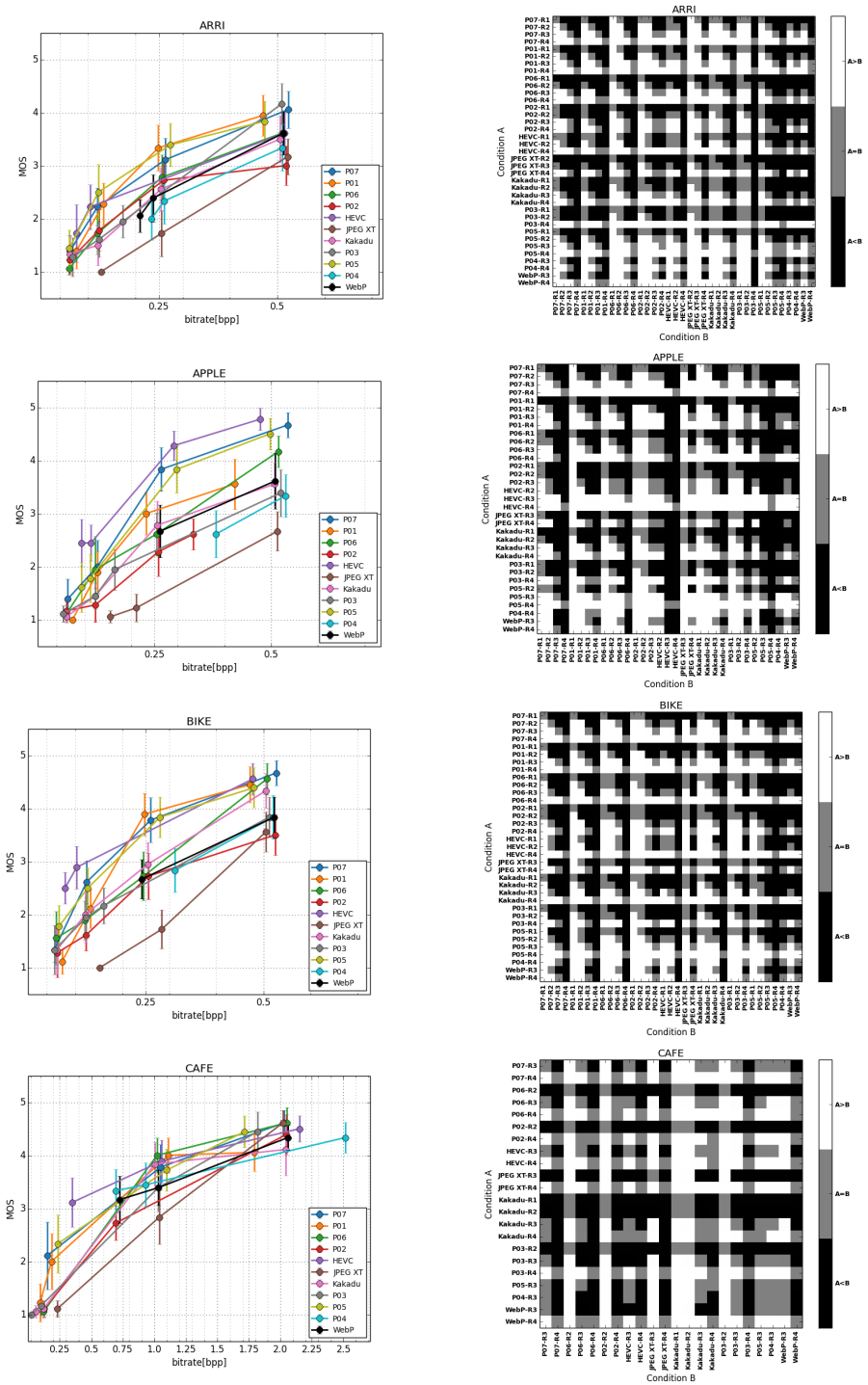


Figure 4.6 – Subjective results for the SDR contents Arri, Apple, Bike and Cafe from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

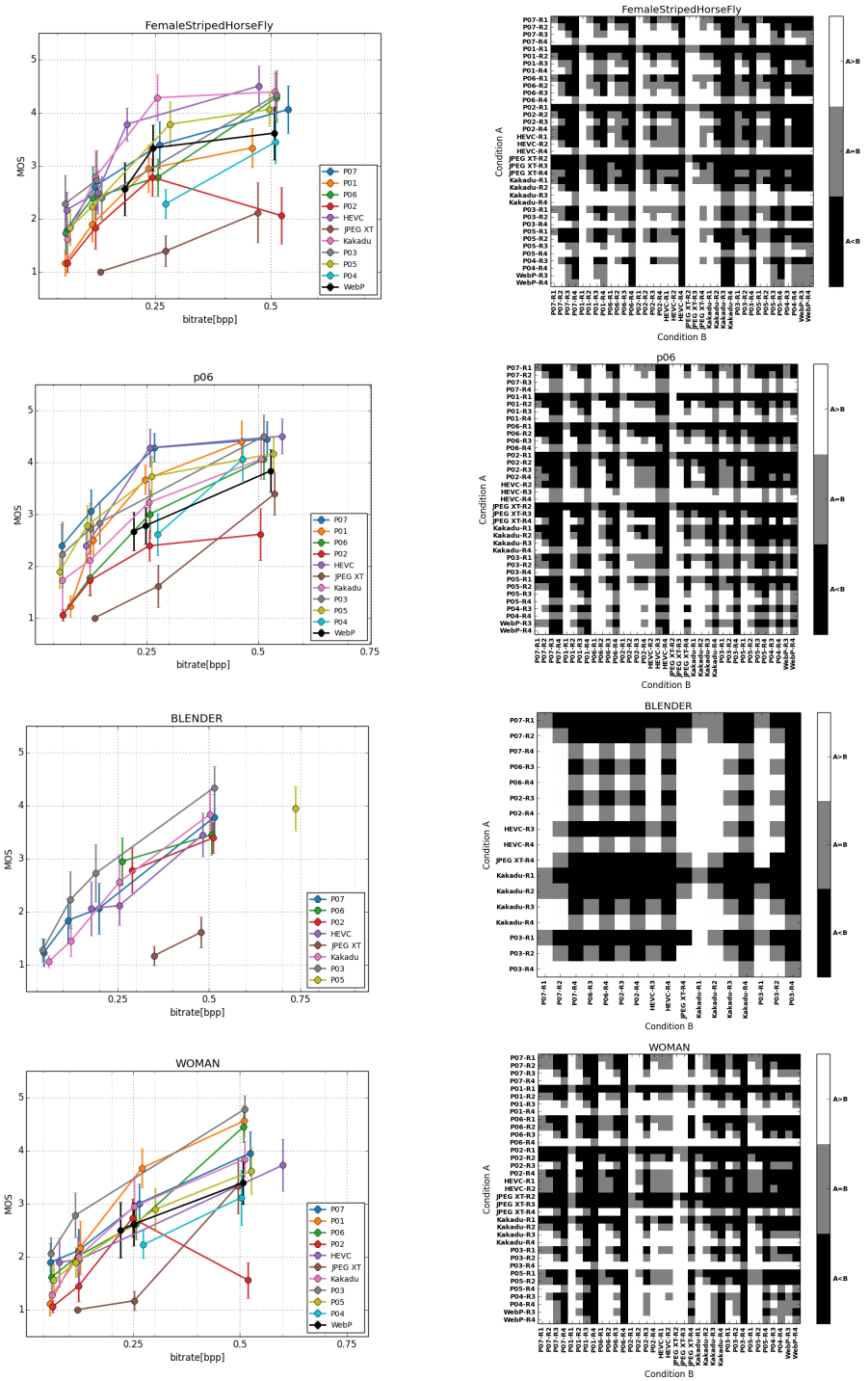


Figure 4.7 – Subjective results for the SDR contents Fly, p06, Blender and Woman from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

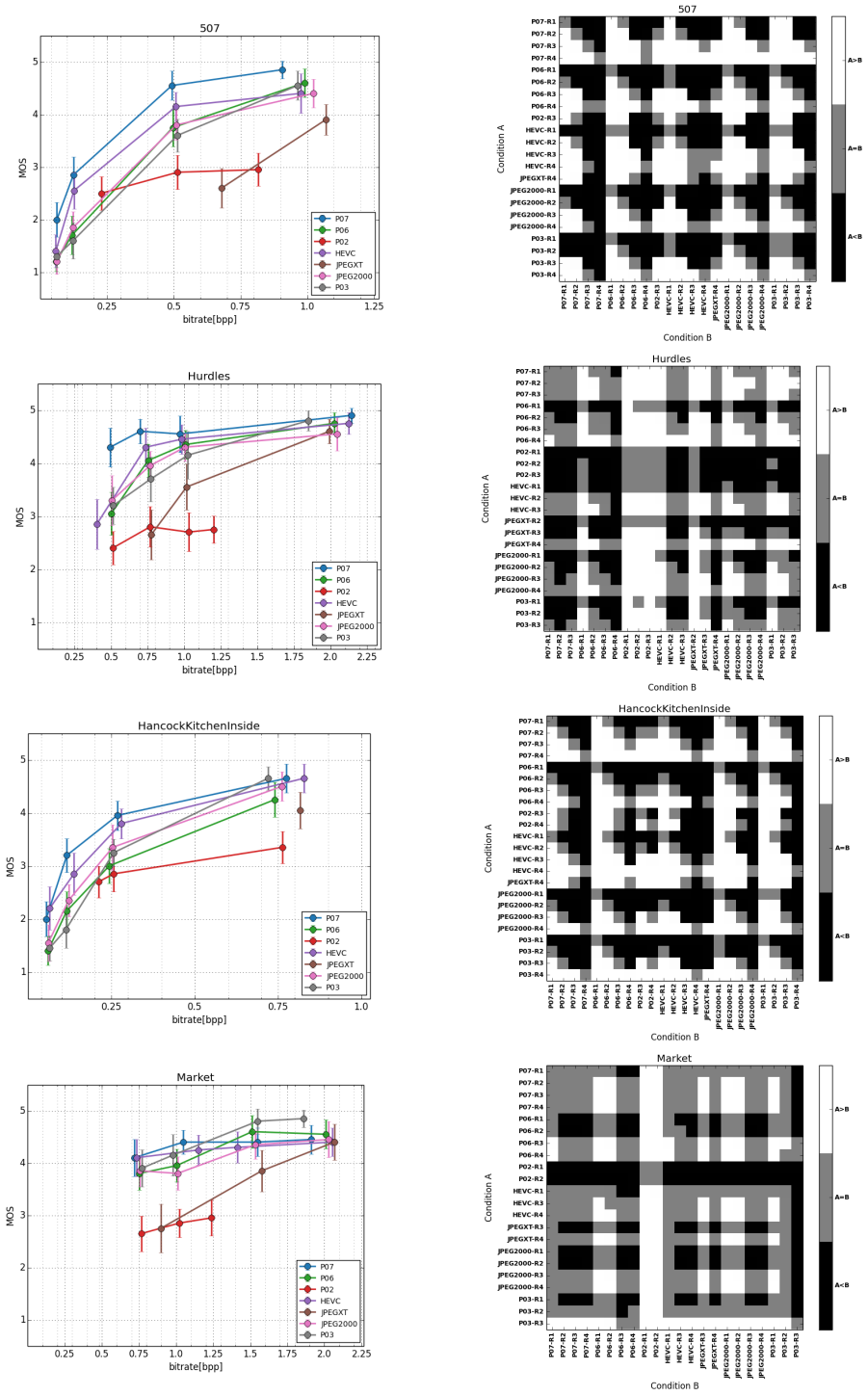


Figure 4.8 – Subjective results for the HDR contents 507, Hurdles, Kitchen and Market from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

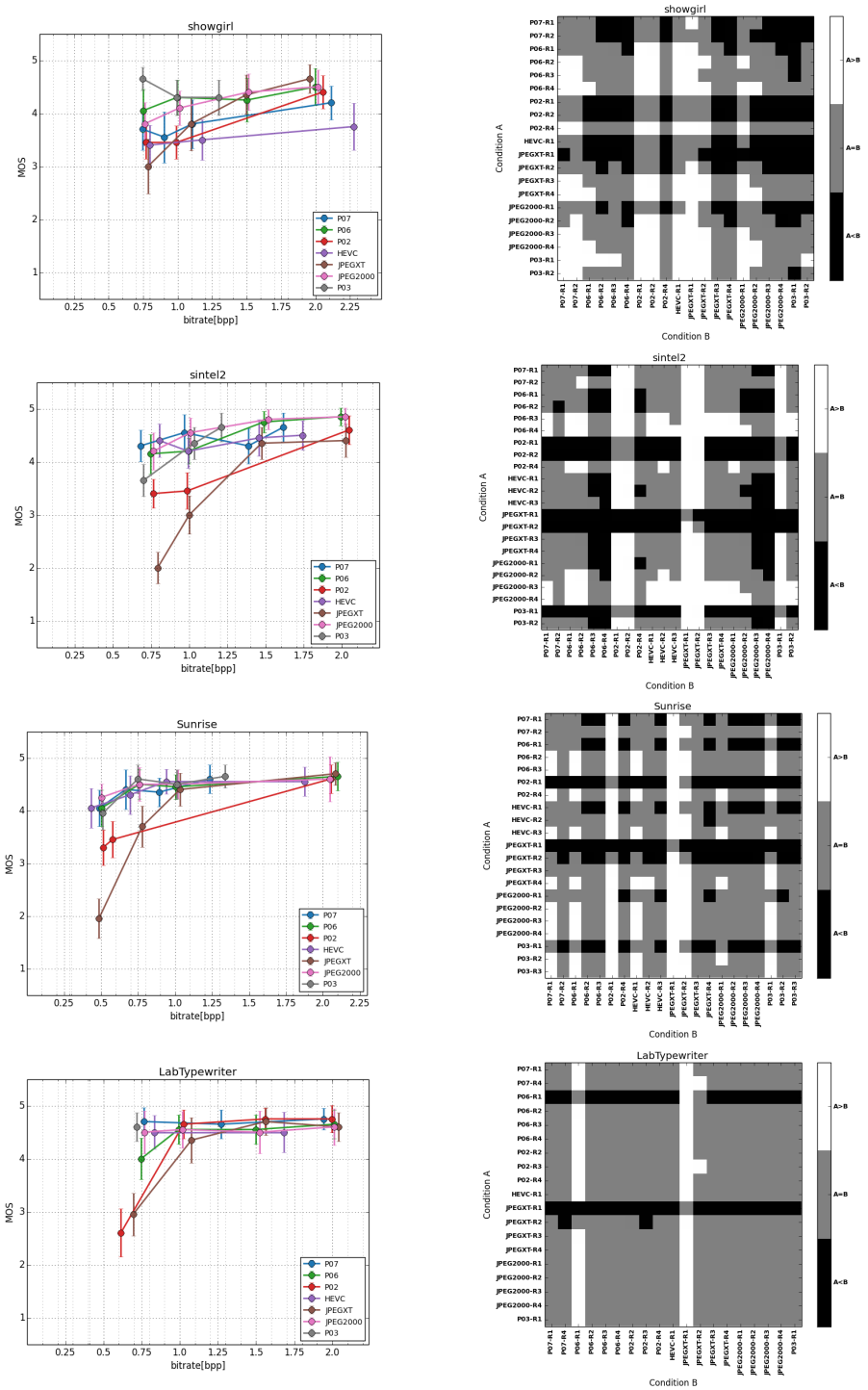


Figure 4.9 – Subjective results for the HDR contents Showgirl, Sintel, Sunrise and Typewriter from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

JPEG and P02 were inferior to other codecs at the lower end of the rate spectrum for most HDR images. Contents Market, Showgirl, Sintel2, Sunrise and Typewriter did not provide statistically significant differences between the performances of other codecs. MOS for 507, Hurdles and Kitchen had more variances along the rate spectrum. P03, P06 and P07 reached transparent quality at the highest bitrate, with P07 having transparent quality at the next lowest bitrate for contents 507 and Kitchen, and remaining at transparent quality at all bitrates for Hurdles. Interestingly, P03's performance at the lowest bitrate for content Showgirl was never inferior to any other codec at any bitrate. These different behaviors indicate the strengths and weaknesses of codecs at certain types of images and regions.

4.2.3 Correlation between results from different labs

It is important to establish the accuracy of the results of the experiments. Objective and subjective tests were therefore conducted in different labs and results were cross checked. Correlation between the results of subjective quality assessment tests of SDR data ran at EPFL and VUB are given in Table 4.6 and Figure 4.10a. Correlation between the results of subjective quality assessment tests of HDR data ran at EPFL, VUB and TPT are given in Table 4.5 and Figures 4.10b-4.10d. The results were compared using multiple metrics such as PLCC, SROCC, Root Mean Squared Error (RMSE), outlier ratio, correct estimation rate, under and over estimation rates, correct decision rate, false ranking, false differentiation and false tie rates. The minimum correct decision rate was 88.15% with PLCC and SROCC going up to 97.68% and 97.75%, respectively.

Table 4.5 – Comparison of subjective quality assessment results for SDR data, gathered by EPFL and VUB.

Key	**Value (%)**	
	Linear fitting	Cubic fitting
PLCC	97.31	97.75
SROCC	97.75	97.75
RMSE	25.25	23.44
Outlier ratio	3.53	2.94
Correct estimation	100.0	100.0
Under estimation	0.00	0.00
Over estimation	0.00	0.00
Correct decision	89.05	90.16
False ranking	0.00	0.00
False differentiation	5.66	5.86
False tie	5.30	3.99

Table 4.6 – Comparison of subjective quality assessment results for SDR data, gathered by EPFL, VUB and TPT.

Key	Value (%)					
	Linear fitting			Cubic fitting		
	EPFL vs. VUB	EPFL vs. TPT	VUB vs. TPT	EPFL vs. VUB	EPFL vs. TPT	VUB vs. TPT
PLCC	97.28	96.43	97.68	98.12	97.15	97.72
SROCC	92.61	91.44	89.46	92.61	91.44	89.46
RMSE	28.92	33.06	23.91	24.08	29.60	23.69
Outlier ratio	5.80	6.25	4.02	5.80	5.80	4.02
Correct estimation	100.0	100.0	100.0	100.0	100.0	100.0
Under estimation	0.00	0.00	0.00	0.00	0.00	0.00
Over estimation	0.00	0.00	0.00	0.00	0.00	0.00
Correct decision	92.10	88.15	90.93	93.17	90.71	90.94
False ranking	0.00	0.00	0.00	0.00	0.00	0.00
False differentiation	2.62	1.74	1.69	3.46	2.18	1.65
False tie	5.29	10.11	7.38	3.38	7.11	7.42

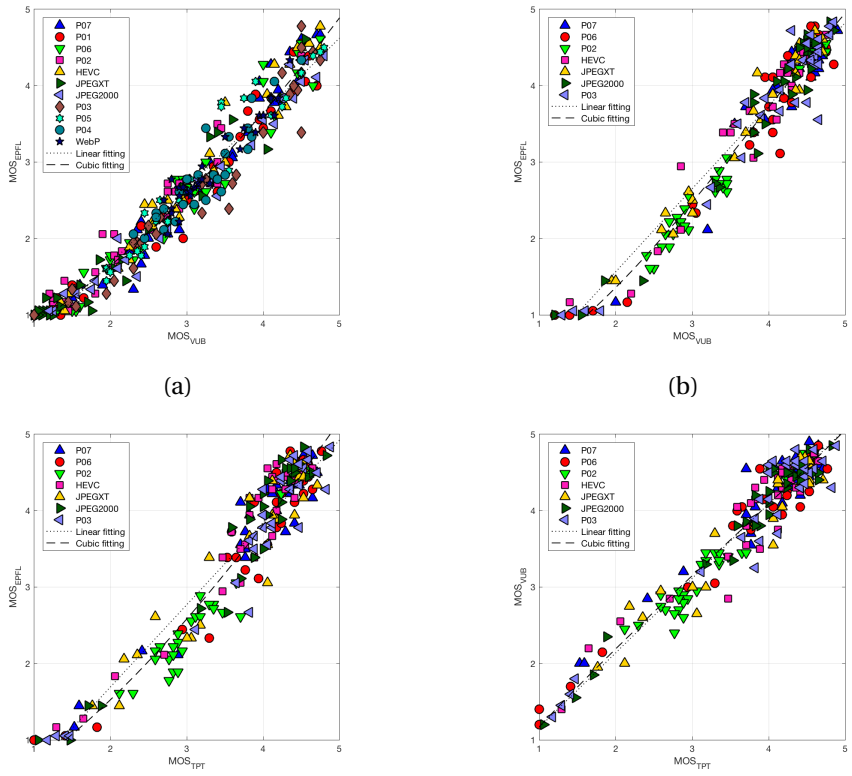


Figure 4.10 – Comparison of subjective quality assessment results gathered by EPFL, VUB and TPT. (a) SDR results from EPFL and VUB, (b) HDR results from EPFL and VUB, (c) EPFL and TPT and (d) VUB and TPT.

4.3 JPEG AI database construction

JPEG Ad Hoc Group on Learning-based Image Coding (JPEG AI) is a new initiative that was formed during the 82nd JPEG meeting in Lisbon, Portugal, and is currently functioning as an Ad-Hoc Group within JPEG. The primary objective of JPEG AI is to evaluate state of-the-art image coding solutions that use machine learning at any or all of the major coding blocks such as feature extraction, quantization, nonlinear transforms and entropy coding. Following the thorough analysis on existing solutions, the group then aims to work towards new coding algorithms and a new learning-based image coding standard. The preliminary action plan of JPEG AI can be listed as follows [Ascenso and Akyazi (2019)]:

1. Analyze and present the state of the art in learning-based image coding in an organized way that involves characterization of the solutions based on, but not limited to
    - network type used in the architecture,
    - target range of quality, i.e. low to transparent,
    - types of rate control used,
    - types of tools and redundancies exploited, i.e. spatial correlation, NSS, intra prediction etc.,
    - the coding unit, i.e. entire image or blocks,
2. Present the open source software implementations of the solutions when applicable,
3. Prepare a comprehensive dataset comprising of uncompressed or pristine quality images to use for training, validating and testing existing solutions,
4. Prepare a Common Test Conditions (CTC) document to carry out reproducible experiments,
5. Conduct an evaluation study to assess the performance of available set of solutions on the prepared dataset,
6. Analyze the results and characterize the compression artifacts.

In the current and next sections, starting with the collection of images for dataset creation, the conduction and evaluation of a subjective quality assessment study is presented, following the current CTC prepared for JPEG AI through past JPEG meetings.

4.3.1 Dataset

Many online resources were browsed with scrutiny to form the complete JPEG AI database, presented in Table 4.7

Table 4.7 – Online databases consulted for preparing the JPEG AI dataset.

Name	Number of images	Type	Resolution	Bit depth
Kodak[a]	25	Uncompressed	SD	8
MCL-JCI[b] [Jin et al. (2016)]	50	Uncompressed	HD	8
Ultra-Eye[c] [Nemoto et al. (2014)]	41	Uncompressed	HD and UHD	8
JPEG XL	67	Uncompressed	HD to UHD	8, 10, 12
CLIC[d]	1017	Uncompressed	HD	8
RAISE[e] [Dang-Nguyen et al. (2015)]	8156	Uncompressed	UHD	12, 14
UCID[Schaefer and Stich (2003)]	1338	Uncompressed	SD	8
Exploration[f] [Ma et al. (2016)]	4744	Pristine	SD	8
DIV2K[g] [Agustsson and Timofte (2017)]	1000	Pristine	HD	8
ImageNet[h] [Deng et al. (2009)]	14M	Distorted	SD	8

[a] http://r0k.us/graphics/kodak/
[b] http://mcl.usc.edu/mcl-jci-dataset/
[c] https://mmspg.epfl.ch/downloads/ultra-eye/
[d] https://www.compression.cc/challenge/
[e] http://loki.disi.unitn.it/RAISE/
[f] https://ece.uwaterloo.ca/ k29ma/exploration/
[g] https://data.vision.ee.ethz.ch/cvl/DIV2K/
[h] http://image-net.org/download

The Kodak image dataset is a small standard test set used in many codec performance evaluations, comprised of SD natural images with urban, nature and people contents. MCL-JCI dataset includes 50 uncompressed images of HD resolution that can be categorized into 10 semantic categories such as people, animals, plants, buildings, water or lake, sky, bridge, transportation vehicles (boats or cars) and indoor. The dataset was initially used for perceived IQA analysis on JPEG compressed contents. The Ultra-Eye dataset was created for eye-tracking experiments with images containing small details with large variations such as close ups with variable background, small objects in large landscapes and city skylines. The images also cover a wide variety of indoor and outdoor scenes with nature, people, animals, historical contents and paintings. JPEG XL dataset, as described previously, was constructed for IQA purposes for compression quality evaluation and involves five distinct classes of images with varying resolutions and bit depths. CLIC dataset was created for the CVPR Workshop and Challenge on Learned Image Compression (2019), with uncompressed natural images of HD resolution to be used for training, validation and testing learning-based compression solutions. RAISE is a real-world image dataset comprised of natural scenes captured in Europe over the course of 4 years, and is primarily designed for the evaluation of digital forgery detection algorithms. The Waterloo Exploration Database, or Exploration in short, was also designed for IQA applications and contains images of people, animals, plants, landscapes, cityscapes, still-life and transportation. UCID was constructed mainly to provide a standard dataset to be used in investigating the effects of image compression on image retrieval performance on Content-Based Image Retrieval (CIBR) systems, and contains mostly natural and urban images. ImageNet contains more than 14 million annotated images within 20k categories for use in visual object recognition research. Such recognition tasks often need to include distorted images, therefore ImageNet database does not have a specification for pristine

quality images. Lastly, the DIV2K database was created for benchmarking example-based single image super-resolution as a complement to existing super-resolution databases, and to increase the content diversity in such datasets.

Besides the aforementioned datasets with well-defined attributes, used for researches and challenges on multiple domains, other public image databases containing images with Creative Commons Zero licenses were also consulted, such as ISO-Republic[7], Pexels[8], Unsplash[9], Pixabay[10], Public Domain Pictures[11], Europeana[12] and Wikimedia Commons[13].

After examination of the available resources with scrutiny, the JPEG AI database was put together with 5264 training, 350 validation and 40 test images. When dividing the images into separate sets, inclusion of well-balanced representatives from various categories such as people, nature scenes, urban, still life were considered. The image resolutions vary from SD to UHD with 8K instances also included, yet very low and very high resolution images are limited in number. Currently, all images in the database are 4:4:4 RGB images in PNG format, and are of SDR with 8-bit depth, which will be extended to include higher bit depths and HDR contents as a next step.

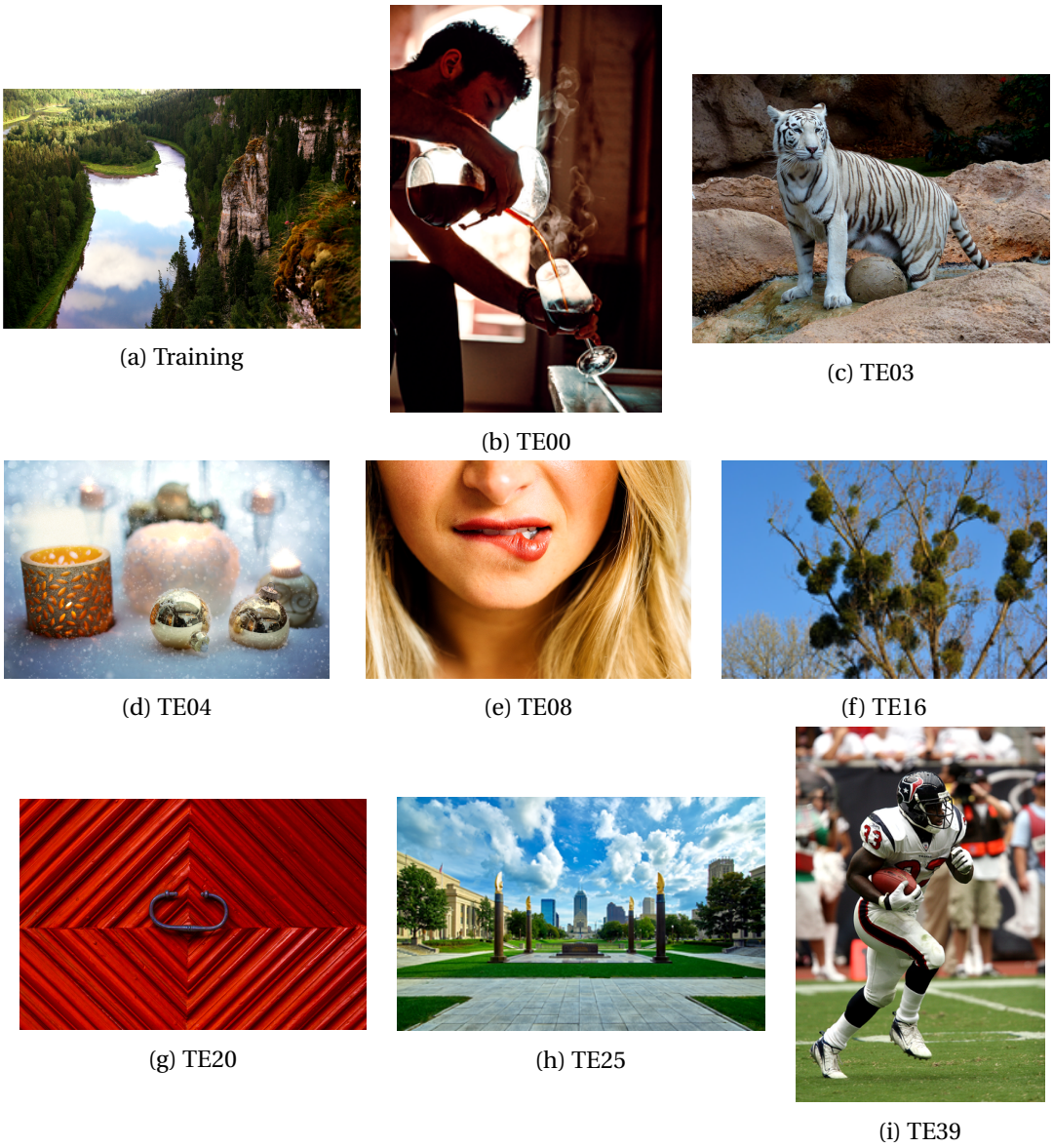
The JPEG AI database was constructed to (i) evaluate the performance of state-of-the-art learning-based image coding solutions and (ii) to be used for training, validation and testing of novel learning-based image coding solutions. As a preliminary step to form an IQA database for objective (i), 8 contents from the test set were selected through expert viewing sessions, to be used for subjective quality assessment experiments. The selected contents are depicted in Figure 4.11.

4.3.2 Encoding images for evaluation

Five learning-based compression algorithms available online were selected for performance assessment against four anchors used during JPEG XL experiments, i.e. JPEG, JPEG 2000, HEVC/H.265 and WebP. The list of tested learning-based image coding solutions is as follows:

- Full Resolution Image Compression with Recurrent Neural Networks (FRICwRNN)[14] [Toderici et al. (2017)]: TensorFlow model for compressing and decompressing images using an already trained Residual GRU model. Although the model is fully convolutional, the input image size needs to be multiples of 32, therefore zero padding was applied when necessary.

[7] https://isorepublic.com/
[8] https://www.pexels.com/royalty-free-images/
[9] https://unsplash.com/public-domain-images
[10] https://pixabay.com/
[11] publicdomainpictures.net
[12] https://www.europeana.eu/portal/en
[13] https://commons.wikimedia.org/wiki/Main_Page
[14] https://github.com/tensorflow/models/tree/master/research/compression/image_encoder



(a) Training (b) TE00 (c) TE03

(d) TE04 (e) TE08 (f) TE16

(g) TE20 (h) TE25 (i) TE39

Figure 4.11 – Thumbnails of JPEG AI contents selected for subjective quality assessment.

- The following models using factorized entropy models (Fact-) or exploiting the dependencies within the latent representation through a scale hyperprior at the encoder (Hyper-), with MSE or MS-SSIM loss as a distortion measure[15] [Ballé et al. (2018)]:
  - Factorized Entropy Model with Mean Squared Error Loss (FactMSE)
  - Factorized Entropy Model with Multi-scale Structural Similarity Index Loss (FactMS-SSIM)
  - Entropy Model with Scale Hyperprior using Mean Squared Error Loss (HyperMSE)

[15] https://github.com/tensorflow/compression

– Entropy Model with Scale Hyperprior using Multi-scale Structural Similarity Index Loss (HyperMS-SSIM)

The reference softwares used for encoding anchors were identical to those used in JPEG XL evaluations, as depicted in Table 4.2. Command lines used for encoding the anchors and learning-based solutions are given in Table 4.8.

Table 4.8 – Selected parameters and settings for anchors and learning-based codecs.

Codec	Input format	Command line
JPEG	YCbCr 4:4:4 8-bit	jpeg -qt 3 -h -v -c -q <qp> -s 1x1,1x1,1x1 <input> <output>
JPEG 2000	YCbCr 4:4:4 8-bit	kdu_v_compress -i <input> -o <output> –rate <bpp> -precise -tolerance 0
HEVC	YCbCr 4:4:4 8-bit	TAppEncoderStatic -c encoder_intra_main_scc.cfg -f 1 -fr 1 -q <qp> -wdt <width> -hgt <height> –InputChromaFormat=<chroma_format> –InternalBitDepth=<bit_depth> –InputBitDepth=<bit_depth> –OutputBitDepth=<bit_depth> –ConformanceWindowMode=1 -i <input> -b <output> -o /dev/null
WebP	YCbCr 4:2:0 8-bit	cwebp -m 6 -q <qp> -s <width> <height> <depth> <input> -o <output>
FRICwRNN	RGB 4:4:4 8-bit	python encoder.py –input_image=<input> –output_codes=<output> –iteration=<qp> –model=residual_gru.pb
FactMSE	RGB 4:4:4 8-bit	python tfci compress bmshj2018-factorized-mse-<qp> <input> <output>
FactMS-SSIM	RGB 4:4:4 8-bit	python tfci compress bmshj2018-factorized-msssim-<qp> <input> <output>
HyperMSE	RGB 4:4:4 8-bit	python tfci compress bmshj2018-hyperprior-mse-<qp> <input> <output>
HyperMS-SSIM	RGB 4:4:4 8-bit	python tfci compress bmshj2018-hyperprior-msssim-<qp> <input> <output>

During the preparation of JPEG AI CTC, YCbCr 4:4:4 color space was preferred to avoid negative bias on anchor results. The learning-based codecs, on the other hand, all operated with RGB 4:4:4 inputs. The color space conversion for JPEG XT was handled inside the codec, so the files were not converted. Conversions from RGB 4:4:4 to YCbCr 4:4:4 for JPEG 2000 and HEVC/H.265-Intra and YCbCr 4:2:0 for WebP were conducted using FFmpeg with the following command, with <pix_fmt> parameter set either to yuvj444p or yuvj420p:

ffmpeg -i <input> -s <width>x<height> -pix_fmt <pix_fmt> <output>

4.3.3 Objective quality assessment

Objective quality assessment was carried out at the bitrates selected for subjective quality evaluation for all codecs, in YCbCr color space. Selected metrics for objective quality assessment were PSNR, SSIM, MS-SSIM, VIF and VMAF. All metrics were computed using FFmpeg with command lines as provided in Table 4.9.

Table 4.9 – Command lines for objective metric computations for JPEG AI experiments.

Metric	Command line
PSNR	ffmpeg -s:v <width>x<height> -i <decoded> -s:v <width>x<height> -i <reference> -lavfi psnr=stats_file=<log_file> -f null -
SSIM, MS-SSIM VIF, VMAF	ffmpeg -s:v <width>x<height> -i <decoded> -s:v <width>x<height> -i <reference> -lavfi libvmaf=ssim=true:ms_ssim=true:log_fmt=json:log_path=<log_file> -f null -

4.3.4 Subjective quality assessment

The subjective quality assessment protocol was identical to that of JPEG XL experiments. The Double Stimulus Impairment Scale (DSIS) Variant I [ITU-R BT.2022 (2012)] was the test methodology selected for subjective quality assessment. The stimulus under assessment and the reference were presented simultaneously to the subject and the subject was then asked to rate the degree of annoyance of the visual distortions in the stimulus under assessment with respect to the reference. The degree of annoyance was divided into five different levels labeled as Very annoying, Annoying, Slightly annoying, Perceptible but not annoying and Imperceptible, corresponding to a quality scale ranging from 1 to 5, respectively.

Content and rate point selection

As was done in the JPEG XL tests, the content and rate selection was carried out during expert viewing sessions prior to setting up the experiments. All contents in the test dataset depicted in Figure 4.11 were encoded using the anchor software at 8 rate points [0.06, 0.12, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00] bpp and the decoded images were viewed by experts. In order to obtain meaningful results from the experiments, the selected rate points needed to span a range that covers very low to high bitrates, corresponding to very low to transparent visual quality. The anchor with the best performance, i.e. HEVC/H.265, was used to select such rate points and the selection was verified using other anchors. Table 4.10 depicts the original resolutions of the test contents and the selected bitrates for subjective evaluation.

Table 4.10 – Original resolutions and selected bitrates for subjective quality assessment of JPEG AI contents.

Name	Resolution	Bitrates
TE00	1486 × 2230	[0.06, 0.12, 0.25, 0.50]
TE03	4000 × 3000	[0.06, 0.12, 0.25, 0.50]
TE04	5000 × 3332	[0.06, 0.12, 0.25, 0.50]
TE08	3400 × 2266	[0.06, 0.12, 1.00, 2.00]
TE16	1280 × 852	[0.06, 0.12, 0.25, 0.50]
TE20	4000 × 2666	[0.06, 0.12, 0.25, 0.50]
TE25	2200 × 1392	[0.06, 0.12, 0.50, 0.75]
TE39	2336 × 3504	[0.06, 0.12, 0.25, 0.50]

4.4 Experiments and results on JPEG AI database

The experiments were conducted in VUB with the participation of 18 volunteering subjects. During data preparation, identical steps to those followed in JPEG XL SDR content evaluation were repeated. The controlled environment was also set as previously defined in Section 4.1.4. A standard outlier detection was performed on all sets of raw scores to remove subjects whose ratings deviated strongly from others [ITU-R BT.500-13 (2012)]. None of the subjects were identified as outliers in the experiments. The MOS and 95% CIs assuming a Student's t-distribution of the scores were computed for each test condition [De Simone et al. (2011)]. To determine and compare the differences among MOS obtained for different codecs and bitrates, a one-sided Welch test at 5% significance level was performed on the scores. Bitrates that deviated more than 20% from the target rates were excluded from statistical significance tests.

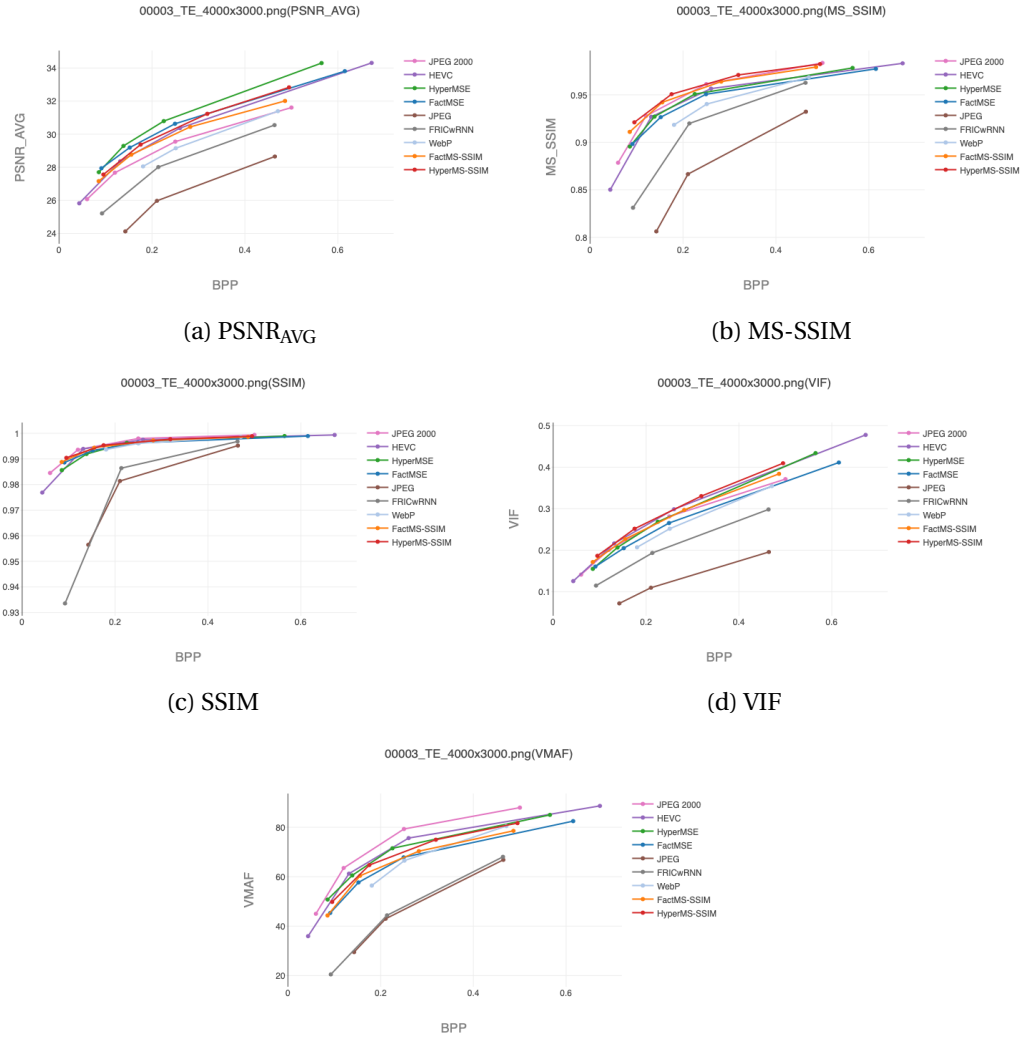
4.4.1 Objective quality assessment results

Objective quality assessment was performed on all 8 contents listed in Table 4.1, at selected bitrates depicted in Table 4.10, for all proponents and anchors and interactive plots were generated. For demonstrative purposes, the objective quality assessment results of the contents TE03 and TE08 are displayed in Figures 4.12 and 4.13.

In Figure 4.12, the leading PSNR codec is HyperMSE, followed by FactMSE, HEVC/H.265 and HyperMS-SSIM performing on par. The leading MS-SSIM codec, however, is HyperMS-SSIM closely followed by FactMS-SSIM and JPEG 2000. VMAF performance of JPEG 2000 curve is above all other codecs at all bitrates. Similar behavior is observed on Figure 4.13, with MS-SSIM-optimized codecs performing better in terms of MS-SSIM and MSE-optimized codecs performing well well in terms of PSNR metrics. Moreover, the performance of all codecs except FRICwRNN and JPEG are very close at all bitrates, which indicates that learning-based codecs are indeed able to reach the performance of their transform-based counterparts in terms of numerous objective metrics.

4.4.2 Subjective quality assessment results

Subjective quality assessment was performed on the selected contents at the screened out bitrates given in Table 4.10, for all proponents and anchors. The MOS vs. bitrate plots and comparisons between pairwise conditions are presented in Figures 4.14-4.15.



(a) $PSNR_{AVG}$

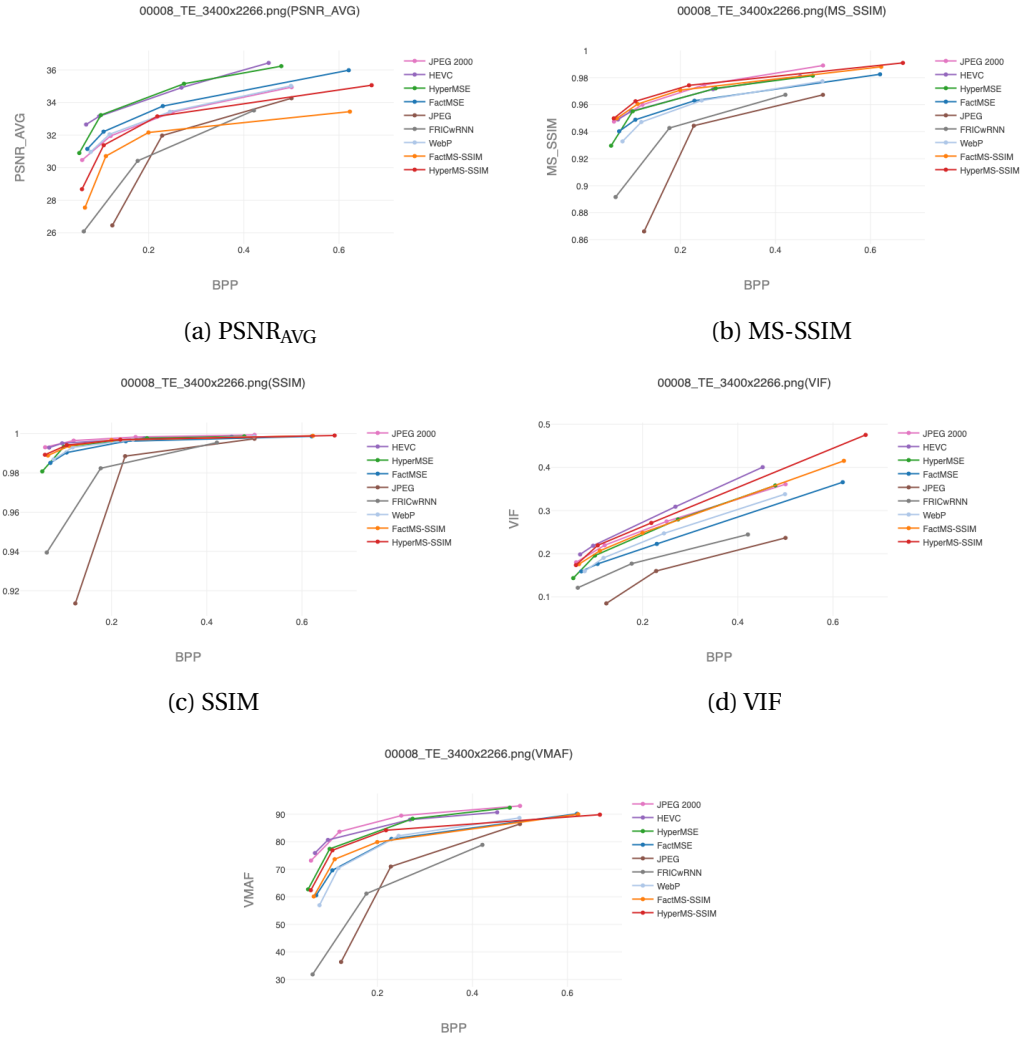
(b) MS-SSIM

(c) SSIM

(d) VIF

(e) VMAF

Figure 4.12 – Objective results for the content TE03.



(a) $PSNR_{AVG}$

(b) MS-SSIM

(c) SSIM

(d) VIF

(e) VMAF

Figure 4.13 – Objective results for the content TE08.

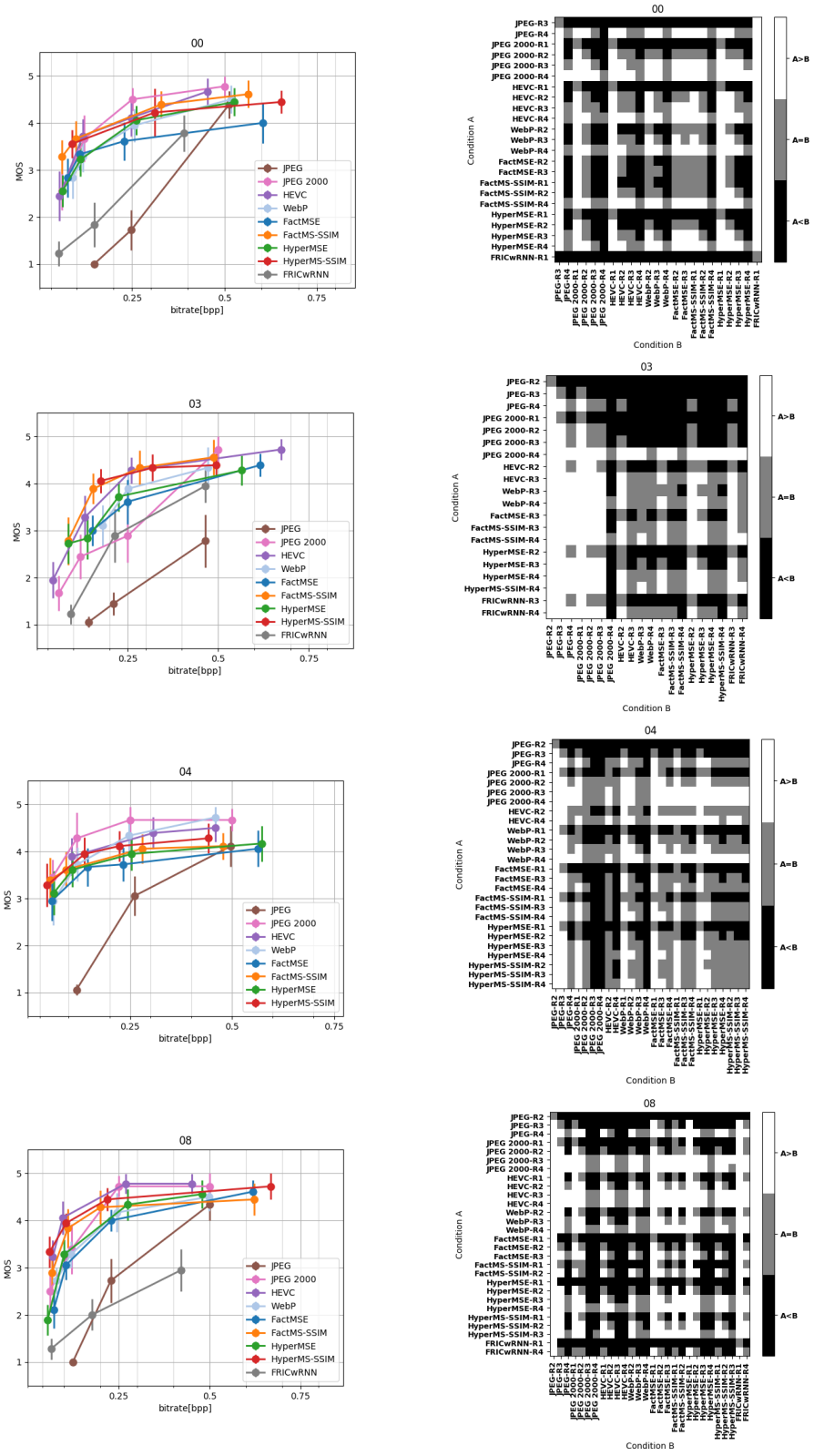


Figure 4.14 – Subjective results for the contents TE00, TE03, TE04 and TE08 from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.

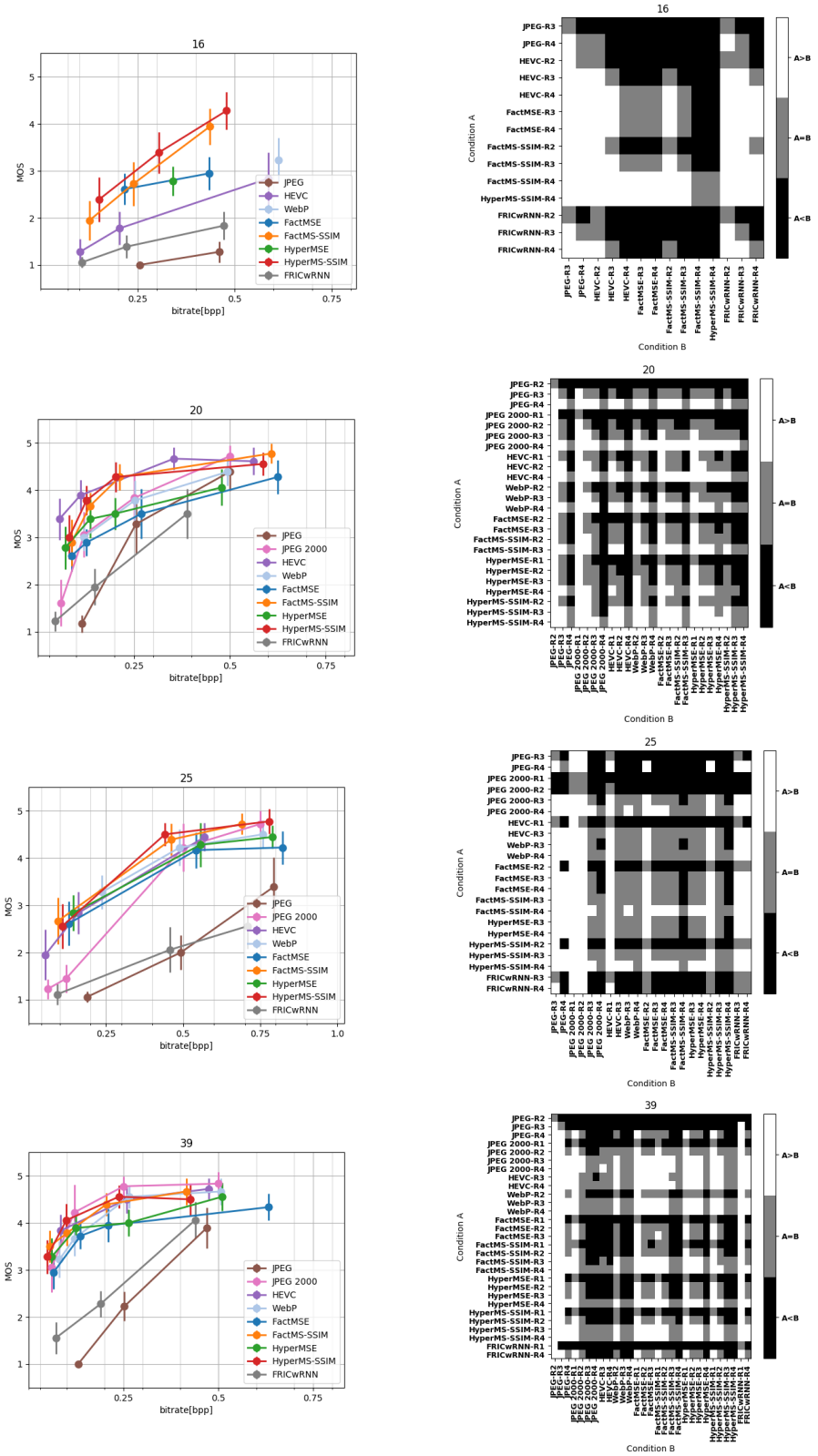
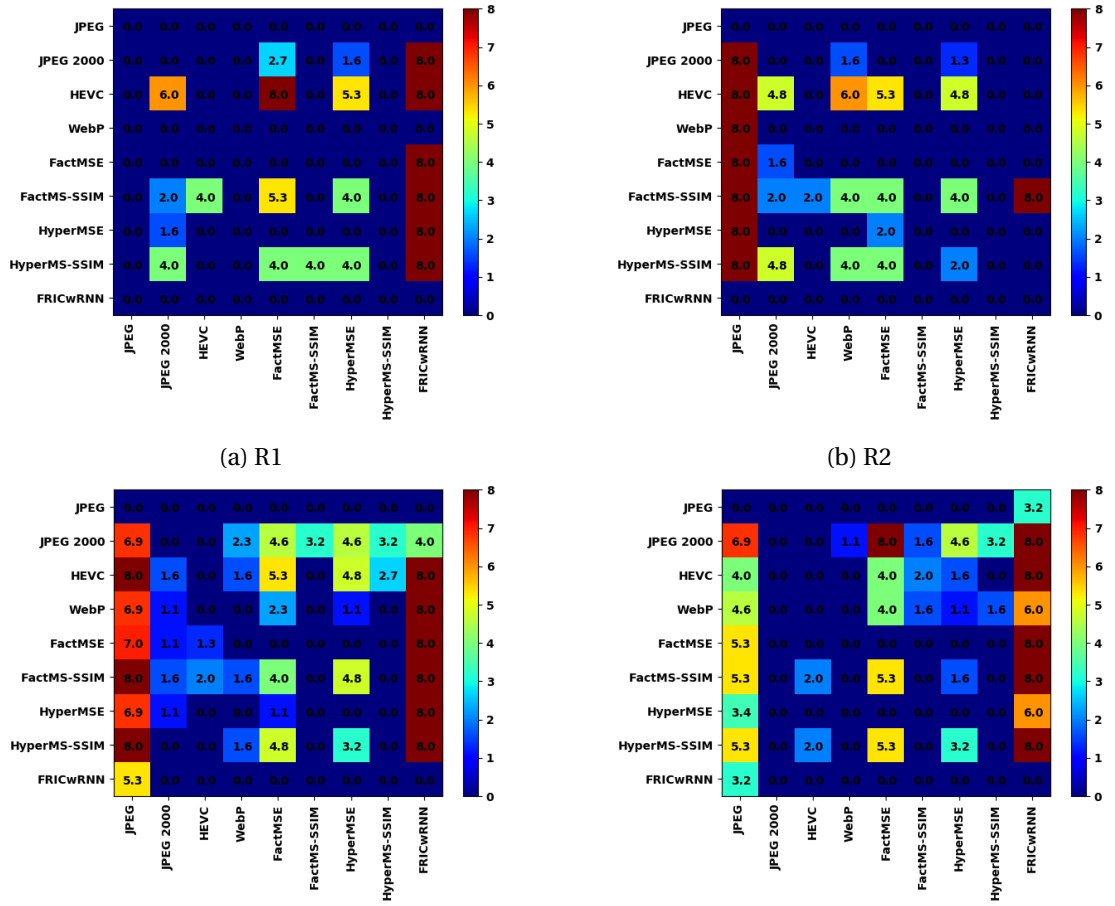


Figure 4.15 – Subjective results for the contents TE16, TE20, TE25 and TE39 from top to bottom. MOS vs. bitrates are presented on the right. Comparisons between pairwise conditions are presented on the left.



(a) R1

(b) R2

(c) R3

(d) R4

Figure 4.16 – Pairwise comparisons between codecs for rates R1, R2, R3 and R4, where R1 is the lowest and R4 is the highest target bitrate. The plots compare how many times codec i was statistically significantly better than codec j at a given bitrate, on a scale out of 8 contents.

Throughout all contents, the performance of JPEG is mostly inferior to all other codecs at the lowest three bitrates. A very similar trend is observed when the performance of FRICwRNN is compared to other codecs except JPEG. The interpolated MOS curves suggest that the subjective ratings of FRICwRNN were higher than JPEG for contents 00, 03 and 39. Comparisons between pairwise conditions presented on the right columns in Figures 4.14 and 4.15 indicate statistically significant differences at comparable rate points. FRICwRNN was superior to JPEG for 2 contents at R3 and R4 namely, TE03 and TE16. Another way of interpreting such a comparison is depicted in Figure 4.16: at target rate point R3, the actual bitrates of FRICwRNN and JPEG were both within the acceptable rate range, i.e. deviated less than 20% from the target bitrate for 3 contents. Out of these 3 contents, FRICwRNN performed statistically significantly better than JPEG twice, which indicates a performance rating of 5.3 out of 8 at R3 in Figure 4.16(c). At the highest target bitrate R4, this ratio fell down to 2 out of 5, indicated as 3.2 out of 8 in Figure 4.16(d).

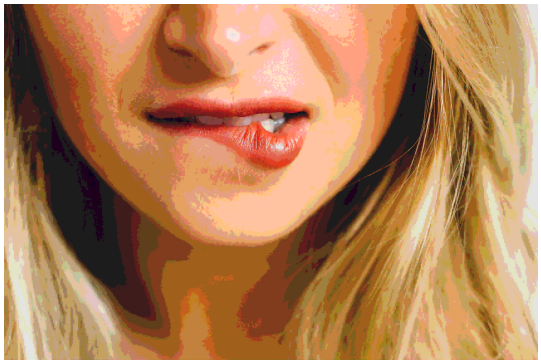
(a) R2 (b) R3 (c) R4 (d) Reference

Figure 4.17 – Section of test content TE16 compressed using HEVC/H.265, FRICwRNN, FactMS-SSIM and HyperMS-SSIM from top to bottom.

The comparison between the rest of the tested learning-based codec performances and anchors is less straightforward. An initial observation is that FactMSE usually performs inferior to the remaining codecs, with exceptions at target rate points R2 and R3 when compared to JPEG 2000, and at target rate point R3 when compared to HEVC/H.265. More specifically, FactMSE was rated statistically significantly higher than JPEG 2000 at R2 for content TE25 and at R3 for content TE03. FactMSE was performing statistically significantly better than HEVC/H.265 at R3 only for content TE16. The results on content TE16 are particularly interesting, indicating an exceptionally low performance on anchors and leading performances of learning-based codecs optimized using MS-SSIM metric. A closer examination on the visuals is provided in Figure 4.17. The abrupt patterns generated by FRICwRNN at R2 are evident, followed by the clear blocking artifacts of HEVC/H.265. HyperMS-SSIM is able to preserve the details better than the other codecs in comparison, with slightly better performance than that of FactMS-SSIM, yet without any statistically significant advantage at comparable rates. It is worth noting that FactMS-SSIM and HyperMS-SSIM are the only codecs for content TE16 that are able to reach transparent quality.

(a) Reference

(b) JPEG 2000, R1



(c) JPEG, R2

(d) FRICwRNN, R1

(e) FactMSE, R1

(f) FactMS-SSIM, R1

(g) HyperMSE, R1

(h) HyperMS-SSIM, R1

Figure 4.18 – Test content TE08 compressed at target bitrate 0.06bpp, using JPEG 2000, JPEG FRICwRNN, FactMSE, FactMS-SSIM, HyperMSE and HyperMS-SSIM.

Contents TE04 and TE39 exhibit similar plots, with codecs except FRICwRNN and JPEG attaining ratings of minimum 3 at all bitrates. TE04 is an image from still-life category while TE39 includes a person, yet both images have focused objects in the foreground and an out-of-focus background. TE08, however, displays a more versatile range of MOS values. All codecs except FRICwRNN reach transparent quality at R4 for TE08, yet codecs FactMSE and HyperMSE have much lower ratings at low bitrates compared to their counterparts optimized using MS-SSIM. The apparent artifacts in TE08 at the lower bitrates are depicted in Figure 4.18.

FRICwRNN possesses artifacts similar to blocks, yet they have a plaid characteristic that is visually more pleasant compared to JPEG. An interesting pattern that is generally observed at low bitrates with codecs optimized using MS-SSIM is the contrast change. The R-D optimization favors preserving structural information at the expense of less fidelity to color components. MSE optimization, on the other hand, is more inclined to introduce blur and introduces only local contrast changes in the form of emphasized colors. In contents like TE20 that exhibit clear structural patterns, MS-SSIM optimized learning-based codecs are performing better than other solutions and are on-par with HEVC/H.265 that is superior to other anchors due to its directional modes. In urban contents like TE25, however, with separate elements of different characteristics scattered all around the image, it is more difficult for both learning-based and transform-based codecs to maintain distinct spatial attributes of the reference throughout the entire scene.

A comparison between the objective and subjective results following the Figures 4.12, 4.13 and 4.14 indicate that both assessments follow similar trends. Just as a codec optimized for MSE metric yields to higher PSNR, the subjective perspective optimized according to HVS and human intelligence yields to higher MOS that is not always matching with each tested metric. Despite these differences, both subjective and objective metrics prove that learning-based codecs proposed in [Ballé et al. (2018)] are performing as well as the state-of-the-art anchors.

4.5 Conclusion

This chapter presented the creation of two IQA databases composed only of compression artifacts, following the framework and results of the quality assessment of proponents submitted to the JPEG XL Call for Proposals for creating the next generation image coding standard, and the JPEG AI experiments for testing the performance of state of the art learning-based image coding solutions. The contributions in this chapter can be summarized as follows:

- During JPEG XL assessment, a total of seven proponents were compared, also with four anchors, at eight different bitrates during objective and four different bitrates during subjective evaluation. Subjective tests were run at three different labs to cross-check the accuracy of the experiments. The performance of some proponents were observed to be as good as or exceeding state-of-the-art codecs for numerous SDR and HDR

contents. The quality assessment tests have lead to the creation of novel SDR and HDR IQA databases, comprised of HD (to UHD for SDR) resolution references and distorted images with compression artifacts from state-of-the-art anchors and proponents, as well as the selection of two proponents, which were then combined to generate the current version of JPEG XL codec. Through objective and subjective test results, it was established that the Call was able to gather solutions superior to the current JPEG standard.

- JPEG AI image quality assessment experiments evaluated the performance of five learning-based image coding solutions against four traditional image codecs, on 8 SD to UHD natural images, at four different bitrates. Results indicated that subjective and objective qualities of state-of-the-art learning-based image coding algorithms were comparable to transform-based codecs. Thorough inspection on the visual results revealed typical artifacts encountered in tested learning-based models. A novel IQA database was created comprised of SD to UHD resolution references and distorted images with compression artifacts from state-of-the-art transform-based and learning-based codecs.

- Owing to the similarity in the methodologies, references and distortion levels in JPEG XL and JPEG AI tests, it is safer to assume a linear mapping between the MOS ratings of the two datasets for 8-bit SDR contents. Thus, the two 8-bit SDR datasets can be combined into a single IQA database with 15 references of resolutions ranging from SD to UHD, and a total of 571 distorted images with distortions due to compression artifacts using 16 different codecs, with 6 of them learning-based. To the author's best knowledge, this is the first IQA database on SDR images that spans such large variety of state-of-the-art compression algorithms on high resolution data, which can be openly used for standardization purposes.

Future work defined in JPEG AI CTC document suggests carrying out SS tests that are expected to reveal different characteristics of the learning-based solutions in the absence of reference images. For example, the contrast changes in MS-SSIM-optimized codecs may be perceived less as artifacts when not presented side-by-side with the references. Similarly, a variant of DSIS test that measures the level of "naturalness" of learning-based solutions perceived by subjects is proposed. Using the same DSIS methodology and changing the rating scale by asking the subjects to rate the naturalness of the images is expected to provide insight into the integration of learning-based features into human vision. These experiments are to involve state-of-the-art codecs, as well as solutions being currently developed as listed in Chapter 2, to present up-to-date results and expand the IQA database.

In the next chapter, the new IQA models are presented which are trained using the JPEG XL IQA database and the combined JPEG XL and JPEG AI IQA databases. First, the JPEG XL IQA database is used to train the WIQM model presented in Chapter 3, which had been pre-trained on TID2013 database. Afterwards, a refined architecture is proposed and trained and tested on the full database constructed in this chapter.

5 Wavelet-based image quality metric for assessment of compression quality

The WIQM metric proposed in Chapter 3 proved the advantages of using wavelet decomposition as a preprocessing step for quality evaluation using convolutional neural networks. The analysis of results highlighted a major shortcoming in the state of the art related to quality evaluation, namely, the lack of a high resolution database that is comprised of ample number of reference images compressed using state-of-the-art codecs, and their respective subjective ratings. This shortcoming was addressed in Chapter 4 with the construction of a new IQA database (XLAI) that is composed of 15 reference images compressed using 16 distinct codecs (of which 6 are learning-based) and 4 bitrate levels, adding up to a total of 571 distorted images of resolutions SD to UHD.

In this chapter, the performance of WIQM is improved through the use of XLAI database, to yield an IQM that is able to assess the impact of compression related distortions on image quality. First, the architecture in Chapter 3 is used with WIQM model as the baseline, and the metric is continually trained using JPEG XL database. After the analysis of this first step, the full XLAI database is used to train a modified network to yield a final metric, WXLAI, whose performance surpasses state-of-the-art image quality metrics when measuring quality of compressed images.

5.1 Performance improvement on wavelet-based image quality metric using JPEG XL database

In order to observe the effect of using a higher resolution database on the quality assessment model, a preliminary experiment was conducted using the JPEG XL database on the WIQM architecture. The WIQM model training was continued with 305 reference-distorted image pairs of the JPEG XL database using 5 splits of the data into training, validation and test sets randomly, using 4, 1 and 2 images, respectively. The MOS of JPEG XL was mapped to range $[0,9]$ to match the MOS of TID2013. Training was otherwise same as in Chapter 3, i.e. MSE

loss was used to compute the error between predicted scores and MOS ratings of each batch. One image was used in each batch, from which $N_p = 128$ patches were randomly extracted. Data augmentation was carried out by flipping each image from left to right and selecting an additional $N_p = 128$ patches at random from each of the flipped images. The Adam optimizer was used for batch optimization with the recommended parameters. A decaying learning rate was preferred starting from 10^{-4} with a decay percentage of 10% every 5 epochs. The models were trained for 140 epochs where the loss had converged to stable values, and the final models used for accuracy tests were the models with the least validation error.

5.1.1 Results

Results on JPEG XL Database

Table 5.1 presents the performance comparison of tested objective metrics in terms of the PLCC and SROCC values with respect to the MOS values of each image, averaged over all test images. Figure 5.1 depicts the performance improvement of WIQM-XL over WIQM. The linear fitting shows that WIQM-XL scores and the MOS correlate highly whereas WIQM scores employ a flatter trend [Akyazi and Ebrahimi (2019b)].

Table 5.1 – Performance comparison of objective quality metrics PSNR, SSIM, MS-SSIM, FSIM, WIQM and WIQM-XL in terms of PLCC and SROCC on the JPEG XL test set. The reported results have been averaged over five randomly selected tests, each consisting of 2 reference images and the corresponding distorted images in the JPEG XL database.

IQM	PLCC	SROCC
PSNR	0.7433	0.7132
SSIM	0.7266	0.7763
MS-SSIM	0.7410	**0.8413**
FSIM	0.5809	0.7770
WIQM	0.7358	0.7791
WIQM-XL	**0.7505**	0.7395

Results on TID2013 Database

The performance of WIQM-XL model was also tested on selected images of the TID2013 test set, that had been distorted by compression-related artifacts, i.e. JPEG and JPEG 2000 compressed images. A total of 50 test images from the TID2013 test set, associated with 5 different reference images were included in the tests. Instead of using $N_p = 128$ as was done for the high-resolution JPEG XL test images, N_p was reduced to 32, as was done during the training of the initial WIQM model. Table 5.2 shows that the performance of WIQM-XL model is superior to WIQM on the compressed images test set, which illustrates that WIQM-XL is more well-suited for evaluating the quality of compressed images in both TID2013 and JPEG

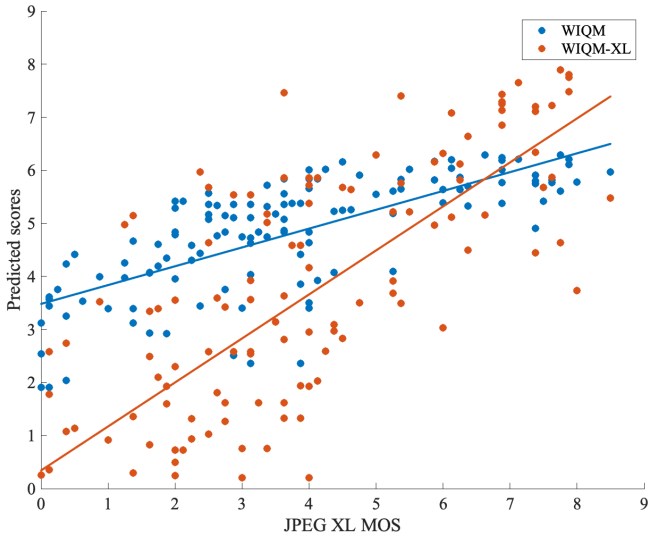


Figure 5.1 – Scatter plot of predicted scores vs. MOS on JPEG XL test set, with linear fitting for WIQM and WIQM-XL models.

XL databases. This improvement is also depicted in Figure 5.2, where the linear fitting on the predicted scores using WIQM-XL correlates more with the underlying MOS compared to WIQM.

Table 5.2 – Performance comparison of objective quality metrics WIQM and WIQM-XL in terms of PLCC and SROCC on the images with compression artifacts in the TID2013 test set. The reported results have been averaged over five randomly selected tests, each consisting of 5 reference images and the corresponding 50 distorted images in the TID2013 database.

IQM	PLCC	SROCC
WIQM	0.8725	0.8743
WIQM-XL	**0.8975**	**0.8954**

5.1.2 Analysis

The results on JPEG XL and TID2013 tests sets indicate that the use of JPEG XL database to tune the WIQM model yielded a metric performance that is superior in predicting the quality of images distorted by JPEG and JPEG 2000 compression in the TID2013 database. The WIQM-XL model is also able to predict the quality of compressed images in the JPEG XL test set effectively, although the performance of MS-SSIM was better than other tested metrics in terms of SROCC measure.

While the correlation between predicted scores and MOS ratings concerning the compressed

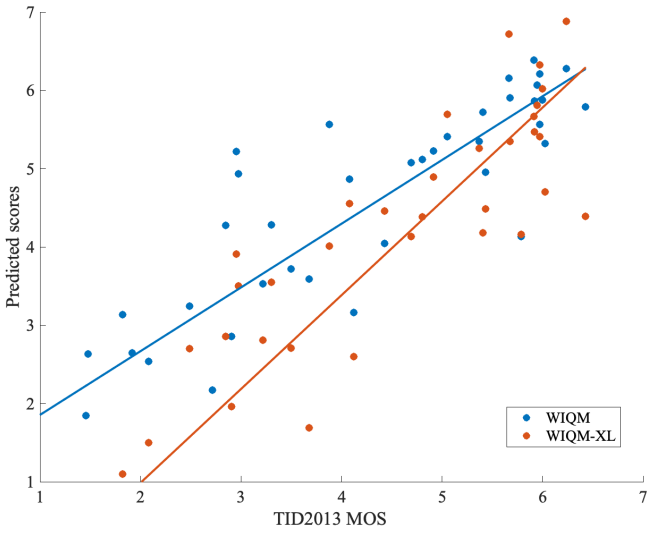


Figure 5.2 – Scatter plot of predicted scores vs. MOS on the images with compression artifacts in the TID2013 test set, with linear fitting for WIQM and WIQM-XL models.

test images in TID2013 database was improved, WIQM-XL model still suffers from the linear mapping between TID2013 and JPEG XL scores. With the resolution of JPEG XL contents roughly five times larger than the TID2013 database, the features extracted from the patches of same dimensions from the two databases are expected to be different. One method to address this issue would be to incorporate multiple resolutions of JPEG XL database into training, however scaling of the data could impair the contents and therefore cause the subjective ratings to become obsolete. Assigning a more accurate mapping of the scores and distortion levels between the two databases would also improve the results. A more practical approach is to use a comprehensive database, such as XLAI, that does not require any mapping and includes an adequate set of images that allow the model to generalize well.

5.2 A wavelet-based image quality metric using convolutional neural networks for assessment of compression quality

Having demonstrated the advantage of using wavelet decomposition for CNN-based IQA models, the final improvement to WIQM follows the preprocessing of the initial model but employs a slightly different architecture at the feature extraction step. Instead of computing the DWT on the grayscale patches and feeding the color image through a separate branch into the network, DWT is computed on the color image channels simultaneously and the architecture is reduced to 3 branches that receive the outputs of the 2D DWT as inputs. The novel feature extraction block is depicted in Figure 5.3. The complexity of the proposed WXLAI model is much lower compared to WIQM, where the number of parameters of the former and the latter are approximately 1.7M and 9.5M, respectively. This is due to the decreased number

of main branches, reduced number of outputs of convolutional layers and removal of residual connections with 1×1 kernels. The overall architecture of the WXLAI network is identical to that of WIQM, as was presented in Figure 3.2, with only the difference feature $f_R - f_D$ included in the feature fusion step for further simplification.

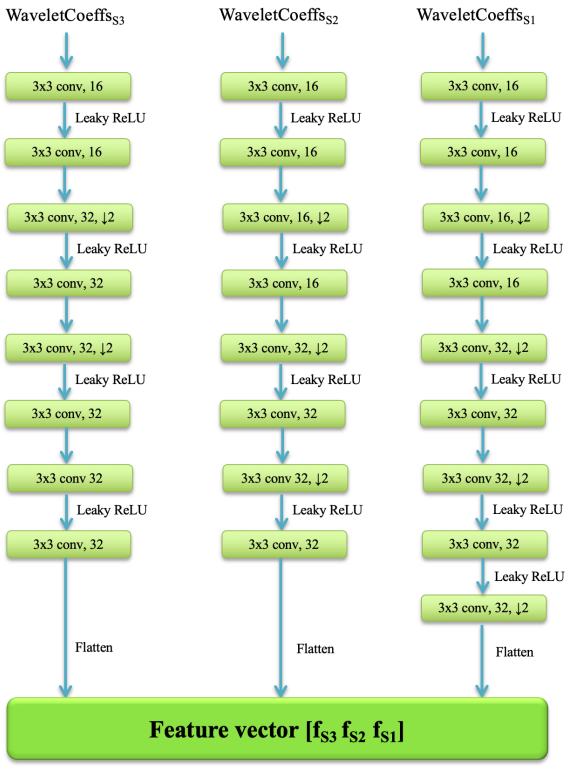


Figure 5.3 – Feature extractor composed of convolutional layers for the WXLAI network. Inputs of the first three branches from left to right are the wavelet coefficients of the 128×128 image patch, where S3 corresponds to the coarsest scale and S1 corresponds to the finest scale. Feature vectors of the three branches are concatenated into a final feature vector of the input image patch.

XLAI database was used for training the model WXLAI, as experiments on WIQM-XL indicated that fine tuning the metric for assessment of compression quality is possible when using such comprehensive training dataset. The proposed model was trained for 200 epochs, with two references out of 15, i.e. TE25 and TE39, separated as test images. MSE loss and Adam optimizer were used with identical hyperparameters as before. Learning rate was kept constant throughout the experiments at 10^4. $N_p = 128$ number of patches of dimensions 128×128 were randomly sampled from each of the 507 training images during each epoch, no data augmentation was used. The patches were normalized to $[0,1]$ range by dividing by 255, instead of a statistical normalization over the training dataset.

5.2.1 Results

The performance of WXLAI model was tested against PSNR, SSIM, MS-SSIM, WaDIQaM-FR and WIQM on both TID2013 and XLAI databases. VIF and VMAF were preferred and added to the tests instead of $FSIM_C$, as $FSIM_C$ had the lowest PLCC levels on WIQM-XL evaluation. The results are presented in Tables 5.3 and 5.4.

Table 5.3 – Performance of WaDIQaM-FR and WXLAI on the TID2013 database, in terms of PLCC and SROCC.

	TID2013			
	Full		**JPEG & JPEG 2000**	
	PLCC	SROCC	PLCC	SROCC
PSNR	0.6667	0.6965	0.9298	0.9436
SSIM	0.7707	0.7592	0.9292	0.9540
MS-SSIM	0.7863	0.7902	0.8973	0.9583
VIF	0.5613	0.5922	0.9097	0.9370
VMAF	0.7791	0.7484	0.9519	0.9527
WaDIQaM-FR	0.9332	0.9239	0.9861	0.9782
WIQM	0.9381	0.9268	0.9663	0.9553
WXLAI	0.7742	0.7541	0.9546	0.9384

Table 5.4 – Performance of WaDIQaM-FR and WXLAI on the XLAI database, in terms of PLCC and SROCC.

	XLAI			
	Full		**Test**	
	PLCC	SROCC	PLCC	SROCC
PSNR	0.6580	0.6594	0.6438	0.6409
SSIM	0.5696	0.7112	0.8086	0.8163
MS-SSIM	0.7501	0.8153	0.8916	0.9149
VIF	0.6486	0.6637	0.7482	0.7992
VMAF	0.7149	0.6951	0.7390	0.7254
WaDIQaM-FR(32)	0.5842	0.6022	0.4244	0.5420
WaDIQaM-FR(128)	0.5799	0.5973	0.4078	0.5356
WIQM(32)	0.5359	0.5856	0.6652	0.5632
WIQM(128)	0.6299	0.6036	0.6764	0.5747
WXLAI	0.8470	0.8437	0.8616	0.7946

5.2.2 Analysis

Table 5.3 shows that the performance of WIQM and WaDIQaM-FR surpass all other metrics on the full TID2013 database, with little difference in their respective performance. This result, however, is expected, as the metrics had been trained on the same database and are able

to rate the characteristic distortion types and levels robustly. The performance of WXLAI is comparable to all other metrics except PSNR and VIF, which are inferior. Performances of the metrics on measuring the quality of compressed images in TID2013 are much higher compared to the full database. Again, WIQM and WaDIQaM-FR achieve the best PLCC and SROCC among all metrics. WXLAI follows WaDIQaM-FR with a PLCC of 95.46%. The difference between the correlations of WXLAI, WIQM and WaDIQaM-FR are small considering that WXLAI was not trained using any instances from TID2013 dataset. The results indicate that the generalization ability of WXLAI metric is substantially high on cross-database evaluation. WXLAI performs almost as good as a metric trained on the test database. Moreover, despite having been trained on a database with distortions caused by 16 different codecs, WXLAI is able to reach the performance of state-of-the-art metrics on a database containing 24 types of distortions, out of which only 2 exemplify compression artifacts.

Table 5.4 presents the performances of tested metrics on the full XLAI database and on the 64 test images reserved for testing. The performance of WXLAI on the full dataset was also included, though it should not be considered as an unbiased metric as the results overlap with training performance. On the XLAI test set, however, WXLAI outperforms WIQM and WaDIQaM-FR, with the latter having surprisingly lower performance both in terms of PLCC and SROCC.

During training, WXLAI extracted 128 patches of size 128×128 from each reference and distorted image, while WaDIQaM-FR and WIQM were trained to extract 32 random patches of dimensions 32×32 and 128×128, respectively. It was stated in Bosse et al. (2018) that increasing the number of patches further did not provide significant advantages. At $N_p = 32$, the performance of WaDIQaM-FR stabilized on both TID2013 and LIVE databases. Considering that the resolution of the images within either database is much lower than XLAI, WaDIQaM-FR and WIQM were also tested using $N_p = 128$. While the prediction accuracy of WIQM improved, a decline in PLCC and SROCC was observed on both full XLAI database and the test images for WaDIQaM-FR. Such undesirable effects highlight the relative importance of the resolutions of training images and the patch size compared to the number of patches used to compute the overall score. WXLAI uses high resolution images during training and extracts higher resolution patches from each image. Although the training image resolutions of TID2013 and XLAI databases are quite different, WXLAI is able to perform well on both databases owing to:

- the increased patch dimensions that are able to capture a larger visual field in images, which is an advantage for predicting the quality of higher resolution images,
- the wavelet decomposition step that helps to analyze visual components within the selected patches on multiple scales, which results in good performance on lower resolution test images.

Table 5.4 indicates that the performances of SSIM and MS-SSIM on the test set are high, with MS-SSIM surpassing XLAI by approximately 3% both for PLCC and SROCC. A likely

interpretation of this result points to the efficacy of multiscale methods on image quality assessment. While SSIM outperforms WXLAI in terms of SROCC measure by 2.17%, the ranking is reversed in favor of WXLAI when PLCC is considered, with a difference of almost 6%.

In order to improve the performance of WXLAI further, one immediate action is to increase the resolution of training patches within the same framework. Practices such as deeper networks and skip connections are expected to result in better quality prediction ability. Increased patch size and network depth, however, bring additional computational complexity. Another important step is to exploit the correlations within the feature vector by processing it through additional layers before regression.

5.3 Conclusion

In this chapter, improvements to WIQM metric were presented in two steps. First, the WIQM model was trained using additional data from JPEG XL database, which is composed of higher resolution contents that have been compressed using numerous algorithms. The WIQM-XL model showed improved performance compared to WIQM on the JPEG XL test contents in terms of PLCC, yet SROCC of WIQM was still higher than WIQM-XL. On the other hand, WIQM-XL was predicting the quality of compressed images in TID2013 more accurately, as was expected with the inclusion of an increased number of compression artifacts.

A linear fitting on the scatter plots of predicted scores versus MOS ratings depicted that WIQM-XL was able to assess the quality of JPEG XL test images effectively. However, since the model was initially trained on TID2013 that has different types and levels of distortion, as well as a wider rating scale, the compatibility of WIQM-XL with the JPEG XL database still needed improvement.

The third model was trained on the full JPEG XL and AI database, referred to as XLAI throughout this chapter, which was composed of an increased number of references compared to JPEG XL and therefore provided enough samples for training. Moreover, the WIQM architecture was simplified by reducing the parameters to almost 6 times less, yet keeping the wavelet decomposition as a preprocessing step. The performance of the third model, WXLAI, was superior to WaDIQaM-FR and WIQM when assessing the quality of compressed images in a full reference framework on XLAI test. Moreover, WXLAI demonstrated superior generalization ability on the cross-database evaluation performed using the compression subset of TID2013.

The contributions of this chapter can be restated as follows:

- The effect of training sets on proposed learning-based IQA networks was explored. Metrics trained on low resolution databases with numerous distortion types and levels were compared to metrics trained on databases involving higher resolution images and artifacts from numerous distinct compression algorithms. It was observed that the latter

performed superior in assessing the quality of compressed images with high resolution, with sufficient generalization ability on experimented test sets.

- The advantage of employing wavelet decomposition as a preprocessing step was demonstrated through the test performances of investigated metrics on HD images, while having been trained on SD databases. Results indicated that such preprocessing yielded higher correlation with subjective ratings on cross-database tests, despite involving different resolutions, distortion types and levels than training data.
- A learning-based metric that is able to assess the quality of images of various resolutions from SD to UHD, compressed using both learning-based and transform-based compression methodologies at different rate points, was established.

Future directions for improving the proposed WXLAI metric point at exploration studies on the network architecture, dependencies within the feature vector, aggregation of features from reference and distorted images and effects of patch dimensions. As was stated in Chapter 3, an exciting direction is to extend the metric to temporal dimension. Incorporating temporal correlation into the problem of quality assessment of compressed video is very challenging, yet at the same time crucial for the high resolution multimedia world that we live in today.

Learning-based image compression Part II

6 Low-rate image compression using convolutional autoencoder and wavelet decomposition

Disclaimer: This chapter was adapted from the following article, with permission from all publishing entities:

Pinar Akyazi and Touradj Ebrahimi, "Learning-Based Image Compression using Convolutional Autoencoder and Wavelet Decomposition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

Image compression is a challenging problem that has been drawing the attention of both researchers and multimedia service providers. The main goal of image compression is to deliver as high visual quality as possible while maintaining the bitrates reasonably low, depending on the system requirements. Image and video processing communities have been proposing different solutions to improve compression efficiency. Traditional image compression algorithms use hand-crafted features and fixed transforms to represent the encoded bitstreams. Recently, the performance of learning-based image compression models have succeeded to reach that of state-of-the-art transform-based approaches, as was highlighted in Chapter 2.

Learning-based methods do not employ hand-crafted features to represent images. Rather, image features are learned from a large set of examples, through extensive training using neural networks. Autoencoders are able to learn approximately invertible mappings from images, which are referred to as the latent representation. The latent representation can then be quantized and entropy coded in order to further reduce storage and transmission costs to a necessary level. In an end-to-end trained system, parameters of the decoder are optimized together with those of the encoder, using a global loss function that spans the whole model.

In Chapter 2, state-of-the-art learning-based compression methods employing various neural network architectures such as CNNs, Recurrent Neural Network (RNN)s and Generative Adversarial Network (GAN)s were presented. The neural networks were used for several tasks within hybrid compression algorithms and autoencoders, including feature extraction, quantization and entropy coding. With many novel approaches being brought forward continuously,

numerous initiatives are being organized in parallel to gather the proposed solutions and advance the field of learning-based image compression. The work of JPEG AI in creating a taxonomy to classify the existing solutions, gathering the state of the art and testing the performance of the proposed methods subjectively and objectively was presented in Chapter 4. Another recent initiative was the Workshop and Challenge on Learned Image Compression (CLIC) organized by the Conference on Computer Vision and Pattern Recognition (CVPR) in 2019. As was the case for the 2018 Workshop, the aim was to encourage research and industry communities in developing novel encoder/decoder architectures, novel ways to control information flow between the encoder and the decoder, and learn how to quantize and process the latent representation better.

Two main tracks were proposed in the challenge: the low-rate track called for solutions that are able to operate with high quality on an average of 0.15bpp on the CLIC test set, whereas the transparent track targeted a minimum 40dB aggregated PSNR or a minimum of 0.993 aggregated MS-SSIM on the same set while trying to keep the rate as low as possible. In this chapter, a convolutional autoencoder (CAE) is presented, that was optimized for the low-rate track. Inspired by the use of multiscale approaches in learning-based image compression [Wang et al. (2019); Rippel and Bourdev (2017); Akbari et al. (2019)] and in transform based image compression (e.g. JPEG 2000) as well as the performances of WIQM and WXLAI presented in Chapters 3 and 5, a preprocessing step involving a 3-scale wavelet decomposition of all input channels was used. A latent representation of the image was obtained during training using convolutional layers and GDN nonlinearities, which was further quantized and entropy coded. The proposed codec, WCAE, was trained end-to-end, using MSE in the loss function. The performance of the proposed method was tested on the CLIC2019 validation and test sets, and compared to the performance of JPEG and JPEG 2000, as well as a similar neural network that does not use wavelet decomposition.

The proposed method and network architecture are described in details in the next section, followed by the presentation and the analysis of the results, and the conclusion of the chapter.

6.1 WCAE framework

The proposed architecture is depicted in Figure 6.1 with the analysis and synthesis blocks shown in details on Figure 6.2. The input color image X was separated into non-overlapping patches of dimensions $N \times M$. Before the analysis stage of the convolutional autoencoder, each color channel of an RGB input image patch was first normalized to have $[-1,1]$ range and then underwent a 3-scale 2D wavelet transform, where Daubechies-1 wavelets were used.

2D wavelet decomposition is known to be effective in various image processing tasks, compression in particular. When compressing an image using convolutional neural networks, although image features are expected to be learned by the network without ideally any preprocessing, using wavelet decomposition has two distinct benefits. First, the image is separated into its high frequency and low frequency components at different scales, allowing more control over

the visual characteristics of the compressed image by giving more or less emphasis on particular frequency components. Second, introducing a wavelet decomposition as a preprocessing step is expected to help the network converge faster.

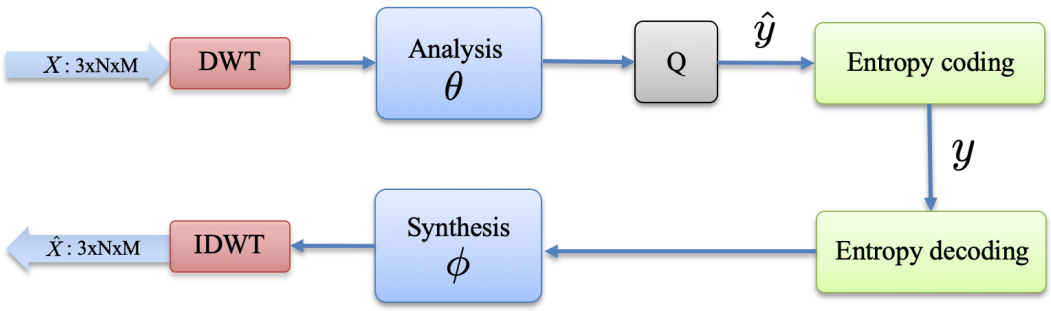


Figure 6.1 – Proposed convolutional autoencoder architecture.

The analysis block was separated into three channels for each scale of the wavelet transform. The coarsest scale had 12 inputs, 4 from each of the 3 color channels. The second and finest scales had 9 inputs each. Convolutional filters of dimensions 3×3 were used at each layer and the number of outputs was doubled once for all scales. The coarsest, second and finest scale inputs were downsampled 2, 3 and 4 times, respectively. Between the convolutional layers, GDN functions provided a nonlinear mapping of the layer outputs. The latent representation was formed by concatenating the 32 outputs of each scale, and had dimensions $32 \times \frac{N \times M}{1024}$. With this representation, the $3 \times N \times M$ input was reduced by a factor of 32, in size.

The latent representation (code) then needed to be quantized. Since quantization is a function with zero gradients almost everywhere, it was replaced by additive uniform noise during training. This is a method preferred at the quantization step of learning-based encoders [Ballé et al. (2016); Cheng et al. (2018)] assuming unit bin size and uncorrelated quantization error between elements. The distribution of the latent representation needs to agree with this assumption for efficient compression, which is expected to be learned by the network during training. When using this design, the bin size of the quantization effect is implicitly regulated through the distribution of the latent representation.

Following quantization, the latent representation can be compressed further using lossless entropy coding. The rate of the quantized and entropy coded latent representation is finally used as a component of the overall loss function. Since the latent representation had been uniformly quantized, an effective entropy coding is expected to reduce the rate optimally. Arithmetic encoders such as range encoder [Theis et al. (2017)] and context-adaptive binary arithmetic coder (CABAC) [Marpe et al. (2003)] are good candidates, however, the entropy coding also needs to be fully differentiable for end-to-end optimization of the CAE. The lower bound of the rate, on the other hand, is equal to the entropy of the quantized code [Shannon (1948)]. It is therefore sufficient to compute the entropy of the quantized code and use it in

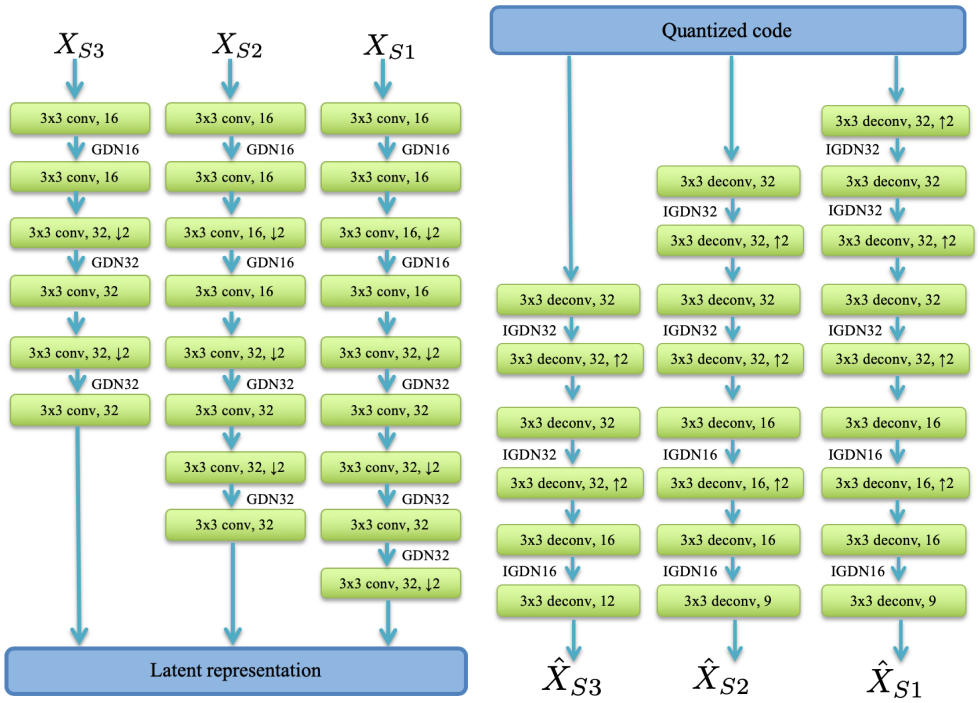


Figure 6.2 – Architecture of the analysis (left) and synthesis (right) blocks. The representation 3x3 conv, [C] depicts a convolutional layer with 3x3 kernels and C outputs. GDN[C] is the generalized divisive normalization function with C inputs. ↑ 2 and ↓ 2 refer to upsampling and downsampling by a factor of 2, respectively. The model has 256606 parameters in total.

the loss function as an estimate of the rate portion.

During this part of the optimization, the network tries to extract meaningful features from the input images such that they can be quantized using unit bin size and then arithmetically coded to reduce rate. The efficiency of this analysis stage is coupled with the synthesis stage, where deconvolutional filters and upsampling operators are used to reverse the compression. In the proposed framework, synthesis mirrored the analysis stage backwards. The quantized code was separated into three equal dimensional components, which represented the coarsest, second and finest scales of the decoded image at the output of the synthesis transform. Inverse GDN function (IGDN) was used between deconvolution layers. At the final layer, the three outputs were used to invert the wavelet transform and yield the decoded image. The distortion between the original and decoded image was measured in terms of MSE and used in the loss function. The overall loss function of WCAE was then:

$$J(\theta, \phi; X) = D(X, \hat{X}) + \lambda R \tag{6.1}$$

where

$$D(X,\hat{X}) = \sum (X-\hat{X})^2 \tag{6.2}$$

$$R = \sum_i P_{\hat{y}}(\hat{y}_i)\log_2 P_{\hat{y}}(\hat{y}_i) \tag{6.3}$$

where X is the input image, $\hat{X}$ is the decoded image, D is the distortion, R is the entropy of the quantized code $\hat{y}$ which has the distribution $P_{\hat{y}}(\hat{y})$. Here, the distribution of $\hat{y}$ is approximated to be Gaussian after the use of multiple GDN nonlinearities and the discrete probability function of $\hat{y}$ is computed as a Gaussian distribution with mean $\mu_{\hat{y}}$ and standard deviation $\sigma_{\hat{y}}$.

After the network was fully trained, the latent representation was quantized by rounding to the nearest integer at test time and entropy encoding was performed by the range encoder [Theis et al. (2017)]. The range encoder expects a positive input, therefore the minimum value of the quantized code was subtracted and then passed to the decoder. In addition, the decoder also needed to receive the input image dimensions, as each input channel was padded with zeros to have dimensions that were multiples of 32. Finally, the cumulative distribution function of the quantized code was also passed to the decoder. The total size of the encoded bitstream was then equal to the sum of these additional parameters sent to the decoder and the output of the range encoder. In order to reach a total rate less than 0.15 bits per pixel (bpp) on the CLIC2019 test dataset, the parameter λ needed to be adjusted during training.

6.2 Experiments and results

In this section, the training of the proposed model is explained in detail, also including a brief description of the dataset. The qualitative results are presented along with visual examples that depict the strengths and the weaknesses of WCAE compared to state-of-the-art transform-based codecs JPEG and JPEG 2000. Another network with similar architecture to that of WCAE which skips wavelet decomposition and uses pixel values as input is also included in the comparison to highlight the effect of the proposed preprocessing method.

6.2.1 Database

For training the network, the mobile and professional training datasets of CLIC2019 were used, which amount up to a total of 585 HD natural images of pristine quality. The validation dataset was comprised of 102 HD images. The test set was released at a later time and was a collection of 330 HD images. During training, the training and validation images were separated into non-overlapping patches of dimensions 256×256. A total of 16750 patches were used for training and 1146 were used for validation. The validation and test results were reported on the whole images. The proposed model needs inputs of dimensions that are multiples of 32, therefore zero padding was employed when necessary. Quality assessment was conducted after cropping the images back to original resolution, however, the rate was reported on the padded images.

6.2.2 Results and discussion

With a selection of $\lambda = 0.0025$, the total rate of CLIC2019 validation set was fixed at 4514814 bytes, which was equal to 0.148bpp on average. The network was trained iteratively using back propagation [LeCun et al. (1998, 2012)] and the Adam [Kingma and Ba (2014)] optimizer with a batch size of 8 and learning rate of 10^{-4} for a total of 100 epochs, where the loss had converged to a stable value. The model with the lowest validation error was selected for testing. The results of the proposed method WCAE were compared to three different codecs, of which two are JPEG and JPEG 2000. To analyze the effects of performing wavelet decomposition in the proposed method, a network with similar architecture was built and trained without the wavelet decomposition. This network is referred to as No-Wavelet Convolutional Autoencoder (NoWCAE), whose analysis and synthesis blocks have the same architecture as that of the finest scale of WCAE. The inputs of NoWCAE were 256×256 image patches during training and the number of output channels that comprised the latent representation were 32. The results in terms of PSNR(dB) and MS-SSIM are given in Table 6.1 and 6.2 on the validation and test sets of CLIC2019, respectively. All results have been averaged on the complete validation and test databases and none of the images in the validation and test sets were used during training updates.

Table 6.1 – PSNR(dB) and MS-SSIM on the validation set for codecs JPEG, JPEG 2000, NoWCAE and WCAE.

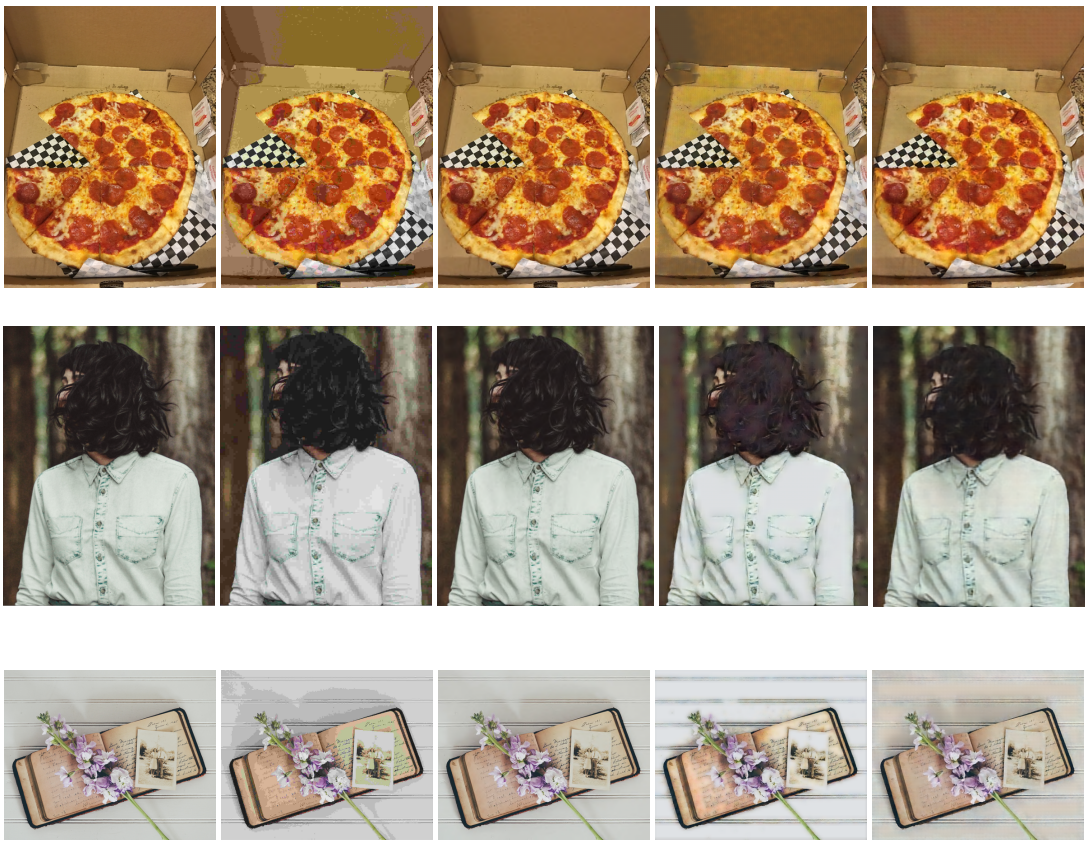
	JPEG	**JPEG 2000**	**NoWCAE**	**WCAE**
PSNR(dB)	30.27	33.29	23.82	25.61
MS-SSIM	0.8208	0.9404	0.8983	0.8965
Rate (bpp)	0.147	0.149	0.206	0.143

The objective results indicate that although WCAE was trained using MSE loss, MS-SSIM values averaged over both validation and test sets are higher compared to those of JPEG. Visual examples from validation and test databases are presented in Figure 6.3, where WCAE clearly has higher subjective quality compared to JPEG and NoWCAE. WCAE also outperformed NoWCAE in terms of PSNR, however the MS-SSIM of NoWCAE is slightly higher than WCAE in the validation database. It must be taken into consideration here that NoWCAE has a higher actual bitrate on both databases, which explains the higher MS-SSIM of NoWCAE on the validation set despite its lower visual quality. Overall, distributions of images compressed with WCAE are more faithful to the distributions of their respective original images. Distinct color changes in Figure 6.3 (d) with respect to Figure 6.3 (e) compared to the originals in Figure 6.3 (a) highlight this advantage of using wavelet transform as a preprocessing step in the proposed network. WCAE has much less low frequency errors compared to JPEG, however the high frequency artifacts are eminent compared to the artifacts in JPEG 2000 results.

WCAE introduces apparent high frequency artifacts and is therefore unable to outperform JPEG 2000 at the targeted rate 0.15bpp neither subjectively nor objectively. In the analysis

Table 6.2 – PSNR(dB) and MSSSIM on the test set for codecs JPEG, JPEG 2000, NoWCAE and WCAE.

	JPEG	JPEG 2000	NoWCAE	WCAE
PSNR(dB)	30.05	32.83	22.25	23.85
MS-SSIM	0.8034	0.9335	0.8743	0.8817
Rate (bpp)	0.149	0.148	0.184	0.138



(a) (b) (c) (d) (e)

Figure 6.3 – Visual examples from the validation and test image datasets (a) and the decoded images of codecs JPEG (b), JPEG2000 (c), NoWCAE (d) and WCAE (e).

.

block, the number of outputs from each wavelet scale was 32. It is possible to attenuate the high frequency artifacts by changing the contribution of outputs from coarse to fine scales. When the outputs of the coarsest scale are doubled to 64 at the fifth convolutional layer, the high frequency artifacts became weaker, however the decoded images had increased low frequency noise. This resulted in lower subjective and objective quality averaged over the validation and test images, depicted on the example in Figure 6.4. Despite the reduced quality, such changes demonstrate how the use of wavelets can be beneficial in order to adjust the frequency characteristics of the output image. Optimization of the contribution from different

(a) (b) (c)

Figure 6.4 – Section of an example image from the validation set (a), WCAE outputs at target bitrate 0.15bpp using 32 outputs at all wavelet scales with PSNR = 31.12dB and MS-SSIM = 0.9264 (b) and 64 outputs instead of 32 at the coarsest scale with PSNR = 29.99dB and MS-SSIM = 0.9093 (c).

scales to the latent representation can adjust the decoded images to have better subjective and objective quality. Optimization of the bit allocation from these scales at the quantization step can increase the quality even further.

6.3 Conclusion

In this chapter, a novel convolutional autoencoder for image compression was proposed, involving a 3-scale 2D wavelet decomposition. The model was trained end-to-end, using MSE in the loss function. Objective results indicate that WCAE outperformed JPEG and NoWCAE, a similar method that excludes the wavelet decomposition step, in terms of MS-SSIM at bitrates lower than 0.15bpp across the full CLIC2019 test set. Visual results of the proposed method also indicate better performance compared to JPEG and NoWCAE. The contributions of this chapter can be summarized as below:

- It was shown that when using the proposed CNN architecture for feature extraction, application of 2D DWT improved the overall compression performance. In general, convolutional neural networks are expected to learn a representation that contains the spatial and frequency information that wavelet coefficients possess and therefore such preprocessing could be redundant. Experimental results, however, indicate that WCAE is able to perform better than its no-wavelet equivalent both objectively and subjectively, at an average bitrate as low as 0.15bpp on the CLIC2019 test set. WCAE also outperforms the legacy JPEG codec subjectively. WCAE does not suffer from blocking artifacts and is able to preserve details better.
- With the help of different wavelet scales, it is possible to achieve different quality trade-offs around the same rate point by changing the number of outputs of scales. When the number of coarse scale outputs were increased, WCAE yielded smoother outputs with slightly less PSNR and MS-SSIM measures.

WCAE was not able to outperform its transform based counterpart JPEG 2000, since it introduced high frequency artifacts that affected the structure and contrast. In the next chapter, improvements to the WCAE architecture are sought using residual blocks and mixed kernel sizes.

7 Wavelet-based convolutional autoencoders with residual blocks and mixed kernels

In the previous chapter, a convolutional autoencoder was proposed, which, with the help of wavelet decomposition as a preprocessing step, achieved low-rate image compression while preserving image details and refraining from blocking artifacts such as the ones posed by DCT. In this chapter, two new end-to-end autoencoders are proposed, as extensions to the WCAE model. The proposed models are referred to as ResWCAE and ResMixWCAE, as the former adds skip connections in WCAE architecture to improve its performance, and the latter uses mixed kernel sizes as an additional modification. Both models were trained at multiple bitrates instead of using a single operating point, by adjusting the rate constraint parameter λ. The performance of the proposed methods were evaluated against JPEG, JPEG 2000 and WebP.

7.1 ResWCAE and ResMixWCAE frameworks

Throughout the evolution of the use of deep convolutional neural networks, it was shown that stacking more layers and thus constructing a deeper network yielded better results in image processing problems such as object detection, image classification [Simonyan and Zisserman (2014); Szegedy et al. (2015); He et al. (2015); Ioffe and Szegedy (2015)] and quality

assessment [Bosse et al. (2018)]. The answer to the question *whether deeper networks always perform better in image related tasks* was investigated in [He et al. (2016)]. It was verified through experiments that beyond a threshold depth, the training error starts increasing and this degradation is not caused by overfitting. The problem was addressed using shortcut connections between intermediate layers, which let the layers fit a residual mapping through the use of identity mapping of the input. These "residual blocks" yielded deeper networks to achieve more accurate results and with higher training efficiency.

The effect of using residual blocks to learning-based image compression was addressed in [Cheng et al. (2019a)]. This work also investigated using different kernel sizes within the same architecture, i.e. using four 3×3, 5×5 and 9×9 dimensional filters in each of the analysis and synthesis stages. It was shown that larger kernel sizes were more effective in providing higher objective quality. When replacing one layer with 9×9 filters with four layers of 3×3 filters to achieve the same receptive field using less parameters, the stack of 3×3 kernels were able to converge using residual connections. The deep residual network provided a computationally efficient alternative to shallower networks with larger kernel sizes.

The answer to the question *whether larger kernel sizes always achieve higher accuracy* was addressed in [Tan and Le (2019)]. Experiments were conducted using MobileNets [Howard et al. (2017); Sandler et al. (2018)], which present a class of models for mobile and embedded vision applications using depth-wise separable convolutions to build lightweight deep neural networks. Experiments on variants of MobileNet showed that larger kernels lead to more parameters and more accuracy, but the accuracy drops down above kernel sizes of 9×9. This observation verified the benefits of convolutional neural networks compared to fully connected networks, which are equivalent in the extreme case when kernel size is equal to the input resolution. It was then suggested to combine the effects of different kernel sizes, provided that both large and small kernels were needed to capture high-resolution and low-resolution patterns for higher accuracy models. MixNets using mixed convolutions of varying sizes improved the accuracy for MobileNets, on both image classification and object detection tasks.

Inspired by these insights, two different analysis/synthesis blocks have been implemented by modifying WCAE, as depicted in Figure 7.1. The architecture with analysis/synthesis blocks in Figure 7.1 is referred to as ResWCAE. ResWCAE has an increased number of outputs compared to WCAE, i.e. 64 outputs at each scale instead of 32, is deeper and has a residual architecture with the use of skip connections [He et al. (2016)]. The same architecture is used to create the third model, referred to as ResMixWCAE. All 3×3 convolutional kernels in the middle and finest scales of ResWCAE are replaced by 5×5 and 7×7 kernels, respectively [Tan and Le (2019)]. ResWCAE and ResMixWCAE reduce the input approximately by a factor of 15, in size.

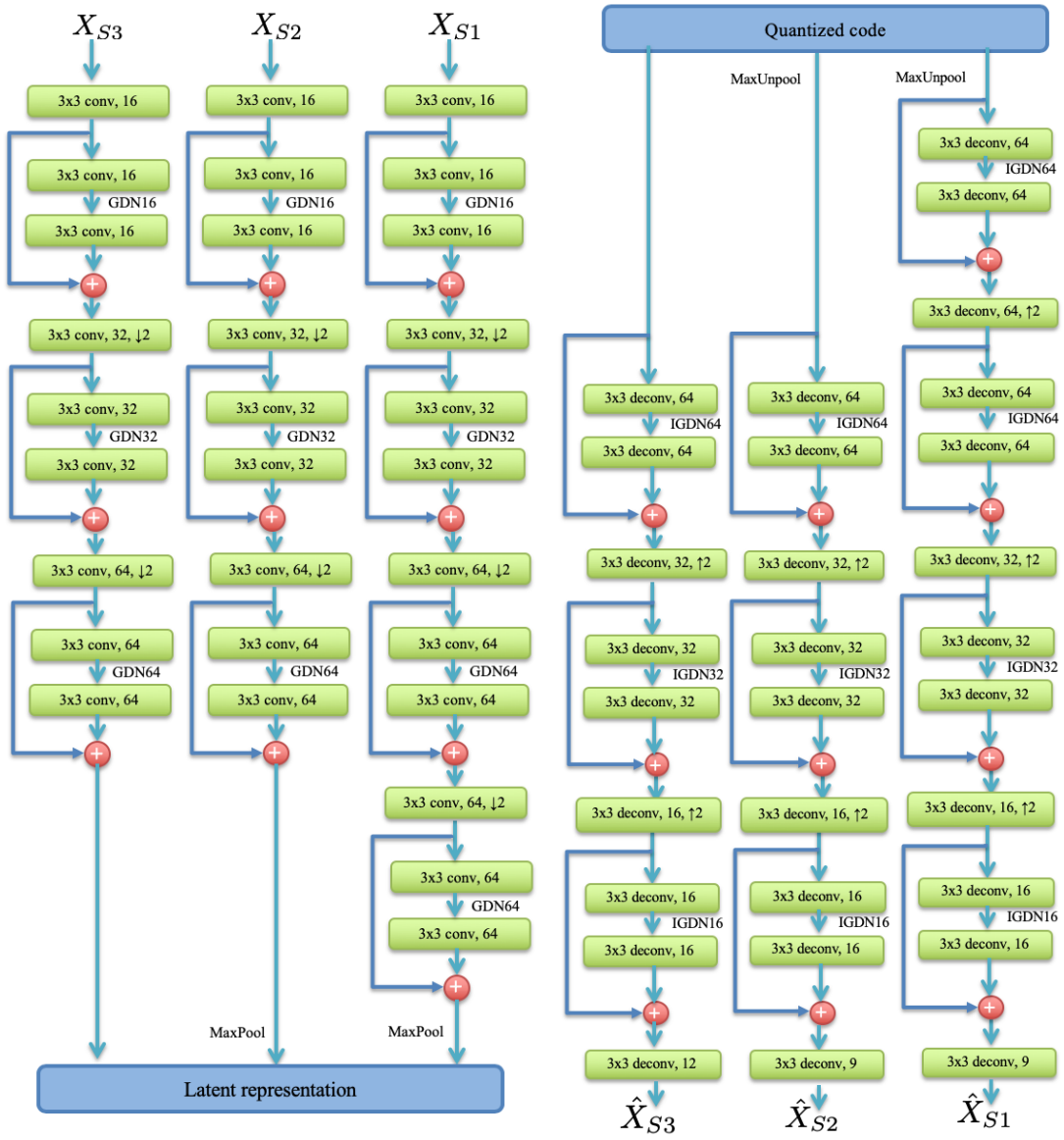


Figure 7.1 – Analysis and synthesis blocks of ResWCAE. The same architecture is used for ResMixWCAE, with the kernel sizes of the middle and fine scales increased to 5 and 7, respectively.

7.2 Experiments and results

For training the networks, the mobile and professional training datasets of CLIC2019 were used and each image was divided into 256 × 256 non-overlapping patches as was done in Chapter 6. The networks were trained iteratively in an identical fashion to that of WCAE, using back propagation and the Adam optimizer with a batch size of 8 and learning rate of 10^{-4}. Training continued for a total of 100 epochs for each network, where losses had converged to

stable values. The parameter λ was tuned to yield target bitrates of [0.12, 0.25, 0.50, 0.75, 1.00] bpp. Early stopping criterion was applied, i.e. models with the lowest validation error were selected as final models for testing.

The results of the proposed methods were compared to three different transform-based codecs, JPEG, JPEG 2000 and WebP, using the objective metrics PSNR, SSIM, MS-SSIM, VIF and VMAF. Performance plots for each metric are depicted in Figure 7.2. All results have been averaged on the complete test dataset, which contained 330 different images.

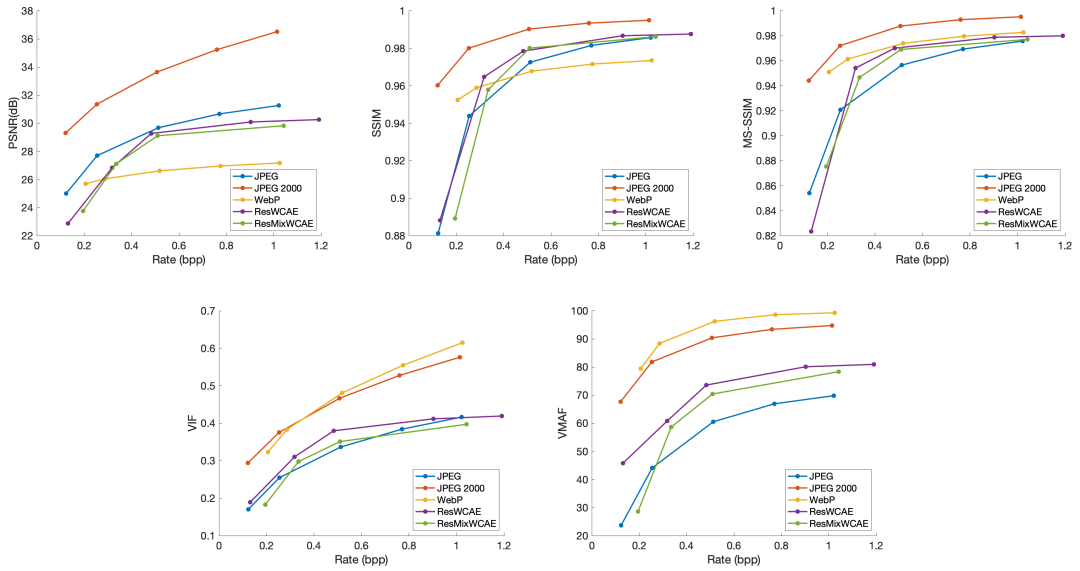


Figure 7.2 – Performances of codecs with respect to selected objective metrics measured at target bitrates.

WCAE was designed for the low bitrate track of the CLIC2019 challenge, and therefore was optimized for bitrates as low as 0.15bpp. The lower number of outputs of the analysis stage of WCAE limit the rate at below 1.00bpp on the test set, when no regularization is applied on the loss function. The latent representation of ResWCAE and ResMixWCAE are almost twice the size of WCAE and consequently perform better at higher bitrates.

Despite the fact that the networks were trained using MSE loss, the PSNR of ResWCAE and ResMixWCAE are lower than JPEG and JPEG 2000 at all target bitrates. The proposed models have higher PSNR than WebP, however this is because images were converted to YCbCr4:2:0 format before encoding with WebP and the conversions add a luminance shift to the decoded WebP image. The visual quality of WebP is superior to the proposed models, as can be verified in Figures 7.4-7.7. On the other hand, the performance of proposed models surpass JPEG in terms of SSIM, MS-SSIM, VIF and VMAF, with ResMixWCAE performing slightly worse than ResWCAE.

(a) (b) (c) (d)

Figure 7.3 – Cropped regions from test images selected for illustration.

Selected examples from the test set are depicted in Figure 7.3, where 7.3(a)-(d) have been cropped from three test images. The corresponding decoded images are presented in Figures 7.4-7.7. A closer examination verifies that the proposed methods perform better than JPEG at all target bitrates. JPEG has significant blocking artifacts at the lower bitrates, whereas at the higher bitrates some details are smoothed out as can be seen in Figure 7.4. A similar effect is observed at the lowest bitrates for WebP, where the contents have been smoothened. This can be seen clearly on Figures 7.4-7.6, also with some blocking artifacts on Figure 7.4. JPEG 2000, on the other hand, suffers from ringing artifacts at the lowest bitrates.

The artifacts of ResWCAE and ResMixWCAE are very different from each other and the transform-based codecs at the lowest bitrate. ResWCAE preserves the high frequency components of the images by sharpening the image, at the cost of adding high frequency noise. ResMixWCAE has less high frequency noise compared to ResWCAE at the expense of loss in details. Distortions in the color channels can be traced on both ResWCAE and ResMixWCAE images, especially on Figures 7.4 and 7.7. Such effects cease as bitrates increase, however some high frequency artifacts remain, especially for ResMixWCAE. This indicates that larger kernels on the finer wavelet scales contribute more to high frequency noise.

Figure 7.4 – Reference image in Figure 7.3a compressed using JPEG, JPEG 2000, WebP, ResWCAE and ResMixWCAE from top to bottom, at target bitrates 0.12, 0.50 and 1.00 bpp from left to right, respectively.

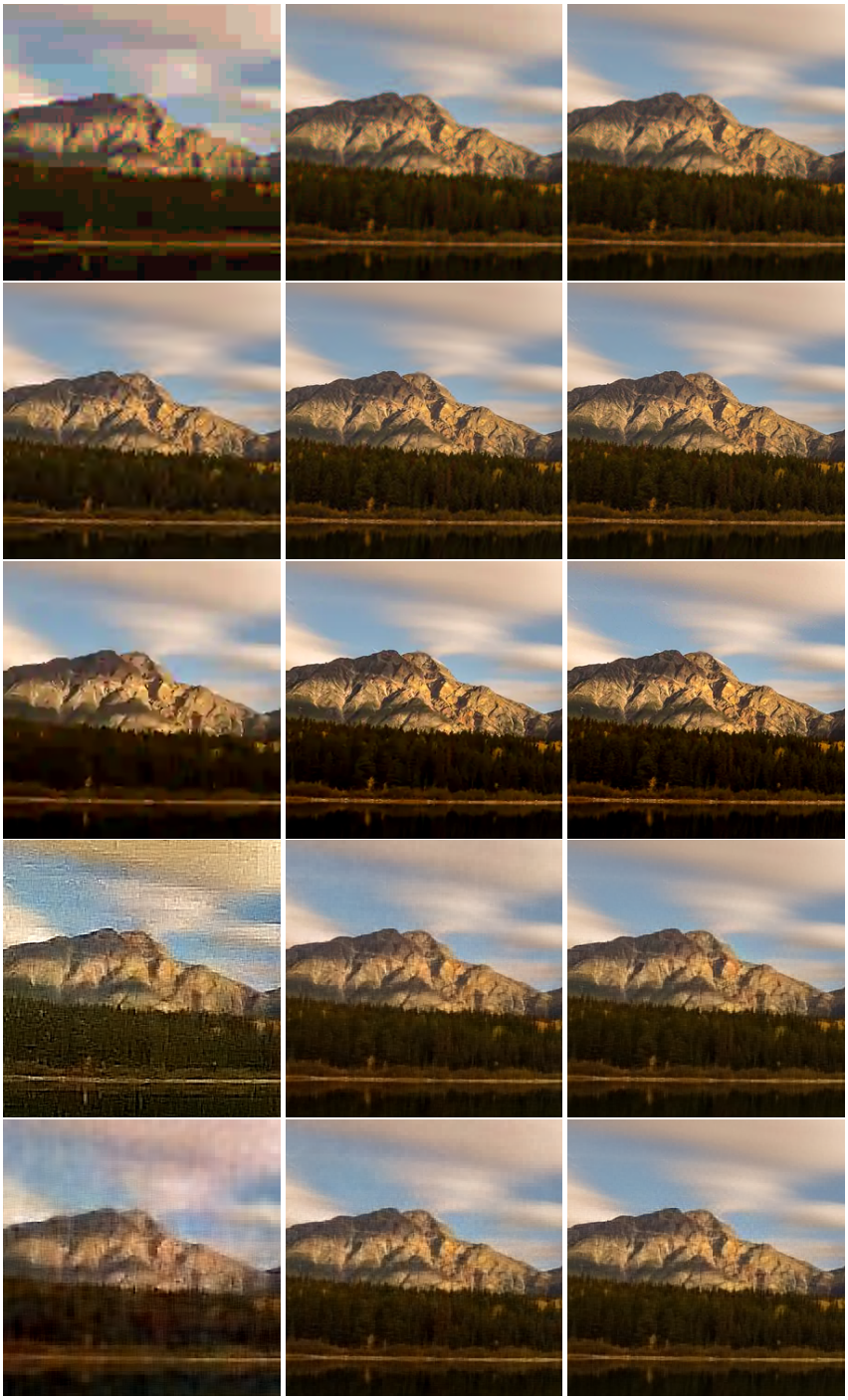


Figure 7.5 – Reference image in Figure 7.3b compressed using JPEG, JPEG 2000, WebP, ResWCAE and ResMixWCAE from top to bottom, at target bitrates 0.12, 0.50 and 1.00 bpp from left to right, respectively.

Figure 7.6 – Reference image in Figure 7.3c compressed using JPEG, JPEG 2000, WebP, ResWCAE and ResMixWCAE from top to bottom, at target bitrates 0.12, 0.50 and 1.00 bpp from left to right, respectively.

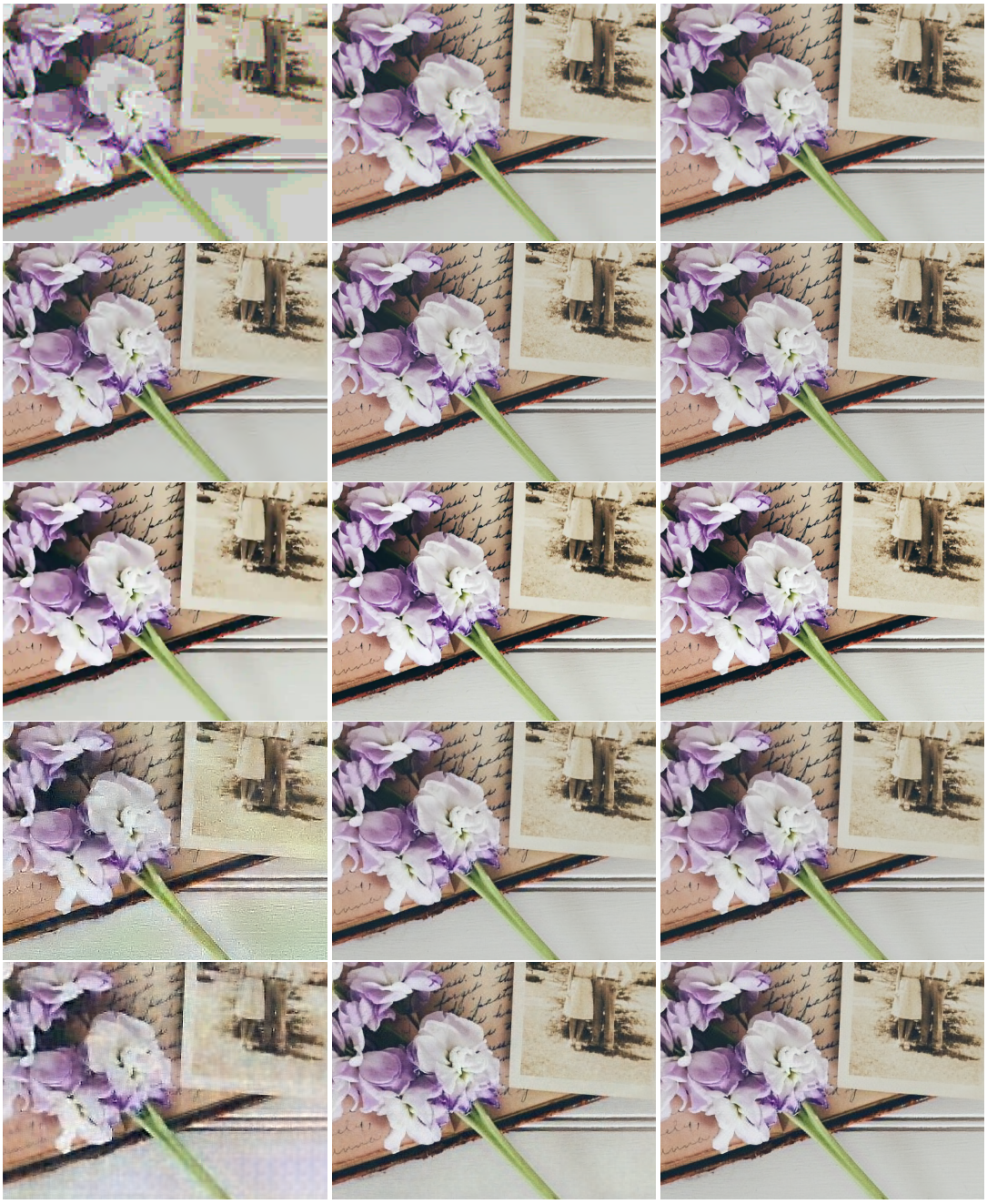


Figure 7.7 – Reference image in Figure 7.3d compressed using JPEG, JPEG 2000, WebP, ResWCAE and ResMixWCAE from top to bottom, at target bitrates 0.12, 0.50 and 1.00 bpp from left to right, respectively.

The use of mixed convolutions is less likely to pose problems for the coarsest wavelet scale, as a low-pass filtered low resolution image resides on the coarse branch that can be sparsified using convolutional neural networks. For the mid and high frequency subbands, however, the network is operating on an already sparsified representation with wavelet coefficients being orthogonal components. This suggests that learning the desired representation for compression efficiency will be more difficult for these scales compared to the coarse scale. Bringing larger convolutional kernels, on the other hand, increases the parameters drastically and therefore fails to improve the compression performance of ResMixWCAE. The number of parameters or WCAE, ResWCAE and ResMixWCAE are 264926, 961886 and 3451998, respectively. The complexities of ResWCAE and ResMixWCAE are approximately 3.5 and 13 times that of WCAE.

The performance gap between ResWCAE and ResMixWCAE suggests that the latter needs to be trained on a larger dataset. A more intuitive improvement, however, would be to carry out an ablation study to see effects of changing kernel sizes at each scale. Another important insight would be obtained by comparing these models to networks with similar architectures, which take images without preprocessing as inputs rather than using wavelet coefficients.

7.3 Conclusion

In this chapter, two new end-to-end image compression architectures based on convolutional neural networks have been presented. The networks have been built as extensions to the model previously introduced in Chapter 6 which uses 2D wavelet decomposition as a preprocessing step before training. ResWCAE and ResMixWCAE have deeper architectures than WCAE, employ residual connections and have a wider bottleneck at the analysis which allow them to be optimized at multiple rate points. ResMixWCAE also has varying kernel sizes along its branches processing wavelet coefficients at different scales. Results show that both models outperform JPEG compression, but are inferior to JPEG 2000 and WebP when compared using objective metrics. Subjective results indicate that ResWCAE and ResMixWCAE are able to preserve high frequency components, reduce blur and introduce no ringing or blocking artifacts. ResWCAE has more high frequency noise at lower bitrates, whereas ResMixWCAE suffers from more high frequency noise at higher bitrates.

The contributions of this chapter are summarized as follows:

- The effects of using a deeper residual network with a wider bottleneck on learning-based compression were tested in comparison to the WCAE model. While the performance of models using pixels as inputs was improved with such alterations, the ResWCAE model that uses 2D wavelet coefficients did not benefit from the proposed architecture at low bitrates. At higher bitrates, however, the MS-SSIM performance of ResWCAE reached the performance of WebP while still being superior to JPEG.
- The effects of using mixed convolutional kernel sizes on the three wavelet scales on top of the residual connections and wider bottleneck were investigated on the ResMixW-

CAE model. While it was hypothesized that using larger kernel sizes would result in capturing higher level features on the finer wavelet scales, it was demonstrated that the transformed representation using wavelet coefficients could not benefit from this increased receptive field. The performance of ResMixWCAE was found to be inferior to ResWCAE, which was more easily optimized owing to its less complex architecture.

ResWCAE was deepened at the expense of more parameters, while ResMixWCAE suffered even more from complexity. Although wavelet decomposition as a preprocessing step had been proven effective at low bitrates for networks of smaller size, this chapter suggests that such preprocessing fails to be as effective for more complex models. More insight can be gained from testing different distributions of kernel sizes and network depths on all three scales. Comparing ResMixWCAE with a similar architecture that processes pixel value inputs through deep residual mixed convolutional neural networks would highlight the strengths and weaknesses of the proposed models thoroughly.

8 Exploration study on the effect of mixed kernels

In the previous chapter, the effects of using convolutional layers with varying kernel sizes when processing the wavelet coefficients of images for compression were explored. ResMixWCAE was motivated by a combination of the findings in [Cheng et al. (2019a)] demonstrating the positive effects of increasing convolutional kernel sizes on PSNR and MS-SSIM of compressed images, and the demonstrated advantages of using wavelet decomposition as a preprocessing step. Instead of using an increased kernel size, employing a combination of mixed kernel sizes was tested on an end-to-end learning-based compression framework. The approach was expected to compound the analysis of spatial correlation and semantic information at multiple scales, however, the performance of ResMixWCAE was shown to be inferior to ResWCAE. Chapter 7 was concluded by expressing the need for an exploratory study on the strengths and weaknesses of the proposed models. Such explorative research is presented in this chapter by investigating the effects of mixed networks on image compression.

8.1 Networks

Three networks were constructed in the form of an ablation study. The most complex network is referred to as TripleNet, which has three branches of convolutional kernels 3×3, 5×5 and 7×7. The second network, DoubleNet, has only two branches, with the 7×7 layers removed. The third and final network is SingleNet, which only has 5×5 kernels. The architectures were inspired by [Ballé et al. (2018)] and [Cheng et al. (2019a)], and are depicted in Figure 8.1. TripleNet is composed of all components of the network. $N1 = N2 = N3 = 32$ outputs are collected from each branch to yield a latent representation of dimensions $(N1 + N2 + N3) \times \frac{W}{2^n} \times \frac{H}{2^n}$. DoubleNet is composed of the first two branches, with $N1 = N2 = 48$ to yield the same number of outputs at the latent representation. SingleNet has only the middle branch and $N2 = 96$ outputs to yield the latent representation by itself. $n = 4$, for there are 4 downsampling operations in total, between convolutional layers of each branch.

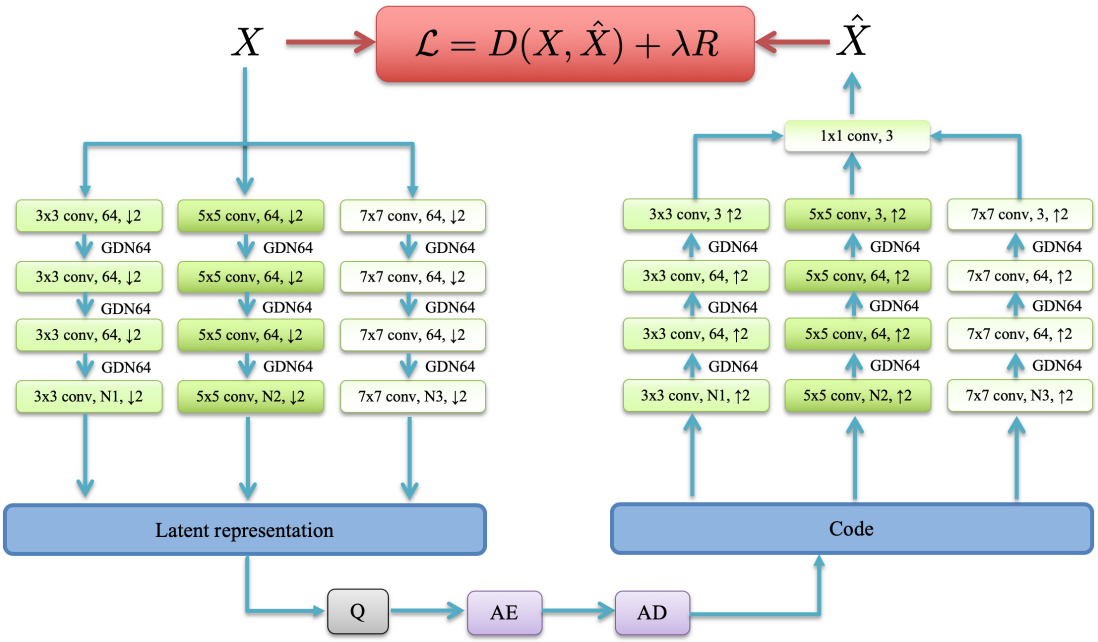


Figure 8.1 – TripleNet, DoubleNet and SingleNet frameworks combined in a single illustration.

While the output size $N1 + N2 + N3 = 96$ was kept constant for all three networks, such small bottleneck is expected to affect the compression performance negatively. However, in order to attain an intermediate range of bitrates, the experiments were conducted using the widest bottleneck possible. The large computational complexity of TripleNet had maxed out the capacity of available resources at an output size of 192, which did not allow the model to reduce the rate beyond 0.7bpp. Therefore, 96 was selected as a reasonable output size to demonstrate the effects of kernel sizes.

For DoubleNet and TripleNet, the reconstructions obtained from the branches need to be combined to yield a single three-channel image. A preliminary study was conducted to investigate the influence of the methodology for dimensional reduction at this final step. Averaging the 9 channels to yield a 3-dimensional output was compared to using a 1×1 convolutional layer with 3 outputs. The latter approach was found to be more effective in delivering better quality, with a small increase in the number of parameters in total.

The networks were trained using a combination of training images in CLIC and JPEG AI dataset. Here, the JPEG AI training dataset refers to the whole set of 5264 images with SD to UHD resolution, as mentioned in Chapter 4, and not the subset of 15 images that was used to construct the IQA database. Again, the maximum patch size allowed by available resources was chosen for training, which was 256×256. In this case, however, the patch size was too small for the network to extract meaningful information from UHD images in the JPEG AI

dataset. Therefore, images with maximum dimension greater than 1024 were downsampled to reduce the maximum dimension to either 1024 or 512 randomly, while preserving the ratio of dimensions. Such downsampling is not expected to affect the quality of the images, on the contrary, it helps to reduce the compression artifacts that might prevail in some of the collected images.

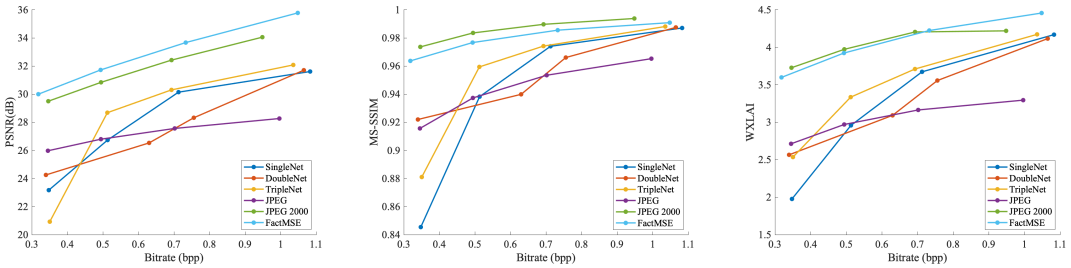
A total of 5760 downsampled images were considered and random patches of dimensions 256×256 were extracted at each iteration. Each network was trained through stochastic gradient descent for approximately 1M iterations, using a batch size of 8 and the Adam optimizer with a learning rate of 10^{-4}. 12 models were trained in total, with 4 models belonging to each architecture at target rates 0.35, 0.50, 0.70 and 1.00 bpp that were attained using $\lambda = 128, 64, 32$ and 8, respectively. During training, quantization was approximated by adding uniform noise in the range [-0.5, 0.5], and rate was approximated as the factorized entropy of the latent representation. At test time, quantization was performed by rounding each element of the latent representation to the nearest integer, and the range coder [Theis et al. (2017)] was used for entropy coding.

8.2 Results and analysis

The performance versus bitrate curves for SingleNet, DoubleNet and TripleNet are provided in Figure 8.2. Objective evaluations were conducted using metrics PSNR, MS-SSIM and WXLAI. JPEG and JPEG 2000 results were included for comparison with transform-based anchors. The FactMSE model was also involved as an example of an autoencoder trained using MSE loss and factorized entropy model. At test time, FactMSE codes were encoded using a non-adaptive binary arithmetic coder. It must be noted that FactMSE was trained on a set of approximately 1 million UHD images downsampled by a randomized factor such that the minimum of their height and width was within the range of 640 to 1200 pixels. In addition, FactMSE had 192 outputs at the latent layer for target bitrates 0.35, 0.50 and 0.70bpp, and 320 outputs for target bitrate 1.00bpp. Due to memory and time constraints during the execution of models investigated in this chapter, utilization of such output numbers and extensive training sets were not attainable.

At first glance, Figure 8.2 suggests that TripleNet is the best performing codec among the three, followed by SingleNet. The performance of DoubleNet is surprisingly lower than the remaining two, except at the highest bitrate where the performances are almost equal, and at the lowest bitrate, where DoubleNet outperforms SingleNet according to all tested metrics and TripleNet with respect to PSNR and MS-SSIM. Moreover, TripleNet's performance plummets in terms of PSNR at the lowest bitrate. Such results require a more in-depth analysis involving visual cues from the output images.

The behavior of the three codecs at the lowest target bitrate is visualized in Figure 8.3. While trying to reach the target bitrate by reducing the entropy of the latent code, SingleNet features processed through 5×5 kernels occupy the total length of the latent representation. In this

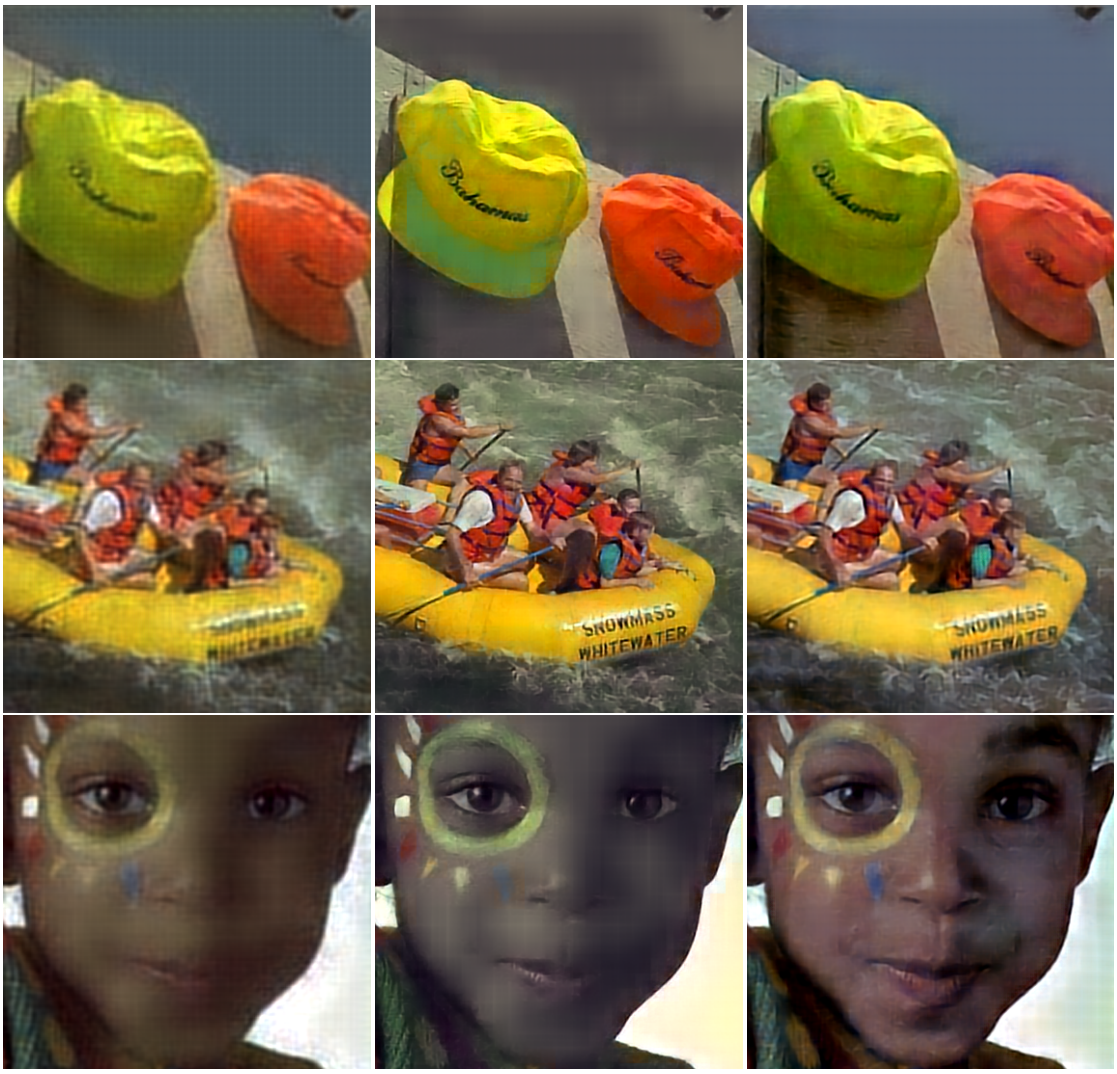


(a) PSNR (b) MS-SSIM (c) WXLAI

Figure 8.2 – Objective quality assessment of SingleNet, DoubleNet and TripleNet at target bitrates.

case, eminent blocking artifacts occur, as well as ringing artifacts that can be traced clearly in kodim03. When 3×3 kernels are added, the code receives a quantitatively equal amount of information from the 3×3 and 5×5 kernels. The code still needs to reach the target bitrate that is relatively low, but now it is able to achieve a trade-off between the information from two scales to attain pixel-wise similarity with the reference. The image is smoother at low frequency areas, yet the details are preserved better. In kodim03 and kodim15, the smoothening effect is observed at the expense of color fidelity. As soon as the third branch with 7×7 kernels is added, the latent code is quantitatively divided equally between the two branches. The additional information from processing of larger blocks helps to restore color fidelity, while preserving low frequency and high frequency content. However, since the code receives less information from 5×5 and 3×3 kernels with respect to DoubleNet, some details are less apparent such as text in kodim03 and kodim14. This explains why at lower bitrates, TripleNet is not always as efficient as DoubleNet. DoubleNet is more effective at preserving the overall structure of the image, as reflected by its high MS-SSIM score, while visual quality of DoubleNet and TripleNet is similar, and superior to SingleNet, as conveyed by both the WXLAI scores in Figure 8.2(c) and the examples in Figure 8.3.

Figure 8.4 depicts visual examples from outputs of the models at target bitrate 0.5bpp. The subjective quality of the models has improved considerably owing to less constraints. SingleNet is able to preserve both low and high frequency components, with reduced overall noise. Blocking and ringing artifacts are still present, albeit reduced. The addition of 3×3 kernels provides information at smaller scales, which results in better preservation of high frequencies. Most apparent examples can be traced on the clouds above the yellow cap and lines on the cap itself in kodim03, and the facial features in kodim15. However, since the contribution of 5×5 kernels is now reduced, the details appear at the expense of high frequency noise. A more pleasing balance is achieved with the addition of 7×7 kernels. TripleNet suffers less from blocking artifacts and high frequency noise, and preserves sufficient amount of details. On average, the performance of TripleNet is superior to that of DoubleNet and SingleNet at target bitrate 0.5bpp. The WXLAI performances of DoubleNet and SingleNet are very close to each other. This indicates that at the targeted bitrate, the trade-off between preserving high



(a) SingleNet (b) DoubleNet (c) TripleNet

Figure 8.3 – Cropped sections of kodim03, kodim14 and kodim15 images from Kodak image database, compressed before cropping using SingleNet, DoubleNet and TripleNet at target bitrate 0.35bpp.

frequency components and introducing high frequency noise is equivalent to the trade-off between reducing high frequency noise and smoothing out the details in terms of perceived quality.

All three networks have increased objective and subjective quality at the third target bitrate. Interestingly, SingleNet and TripleNet perform very similarly subjectively and objectively. SingleNet, however, has more blocking artifacts that can be traced clearly on kodim15. Again, DoubleNet is preserving more high frequency components, but the amount of details conserved in SingleNet are now abundant and visually more pleasing at the absence of noise.

(a) SingleNet (b) DoubleNet (c) TripleNet

Figure 8.4 – Cropped sections of kodim03, kodim14 and kodim15 images from Kodak image database, compressed before cropping using SingleNet, DoubleNet and TripleNet at target bitrate 0.50bpp.

Table 8.1 – PSNR(dB) and MS-SSIM and WXLAI results on images kodim02, kodim04 and kodim15 at the highest target bitrate, compressed using FactMSE, TripleNet and JPEG as depicted in Figure 8.7.

	FactMSE			TripleNet			JPEG		
	PSNR(dB)	**MS-SSIM**	**WXLAI**	**PSNR(dB)**	**MS-SSIM**	**WXLAI**	**PSNR(dB)**	**MS-SSIM**	**WXLAI**
kodim02	38.23	0.9926	4.4746	35.88	0.9842	4.3301	31.26	0.9720	3.7749
kodim04	38.45	0.9935	4.6624	33.75	0.9868	4.1540	31.11	0.9732	3.5103
kodim15	38.68	0.9943	4.5855	33.16	0.9883	4.3659	30.45	0.9806	3.7153

(a) SingleNet (b) DoubleNet (c) TripleNet

Figure 8.5 – Cropped sections of kodim03, kodim14 and kodim15 images from Kodak image database, compressed before cropping using SingleNet, DoubleNet and TripleNet at target bitrate 0.70bpp.

At the highest target bitrate 1.00bpp, the three architectures yield decoded images of equivalent subjective and objective quality, as depicted in Figures 8.2 and 8.6. All tested codecs reach transparent quality at the highest bitrate except JPEG. Figure 8.7 depicts examples from tested codecs at target bitrate 1.00bpp. FactMSE and TripleNet have subtle differences, in fact compared to the original, TripleNet is preserving textures better than FactMSE. While FactMSE smooths out the images, TripleNet conserves details better and achieves a more natural subjective experience.

(a) SingleNet (b) DoubleNet (c) TripleNet

Figure 8.6 – Cropped sections of kodim03, kodim14 and kodim15 images from Kodak image database, compressed before cropping using SingleNet, DoubleNet and TripleNet at target bitrate 1.00bpp.

8.3 Conclusion

This chapter presented an exploratory study on the effect of combining varying kernel sizes in convolutional autoencoders. Motivated by the results in Chapter 7 and the findings in literature promoting increased kernel sizes for autoencoder architectures, the contributions of 3×3, 5×5 and 7×7 convolutional layers were investigated in the form of an ablation study. Three autoencoders were designed with the same number of outputs at the latent stage, formed by the outputs of convolutional layers employing varying sized kernels. As increased kernel sizes yielded enhanced performance in [Cheng et al. (2019a)], an intuitive expectation

(a) Reference (b) TripleNet (c) FactMSE (d) JPEG

Figure 8.7 – Cropped sections of kodim02, kodim04 and kodim15 images from Kodak image database, compressed before cropping using TripleNet, FactMSE and JPEG at target bitrate 1.00bpp. Bitrates for kodim04, kodim02 and kodim15 for TripleNet, FactMSE and JPEG are as follows: TripleNet = {0.9840, 0.9618, 0.9585} bpp, FactMSE = {1.2105, 1.1739, 1.454} bpp, JPEG = {0.9926, 1.0031, 1.0034} bpp.

was to observe increased quality as a results of information from multiple scales. It was shown, however, that the trade-off between quality and rate divided among multiple scales is not as straightforward. The findings of this chapter are summarized as follows:

- SingleNet with kernel size of 5×5 at all four convolutional layers was performing poorly at the lowest bitrate. Contributions of features from smaller kernels resulted in preservation of more details at the same bitrate. Adding more features from larger kernels resulted in a trade-off between preservation of low-level and higher level features. As the output size was constant, TripleNet's performance was better than SingleNet, but equivalent to DoubleNet. DoubleNet was preserving the details aggressively, whereas TripleNet found a better balance between compressing high and low frequency components.
- With increasing rate budget, the contribution of 3×3 convolution outputs kept preserving high frequency noise. This might be considered as an undesirable effect at times, however when the goal of compression is to preserve components such as film grain, the

combination of 3×3 and 5×5 kernels would construct an effective solution, provided that a good balance between the output sizes of 3×3 and 5×5 branches.

- At the highest target bitrate tested, i.e. 1.00bpp, all networks reached equivalent performance. The performance of the proposed models was then comparable to FactMSE which had been trained on the same loss and a very similar architecture to that of SingleNet. SingleNet, however, was still distinguishably noisier with respect to the reference compared to TripleNet and FactMSE. Figure 8.7 provided example image sections comparing TripleNet, FactMSE and JPEG. Objective results in Table 8.1 show that there is little difference between the performances of TripleNet and FactMSE, especially in terms of WXLAI ratings. Subjective examination proves that using mixed kernel sizes together with a high rate budget resulted in a more loyal representation of the decoded image with respect to the original, in comparison to FactMSE which tended to over-smooth contents.

With the insight provided in this chapter, changing the architectures of autoencoders toward more efficient compression modes depending on the context at hand is possible. The complexities of models increased from 751K to 829K and to 1.8M parameters in total through adding layers from SingleNet to DoubleNet to TripleNet. One alternative for reducing complexity while maintaining TripleNet architecture is exchanging larger kernels with cascaded layers of smaller kernel sizes. In any case, an optimized architecture would need a balance between the contributions from the branches. The autoencoder can be tuned to put more emphasis on the contribution of filters of smaller dimensions when elements such as film grain are desired to be preserved. An autoencoder with different "modes" could be achieved, that preserves different qualities of images by tuning the architecture based on input parameters.

The presented study provides valuable insight on the effect of multiscale processing of images, which has proven to be very useful for compression through codecs such as JPEG 2000, HEVC/H.265-Intra and VVC. In particular, HEVC/H.265-Intra and VVC search for a block partitioning of the image by splitting coding tree units into multiple coding units, depending on the local characteristics of the image. The optimal partitioning is determined through an exhaustive search, which could also be accelerated by machine learning. Such partitioning can be combined with locally optimized kernel sizes to extract features from coding units and achieve higher compression performance.

The exploration study on mixed kernels needs to be extended further by investigating the contributions of meaningfully higher sized kernels, i.e. 9×9, and checking the effects of other combinations of the branches, i.e. 5×5 and 7×7, and higher dimensions. The small bottleneck in this study needs to be increased in order to achieve an improved balance between features from branches. Training the models on a higher number of examples would yield more robust results. Finally, entropy estimation using a scale hyperprior [Ballé et al. (2018)] is expected to increase the performance, but exploring the extent of hyperprior's contribution to compression quality in the presence of mixed networks poses an intriguing research problem.

Part III

Towards unified learning-based image compression solutions

9 Learning-based image compression using learning-based image quality metric

In Part I and Part II, learning-based solutions to image compression and image quality assessment, particularly within the context of compression related degradations, were presented. Autoencoders comprised of different architectures, i.e. WCAE, ResWCAE, ResMixWCAE, SingleNet, DoubleNet and TripleNet were evaluated at various rate points and compared to state-of-the-art compression algorithms. The presented autoencoders were all trained using MSE loss between original and decoded images, which measures the average pixel-wise difference between the former and the latter. MSE, however, is not a sophisticated quality metric that correlates well with subjective metrics, as it does not take into account any HVS-related features, structural similarities or higher order pixel statistics, which also impact the subjective quality of images. State-of-the-art autoencoders also employ MS-SSIM loss instead of MSE, which helps more in preservation of structural similarity between reference and decoded image, however, there are two main drawbacks observed in the literature as well as the experiments presented in Chapter 4: (i) while preserving the local structures, MS-SSIM loss results in a degradation in the overall contrast particularly at lower bitrates as shown in 4.18, (ii) autoencoders trained with MS-SSIM loss expectedly yield to compressed images with higher MS-SSIM scores, yet the correlation between MS-SSIM and subjective ratings is not always high. In Figure 4.13(b), HyperMS-SSIM, FactMS-SSIM and JPEG 2000 are the best performing codecs in terms of MS-SSIM, however in Figure 4.14 for content TE08, the subjective performance of HEVC/H.265-Intra stands out as well, leaving HyperMS-SSIM and HyperMSE behind. In addition, Table 5.3 indicates that PLCC of MS-SSIM with subjective scores on compressed images is still lower than that of PSNR, SSIM, VMAF and WXLAI.

Of course, such evidence on a limited number of examples is not enough to reach conclusive remarks on subjective accuracies of metrics and their effects on the performance of autoencoders. However, with WXLAI metric proven to be effective in quality assessment of compressed images within a full-reference framework, a natural direction of research is to study the effects of WXLAI on autoencoders when used as the distortion measure in the loss

function.

To the author's best knowledge, autoencoders employing learning-based metrics has not been investigated in the state of the art concerning learning-based image compression. In this chapter, WXLAI is used in the loss function of an autoencoder instead of a transform-based metric. The SingleNet architecture introduced in Chapter 8 was preferred due to its low computational complexity among other models. The effects of WXLAI metric on training the autoencoder were analyzed from a number of perspectives. The autoencoder was trained firstly using weights initialized with He et al. (2015) and then on weights of a pretrained model. Afterwards the effects of a loss function as a weighted combination of PSNR and WXLAI were examined. Results indicate that the feedback provided by WXLAI metric achieves high sensitivity yet low precision, as explained in more detail in the following sections.

9.1 Convolutional autoencoder with learning-based objective metric (CAE-LM)

The CAE-LM architecture is depicted in Figure 9.1. The input image X is fed into SingleNet and decoded image $\hat{X}$ is constructed. X and $\hat{X}$ are then randomly separated into 128 corresponding patches of dimensions 128×128 and fed into the WXLAI network. The output of WXLAI network is the quality score between images X and $\hat{X}$, which needs to be maximized to ensure good compression quality. This corresponds to minimizing an inverse mapping of the score through stochastic gradient descent. While weights of SingleNet in CAE-LM are updated at each batch iteration, weights of WXLAI network are fixed to those of the model presented in Chapter 5. During training, the connection between input image X and the loss function need to be traced properly. Equation 9.3 shows that the score S is a function of input X, which ensures that in the case S is involved in the loss function $\mathscr{L}$, the gradients will propagate back to the input.

$$\hat{X} = \text{SingleNet}(X) \tag{9.1}$$

$$S = \text{WXLAI}(X, \hat{X}) \tag{9.2}$$

$$= \text{WXLAI}(X, \text{SingleNet}(X)) \tag{9.3}$$

The distortion measure between X and $\hat{X}$ is then:

$$D(X, \hat{X}) = \alpha \text{MSE}(X, \hat{X}) + \beta S(X, \hat{X}) \tag{9.4}$$

In Chapter 8, SingleNet was trained with an equivalent distortion measure, with $\alpha = 1.0$ and $\beta = 0.0$. In this chapter, this condition is reversed to the other extreme, i.e. $\alpha = 0.0$ and

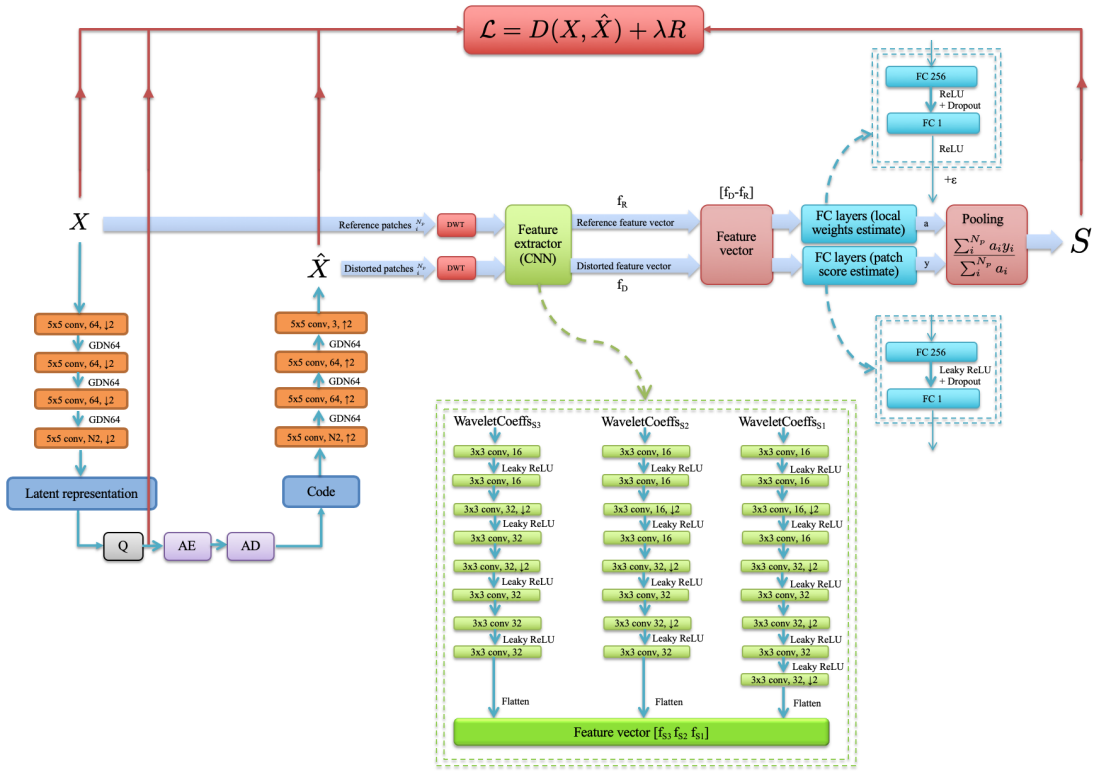


Figure 9.1 – CAE-LM architecture.

$\beta = -1.0$. Initially, the effects of this configuration are analyzed on an untrained network and a pre-trained network. A different balance between (α, β) is then proposed to ensure good compression quality.

9.2 Experiments and results

9.2.1 CAE-LM training on He initialization

In extreme case when MSE loss and rate optimization are completely disregarded with setting $(\alpha, \beta) = (0.0, -1.0)$, the overall batch loss function is governed solely by the batch score. $\lambda = 0$ was used to employ unconstrained rate in order to see the maximum attainable performance of the model. 512×512 input patches extracted randomly from downsampled CLIC and JPEG AI training sets were used during training with a batch size of 8. 128 mini patches of dimensions 128×128 were randomly extracted from each patch for score computation. This setting was preferred since WXLAI metric was trained by extracting 128 patches of dimensions 128×128 from SD to UHD images, and performed well on the 512×384 images of TID2013 dataset. Patch dimensions of 256×256 could be less effective in score computation and therefore yield to a decrease in performance. The model was trained using Adam optimizer

with a learning rate of 10^{-4}. Validation was conducted on the 452 downsampled CLIC and JPEG AI validation images. Tests were evaluated on the Kodak image database.

Through iterations, the score of training images increased while the loss decreased, verifying that CAE-LM model was able to learn from the data and reduce the cost function accordingly. However, the PSNR of the images were extremely low, as illustrated in Figure 9.2. The decoded image in Figure 9.2 is composed of repeating patterns, with very low subjective quality. WXLAI score between reference and the decoded images is, however, 10.0761. Considering that WXLAI scores between the same inputs, i.e. reference image and reference image or distorted image and distorted image, is 4.8944, a score of 10.0761 means the quality of the distorted image is even better than if it were the reference.

Figure 9.2 – Reference image (right) and decoded image (left) from training set, with S = 10.0761 and PSNR = 7.5470dB, at rate 2.60bpp.

During the training of WXLAI model, reference and corresponding distorted images were introduced to the network. Although varying levels and types of distortions were used, the network had never experienced completely corrupted images with PSNR scores below 23dB. All the reference-distorted pairs that were introduced to the WXLAI network during training had a certain level of correspondence in terms of structure, contrast and pixel statistics, albeit at changing levels. As the CAE-LM network is initialized, however, the first iterations generate images with very low pixel values, resulting in seemingly black images with a PSNR of 6dB on average. When MSE loss was employed, the autoencoder tried to adjust pixel values as to match the reference image. On the contrary, WXLAI tries to extract information from the wavelet coefficients of the reference and distorted images and computes scores based on a learned mapping of the difference of these features. The reference and distorted image features are constructed using convolutional layers of kernel size 3×3, indicating that the visual field is very small. Although these settings were efficient for training a metric to evaluate the quality of compressed images of various resolutions, the same architecture and weights iteratively converted the decoded images into the patterned images presented in Figure 9.2. Further training pushed the dark borders outward and filled the image completely with saturated patterns, resulting in increased score and low PSNR. Reducing the learning rate from 10^{-4}

to rates as low as 10^{-7} or using stochastic gradient descent optimizer instead of Adam to fix the learning rate did not result in any favorable outcomes. The influence of WXLAI metric on reducing the distortion between reference and compressed images did not correspond to pixel-wise similarity when CAE-LM was initialized using He initialization [He et al. (2015)].

In order to impose pixel similarity as a side condition, nonzero weights for MSE in the loss function were tried. The parameter setting $(\alpha, \beta) = (0.3, -0.7)$ was selected to observe the relative influence of MSE and WXLAI scores on visual quality of images. Preliminary results showed that the learning curve improved, with increasing PSNR and WXLAI scores as opposed to the $(\alpha, \beta) = (0.0, -1.0)$ training. As the training time of the model was much higher than the previous models, preliminary results on the validation set during the first 10K iterations is depicted in Figure 9.3. The presented plots have been smoothed using a Gaussian filter of window size 5, while the original data was kept in the background for illustrative purposes. Randomization of the training patch extraction contributed to the fluctuations, however the dominant cause was the validation procedure, which constituted an evaluation over consecutive subsets of 20 images from the validation set every 50 iterations, in order to increase training speed.

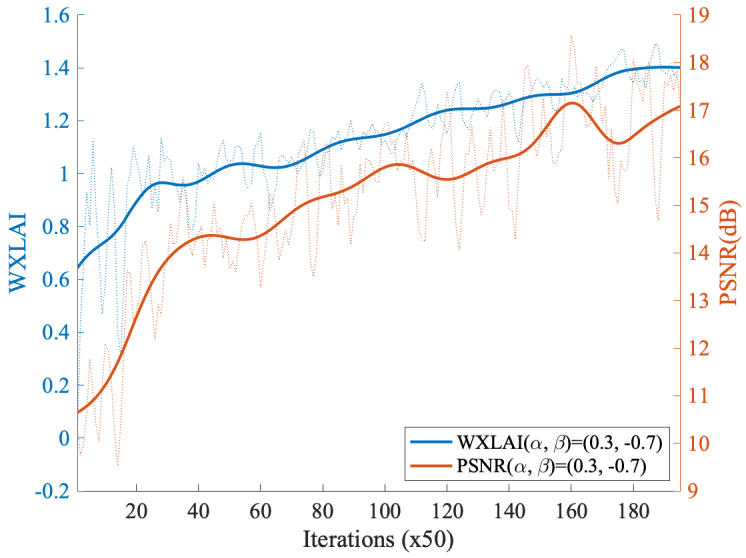


Figure 9.3 – WXLAI and PSNR values on the validation set while training CAE-LM after He initialization

In the beginning, MSE loss is much greater than WXLAI loss, therefore the loss function is governed by MSE during early iterations. If the balance is heavily shifted towards WXLAI, however, the results easily diverge and PSNR decreases, as was explained with Figure 9.2. With the selected (α, β) configuration, such effects were not observed as the values of PSNR and WXLAI indicated steady training behavior.

As training continues, MSE loss decreases and WXLAI score increases. The rate of change of

the MSE loss is much faster than that of WXLAI scores. It is therefore important to find the right balance between the two functions and employ a dynamic assignment of (α, β) during training as to decrease α and increase β gradually in order to avoid visual divergences, i.e. as in the case $(\alpha, \beta) = (0, -1)$, and to set the network to rely more on WXLAI when MSE is sufficiently low.

9.2.2 CAE-LM tuning using pre-trained compression model

Considering that CAE-LM was unable to reach a favorable operating point under the sole influence of WXLAI metric, training was modified as to start from weights that yield decoded images with features similar to the reference, according to human vision. This approach would be more practical for testing the influence of WXLAI during training, since the lengthy training times required to tune (α, β) continuously would delay the observation of meaningful results.

The selected baseline was SingleNet trained using patches of dimensions 256 × 256 at target rate point 1.00bpp. At this level, SingleNet had a PSNR of 31.6100dB and the WXLAI score was 4.1689 on the Kodak test database. Table 9.1 shows the performance of CAE-LM as a function of (α, β), evaluated using PSNR and WXLAI.

Table 9.1 – PSNR(dB) and WXLAI performance of CAE-LM on the Kodak image database when resuming from pre-trained model weights.

(α, β)	**PSNR(dB)**	**WXLAI**	**Rate(bpp)**
(1.0, 0.0)	32.2368	4.2293	2.0299
(0.0, -1.0)	32.4106	4.2888	1.9854
(0.5, -0.5)	32.2677	4.3415	2.0277
(0.3, -0.7)	32.4573	4.3567	2.0275

The formerly 256 × 256 patch size was doubled to 512 × 512 while resuming training, in order to fortify prediction performance of WXLAI metric. A decrease in the starting performance of SingleNet at 1.00bpp was caused by this change. Resuming from baseline SingleNet model towards the new target rate point 2.00bpp using $(\alpha, \beta) = (1.0, 0.0)$ reached a 32.2368dB PSNR and a WXLAI score of 4.2293. When the order of loss function was reversed to $(\alpha, \beta) = (0.0, -1.0)$ both PSNR and WXLAI scores increased to 32.4106dB and 4.2888. Contrary to the inability of CAE-LM training with He initialization in improving compression performance, the model was able to increase the ratings of both test metrics when only WXLAI was used during training resumed from SingleNet weights.

When weights of MSE and WXLAI were distributed evenly in the loss function as $(\alpha, \beta) = (0.5, -0.5)$, PSNR and WXLAI score on the test set increased again with respect to the baseline. The increase in PSNR was less compared to when $(\alpha, \beta) = (0.0, -1.0)$. This indicates that adding the influence of MSE in the loss function resulted in decoded images with higher PSNR and WXLAI, and WXLAI benefited more from the pixel-wise similarity constraint. To further

observe the influence of WXLAI, the weights of the coefficients were shifted in favor of WXLAI by setting $(\alpha, \beta) = (0.3, -0.7)$. The highest PSNR and WXLAI scores were observed, verifying that WXLAI is able to amplify the compression performance when used as the loss function, and the effect is strengthened with help from MSE.

9.2.3 CAE-LM tuning at lower rate points

The results presented in the previous section provided evidence that WXLAI loss is effective objectively, yet the visuals were all of transparent quality and therefore indistinguishable. In order to visually observe the distinction between models, two lower rate points, 0.5 and 1.0bpp, were targeted.

Table 9.2 – PSNR(dB) and WXLAI performance of CAE-LM on the Kodak image database at lowered target rates.

(α, β, λ)	**PSNR(dB)**	**WXLAI**	**Rate(bpp)**
(1.0, 0.0, 8.0)	27.3931	3.2954	1.0314
(0.3, -0.7, 8.0)	27.7432	3.4030	1.0392
(0.3, -70.0, 8.0)	26.1067	3.6699	0.9977
(1.0, 0.0, 32.0)	25.8224	2.6751	0.5264
(0.3, -0.7, 32.0)	26.2454	2.7408	0.5023

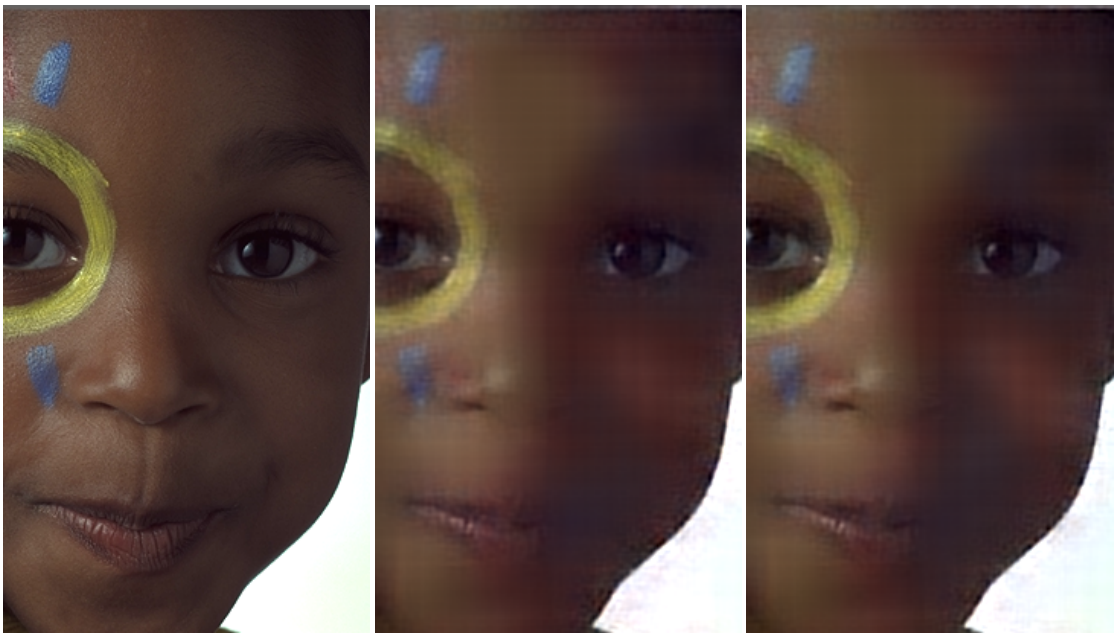
Table 9.2 depicts the results for multiple rate points and (α, β) configurations. WXLAI influence on the loss function within $(\alpha, \beta) = (0.3, -0.7)$ setting yields to an increase in PSNR and WXLAI scores at the target bitrates, compared to $(\alpha, \beta) = (1.0, 0.0)$. When the influence of WXLAI is increased by setting β to -70.0, the WXLAI score increases, accompanied by a decrease in PSNR at target rate 1.0bpp. Visual results are demonstrated in Figures 9.4 and 9.5.

Negligible difference was observed between $(\alpha, \beta) = (1.0, 0.0)$ and $(\alpha, \beta) = (0.3, -0.7)$ settings at target rate point 0.5bpp. At low rates, the influence of MSE is higher than that of WXLAI, therefore the effect of WXLAI is not evident. Very subtle differences are observed between the images at target rate point 1.0bpp. WXLAI influence contributes slightly more to the changes in contrast when $(\alpha, \beta) = (0.3, -0.7)$. However, further increase in the influence of WXLAI metric results in pattern noise. This proves that even at optimization points where MSE is relatively low, increasing WXLAI influence too much pushes the autoencoder towards configurations to encode and decode images with low level features similar to the original from the metric's perspective.

This interesting phenomenon highlights the adverse effects of WXLAI metric that had not been observed in the previous chapters, where the metric's eminent ability to evaluate the quality of compressed images was demonstrated. However, when the metric is asked to convey information as to how the images need to be for higher scores, it presents what an "ideal" image looks like according to its learned weights. During training of WXLAI metric, there

(a) Original (b) $(\alpha, \beta) = (1.0, 0.0)$ (c) $(\alpha, \beta) = (0.3, -0.7)$ (d) $(\alpha, \beta) = (0.3, -70.0)$

Figure 9.4 – Cropped section of kodim15 from Kodak image database, compressed at target rate 1.0bpp before cropping using CAE-LM.



(a) Original (b) $(\alpha, \beta) = (1.0, 0.0)$ (c) $(\alpha, \beta) = (0.3, -0.7)$

Figure 9.5 – Cropped section of kodim15 from Kodak image database, compressed at target rate 0.5 before cropping using CAE-LM.

was no side information as to inform the metric that score 5.0 is the maximum attainable score, and the highest ranking needs to be assigned when reference image is compared to itself. The fact that the metric was able to evaluate the quality of previously tested compressed images accurately does not necessarily imply that it has a unique understanding of what

the characteristics of a high quality image are. When the metric is asked to impose such characteristics, the quality of results diverge from subjective opinion.

9.3 Conclusion

In this chapter, the anticipated combination of a learning-based compression solution and a learning-based objective metric was experimented. The WXLAI objective metric, with its demonstrated ability to evaluate the quality of compressed images compared to the reference, was used as the distortion measure in the loss function of SingleNet codec, which was preferred mainly due to its adequate trade-off between computational complexity and quality performance. To the author's best knowledge, the use of a learned metric as the autoencoder loss is a novelty in the field and had not been published previously. The contributions of this chapter are emphasized as follows:

- The most evident outcome of the experiments was the detection of a defect in the operation of WXLAI metric. This shortcoming did not present itself previously during the evaluation of compressed images, but was apparent when WXLAI metric was employed as part of encoding solutions. Although the training database was improved in Chapter 5 and in turn the metric's performance was enhanced also on cross-database evaluations, the training database turned out to be not as comprehensive as needed for the autoencoder loss. WXLAI metric was shown to have high sensitivity but low precision when assessing the quality of compressed images in a full-reference framework. It was validated that the training database needs to include more instances such as different contents paired as inputs and much higher levels of distortion, as well as specific limits as to explicitly convey the characteristics of image pairs with minimum and maximum ratings.

- In order to overcome this adversarial effect, MSE loss was added as an additional cost to the loss function so that minimization of average pixel-wise differences could be taken into account. It was validated that the loss function weighed by MSE and WXLAI metric improved the objective performance of CAE-LM in terms of PSNR and WXLAI, however, subjective differences were too subtle to be conclusive on the precise effect of WXLAI.

The shortcomings encountered in this chapter need to be addressed thoroughly. As stated previously, the training set of any objective metric needs to be modified in order for the metric to be robust as autoencoder loss and enhance codec performance objectively and subjectively. The proposed model needed tuning parameters α, β and γ that need to be adjusted during training. Moreover, a dynamic calibration of α and β was necessary for WXLAI feedback to take effect robustly. The ideal metric for CAE-LM configuration needs to eliminate the need for such calibration, which can be achieved through extended training sets and side information on minimum and maximum ratings.

One future step is to investigate the effect of wavelet decomposition in the CAE-LM setting. An immediate future work is to check the performance of an objective metric that operates on pixel values, such as WaDIQaM-FR, in a similar CAE-LM architecture. Similarly, the performance of CAE-LM architecture needs to be tested using both the autoencoder and the metric employing wavelet decomposition as a preprocessing step, such as ResWCAE and WXLAI metric. Clearly, the wavelet decomposition plays an important role in the extraction of features from the reference and decoded images at the metric side, which in turn changes the weights of the autoencoder as to attain the desired similarity measure of the metric. If both autoencoder and metric operate on wavelet coefficients, the adversarial effect of WXLAI may be reduced. In the opposite setting where wavelet decomposition is discarded from both autoencoder and metric, the pattern noise encountered throughout presented experiments may be avoided. However, without the extensions on the database and necessary side information concerning quality range, introduction of noise could still be influenced, yet the characteristics of this noise would be different.

During training, the metric extracted a number of corresponding patches from the source and distorted image. Features of patches from images of multiple resolutions were extracted using 3×3 kernels. SingleNet, on the other hand, employed 5×5 kernels for encoding and decoding images of various resolutions. Although both networks were trained using SD to UHD contents, the scaling differences between the autoencoder and metric layers may influence the performance of CAE-LM. Selecting different combinations of kernel sizes are likely to affect the performance of the proposed model. Alternatively, more complex solutions such as mixed kernel sizes may be explored to involve multiscale processing both at the autoencoder and the distortion metric.

Part IV Conclusion and future directions

10 Conclusions

This thesis addressed the challenges of objective quality assessment and compression of images using machine learning. Convolutional neural networks were employed to construct and present novel solutions that enhance the state of the art on the selected topics. The major contributions of this dissertation are outlined in this chapter, followed by proposals concerning future work on the areas of interest.

10.1 Outcomes and accomplishments

Three major topics of research were identified in this thesis:

1. Learning-based image quality assessment, particularly focusing on the objective quality evaluation of compressed images,
2. Learning-based image compression and analysis of factors affecting autoencoder performance,
3. Preliminary study towards a unified autoencoder that employs a robust, learning-based objective metric.

A summary of the outcomes and accomplishments related to each topic is provided in the following subsections.

10.1.1 Learning-based image quality assessment

1. A novel image quality assessment model, WIQM, was presented, which employs 2D wavelet decomposition as a preprocessing step. The advantages of using such preprocessing were demonstrated. WIQM model was shown to be superior to other state-of-

the-art learning-based IQM modules on TID2013 test set. The performance of WIQM was also compared to an equivalent model without the pixel valued inputs and the representation ability of wavelet coefficients for IQA was verified. Although the computational complexity of WIQM was higher than WaDIQaM-FR, its convergence time was shorter. Analysis on the low generalization ability of WIQM showed that the lack of mapping between cross-database distortion levels and scores result in poor prediction ability across datasets.

2. A novel IQA database composed of 15 references of resolutions SD to UHD was constructed. Each SDR reference was compressed at various bitrates using 16 different state-of-the-art compression algorithms, of which 6 were learning-based. In total, 571 distorted SDR images were rated by subjects using DSIS Variant I experiments. An open source framework for IQA was created for analysis of the results, and to be used in future IQA tasks. The performance analysis of state-of-the-art codecs involved in the experiments was presented. Although not exploited in this work, the evaluations included one 10-bit SDR content and 8 HDR contents with 12-bit depth at HD resolution. This additional part of the dataset may also be used for IQA on higher bit-depth images and HDR contents.

3. An improved IQM was presented, which was trained using the novel XLAI database on an architecture similar to that of WIQM, but with approximately 6 times reduced computational complexity. The WXLAI metric was shown to have good generalization ability across databases and was robust to changes in resolution. The correlation of WXLAI ratings with subjective scores on TID2013 database was almost as high as that of metrics trained on the database, and higher than other state-of-the-art metrics employing hand-crafted features. The performance of WXLAI metric on XLAI test set was much better than other learning-based metrics tested, and WXLAI was challenged mostly by SSIM and MS-SSIM, with minor differences between the scores.

10.1.2 Learning-based image compression

4. A novel convolutional autoencoder for image compression at low rates, i.e. 0.15bpp, was proposed, involving a 3-scale 2D wavelet decomposition as a preprocessing step. The model was trained end-to-end, using MSE in the loss function. The proposed WCAE model outperformed JPEG and NoWCAE, a similar method that excludes the wavelet decomposition step, in terms of MS-SSIM across the full CLIC2019 test set. WCAE did not suffer from blocking artifacts and was able to preserve details better, however introduced high frequency noise. The contribution of different wavelet scales to compression performance was analyzed and the possibility of achieving different quality trade-offs at a rate point was demonstrated by changing the number of outputs of scales.

5. The low rate compression model was extended to achieve higher number of rate-distortion trade-offs by using a wider bottleneck. Additional layers that deepen the network were employed and skip connections were used to improve the performance. Kernel sizes of branches processing wavelet coefficients in ResWCAE model were modified in ResMixWCAE to process the coarsest scale with smaller kernels and increase the convolutional window towards finer scales. At low bitrates, neither of the two models were as efficient as WCAE. Although residual connections and deeper networks helped the performances of autoencoders operating with pixel valued inputs, it was validated that such design in the presented models amplified high frequency noise at low rates. At higher bitrates, the performances were objectively and subjectively comparable to JPEG and JPEG 2000, with ResMixWCAE performing inferior to ResWCAE. Demonstrations proved that when extracting features from wavelet coefficients of finer scales, using larger filters with the same output size was counterproductive. A compounded multi-scale feature extraction through wavelet transform and varying kernel sizes was shown to reduce compression performance.

6. Using mixed kernels in compressive autoencoders not accompanied by wavelet transform was found to be more advantageous. For a fixed output dimension, using mixed kernels at analysis and synthesis stages of the autoencoder resulted in a better rate-distortion trade-off and the effect was amplified at higher bitrates. More specifically, noise brought by small kernels were balanced by the receptive field of larger kernels, and details were preserved better when features from distinct kernel sizes were provided. TripleNet was shown to preserve original image features better than the state-of-the-art learning-based image codec FactMSE at the highest target bitrate, despite having a much smaller bottleneck and having been trained on a much smaller dataset.

10.1.3 Towards unified learning-based image compression solutions

7. A convolutional autoencoder that employs a learning-based objective quality metric to assess the distance between reference input and decoded output was implemented. Preliminary results confirmed that WXLAI metric has a high true positive rate at test time, but revealed that it also predicts many false positives when rating query decoded images during autoencoder training. WXLAI influence on the autoencoder loss was able to ameliorate encoder performance when CAE weights were initialized from a pre-trained model as to provide distorted images with characteristics similar to those encountered during training of WXLAI. Weighing the loss function using MSE prevented WXLAI from updating the autoencoder weights towards false positives. The need for a careful adjustment of hyperparameters was highlighted for the presented CAE-LM model to influence the autoencoder efficiently, i.e. for the decoded images to have high visual quality and high WXLAI scores.

10.2 Future directions

The research conducted throughout this work also presented shortcomings that need to be addressed as future work. The insights provided through the chapters would help in improving the topics of interest as follows:

- The performance of the quality metrics presented in this work can be enhanced using several methods. Changing the number of wavelet scales used in the decomposition and analyzing the outputs that are formed as weighted combinations of features of scales is one future direction that needs to be explored. Using saliency models can help not only when weighing patches but also when sampling patches for quality evaluation. Such saliency models may be used to learn an optimal local sampling of patches in terms of quantity, and also to sample patches in a multiscale fashion, i.e. sampling patches of varying dimensions from an image at regions with distinctive saliency measures. Alternatively, the effects of loss functions other than mean squared error, such as batch correlation between predicted scores and subjective ratings, can be explored. In this framework, the patches forming the batch need to be selected carefully.
- Extensions of both learning-based image quality assessment and compression to video is possible by incorporating the temporal dimension into the frameworks. The correlation between consecutive time frames need to be exploited for video compression. As for video quality evaluation, one future experiment is to detect scene changes and compute the video quality as a weighted sum of a number of randomly selected frames from each scene. Again, temporal correlations may be used to determine how to sample the said frames effectively. Additionally, with the help of prediction algorithms, the change of quality across frames can be estimated.
- Another immediate future research direction is towards no-reference image quality assessment. Many learning-based NR-IQA frameworks have been proposed in this domain, which would benefit from applying the methodologies presented in this dissertation, such as the wavelet decomposition preprocessing or multiscale processing of the image through mixed convolutional layers. The use of mixed kernel sizes in NR and FR-IQA instead of wavelet transform is an alternative method for multiscale processing of image features for quality evaluation.
- The precision of the proposed quality metric needs to be enhanced by introducing more examples of bad quality during training, such as the false positives generated by the CAE-LM network. Lower and upper quality limits and conditions need to be conveyed during training. An improved metric with such qualities is expected to perform well within the CAE-LM framework and achieve favorable results.
- Following the above step, the compatibility of the autoencoder and the distortion metric should be ensured by analyzing the effects of input domains (transform features or pixel values) and the impact of kernel sizes of the two networks.

- When enhancing the IQA database for training an IQM metric, ratings that assess how "natural" an image looks compared to the reference can also be included. The overall quality of the image can then be weighed as to avoid artificial looking outputs.
- Changing the "modes" of the autoencoder using the presented architectures is possible. For the wavelet-based encoders with uniform kernels, the features of the image can be adjusted by changing the number of wavelet scales and the output sizes of of scale features. Similarly, for pixel values inputs within mixed convolutional network architectures, the contributions of each unique sized kernel may be adjusted to yield a desired mode of operation. Moreover, these "modes" can be learned through training of labeled instances, e.g. whether attributes such as film grain or blur need to be preserved.
- Both for compression and image quality evaluation, the effect of different color spaces needs to be explored. For example, CIELAB color space employs nonlinear relationships between its components that are intended to mimic human visual system response. Image quality evaluation in this domain is worth exploring. Similarly, chroma subsampling in YCbCr color space can provide additional benefits to autoencoders.
- The computational limitations encountered during this work prevented the use of higher size kernels, wider bottlenecks and deeper architectures in mixed convolutional networks. At the availability of more resources and time, novel configurations, i.e. analysis on different combinations of kernel sizes and bottleneck distribution, should be explored using a larger training set.

This dissertation explored novel approaches to the challenges of objective quality assessment and compression of images using convolutional neural networks. Future directions addressed in this work highlight numerous novel aspects in the field that need to be explored. The presented work may also influence quality evaluation and compression frameworks in other multimedia domains such as omnidirectional imaging, point clouds, light field images, virtual reality and augmented reality.

Bibliography

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., and Gool, L. V. (2017). Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, pages 1141–1151.

Agustsson, E. and Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Van Gool, L. (2018). Generative adversarial networks for extreme learned image compression. *arXiv preprint arXiv:1804.02958*.

Aja-Fernandez, S., Estepar, R. S. J., Alberola-Lopez, C., and Westin, C.-F. (2006). Image quality assessment based on local variance. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 4815–4818. IEEE.

Akbari, M., Liang, J., and Han, J. (2019). Dsslic: deep semantic segmentation-based layered image compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2042–2046. IEEE.

Akyazi, P. and Ebrahimi, T. (2018a). Comparison of compression efficiency between hevc/h. 265, vp9 and av1 based on subjective quality assessments. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

Akyazi, P. and Ebrahimi, T. (2018b). A new objective metric to predict image quality using deep neural networks. In *Applications of Digital Image Processing XLI*, volume 10752, page 107521Q. International Society for Optics and Photonics.

Akyazi, P. and Ebrahimi, T. (2019a). Assessment of quality of jpeg xl proposals based on subjective methodologies and objective metrics. In *Applications of Digital Image Processing XLII*, volume 11137, page 111370N. International Society for Optics and Photonics.

Akyazi, P. and Ebrahimi, T. (2019b). An improved objective metric to predict image quality using deep neural networks. *Electronic Imaging*, 2019(12):214–1.

Akyazi, P. and Ebrahimi, T. (2019c). Learning-based image compression using convolutional autoencoder and wavelet decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Akyazi, P. and Ebrahimi, T. (2019d). A new end-to-end image compression system based on convolutional neural networks. In *Applications of Digital Image Processing XLII*, volume 11137, page 111370M. International Society for Optics and Photonics.

Ascenso, J. and Akyazi, P. (2019). Report on the state-of-the-art of learning based image coding. In *ISO/IEC JTC 1/SC29/WG1 N83058, 83rd JPEG Meeting, Geneva, Switzerland*.

Baig, M. H., Koltun, V., and Torresani, L. (2017). Learning to inpaint for image compression. In *Advances in Neural Information Processing Systems*, pages 1246–1255.

Ballé, J. (2018). Efficient nonlinear transforms for lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pages 248–252. IEEE.

Ballé, J., Laparra, V., and Simoncelli, E. P. (2015). Density modeling of images using a generalized normalization transformation. *arXiv preprint arXiv:1511.06281*.

Ballé, J., Laparra, V., and Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. (2018). Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*.

Bosse, S., Becker, S., Müller, K.-R., Samek, W., and Wiegand, T. (2019). Estimation of distortion sensitivity for visual quality prediction using a convolutional neural network. *Digital Signal Processing*, 91:54–65.

Bosse, S., Maniry, D., Müller, K. R., Wiegand, T., and Samek, W. (2018). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

Cai, J. and Zhang, L. (2018). Deep image compression with iterative non-uniform quantization. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 451–455. IEEE.

Chang, J., Mao, Q., Zhao, Z., Wang, S., Wang, S., Zhu, H., and Ma, S. (2019). Layered conceptual image compression via deep semantic synthesis. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 694–698. IEEE.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2018). Deep convolutional autoencoder-based lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pages 253–257. IEEE.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019a). Deep residual learning for image compression. *arXiv preprint arXiv:1906.09731*.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019b). Energy compaction-based image compression using convolutional autoencoder. *IEEE Transactions on Multimedia*.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2019c). Learning image and video compression through spatial-temporal energy compaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10071–10080.

Cho, S., Lee, J., Kim, J., Kim, Y., Kim, D.-W., Chung, J. R., and Jung, S.-W. (2019). Low bit-rate image compression based on post-processing with grouped residual dense network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546. IEEE.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

CIE, C. (1986). Official recommendations of the international commission on illumination. *Publication CIE No. 15.2*.

Clarke, F., McDonald, R., and Rigg, B. (1984). Modification to the jpc79 colour–difference formula. *Journal of the Society of Dyers and Colourists*, 100(4):128–132.

Dang-Nguyen, D.-T., Pasquini, C., Conotter, V., and Boato, G. (2015). Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 219–224. ACM.

De Simone, F., Goldmann, L., Lee, J.-S., and Ebrahimi, T. (2011). Towards high efficiency video coding: Subjective evaluation of potential coding technologies. *Journal of Visual Communication and Image Representation*, 22(8):734–748.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Dufaux, F., Sullivan, G. J., and Ebrahimi, T. (2009). The jpeg xr image coding standard [standards in a nutshell]. *IEEE Signal Processing Magazine*, 26(6):195–204.

Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., and Carli, M. (2006). New full-reference quality metrics based on hvs. In *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, volume 4.

Gao, F., Wang, Y., Li, P., Tan, M., Yu, J., and Zhu, Y. (2017). Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

ITU-R BT.2022 (August 2012). General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays. International Telecommunication Union.

ITU-R BT.2100 (July 2018). Image parameter values for high dynamic range television for use in production and international programme exchange.

ITU-R BT.500-13 (January 2012). Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union.

ITU-R BT.814-4 (2018). Specifications of pluge test signals and alignment procedures for setting of brightness and contrast of display. International Telecommunication Union.

ITU-R BT.815-1 (1994). Bt.815 : Specification of a signal for measurement of the contrast ratio of displays. International Telecommunication Union.

ITU-T P.910 (April 2008). Subjective video quality assessment methods for multimedia applications. International Telecommunication Union.

Jayaraman, D., Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). Objective quality assessment of multiply distorted images. In *2012 Conferene on Signals, Systems and Computers (ASILOMAR)*, pages 1693–1697. IEEE.

Jiang, J. (1999). Image compression with neural networks–a survey. *Signal processing: image Communication*, 14(9):737–760.

Jin, L., Lin, J. Y., Hu, S., Wang, H., Wang, P., Katsavounidis, I., Aaron, A., and Kuo, C.-C. J. (2016). Statistical study on perceived jpeg image quality via mcl-jci dataset construction and analysis. *Electronic Imaging*, 2016(13):1–9.

Kim, D.-W., Ryun Chung, J., and Jung, S.-W. (2019). Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lainema, J., Hannuksela, M. M., Vadakital, V. K. M., and Aksu, E. B. (2016). Hevc still image coding and high efficiency image file format. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 71–75. IEEE.

Laparra, V., Muñoz-Marí, J., and Malo, J. (2010). Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A*, 27(4):852–864.

Larson, E. C. and Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Lee, J., Cho, S., and Beack, S.-K. (2018). Context-adaptive entropy model for end-to-end optimized image compression. *arXiv preprint arXiv:1809.10452*.

Li, Zhi, A. A. K. I. M. A. and Manohara, M. (2019). Toward a practical perceptual video quality metric. *The Netflix Technology Blog*.

Li, M., Zuo, W., Gu, S., Zhao, D., and Zhang, D. (2018). Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3223.

Li, S., Zhang, F., Ma, L., and Ngan, K. N. (2011). Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13(5):935–949.

Li, Z., Norkin, A., and Aaron, A. (2016). Vmaf-video quality metric alternative to psnr. *Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11*.

Lin, W. and Kuo, C.-C. J. (2011). Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312.

Luo, M. R., Cui, G., and Rigg, B. (2001). The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350.

Ma, H., Liu, D., Xiong, R., and Wu, F. (2019). A cnn-based image compression scheme compatible with jpeg-2000. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 704–708. IEEE.

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., and Zhang, L. (2016). Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016.

Mannos, J. and Sakrison, D. (1974). The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536.

Marpe, D., Schwarz, H., and Wiegand, T. (2003). Context-based adaptive binary arithmetic coding in the h. 264/avc video compression standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):620–636.

Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., and Van Gool, L. (2018). Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4394–4402.

Minnen, D., Ballé, J., and Toderici, G. D. (2018a). Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10771–10780.

Minnen, D., Toderici, G., Singh, S., Hwang, S. J., and Covell, M. (2018b). Image-dependent local entropy models for learned image compression. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 430–434. IEEE.

Mitsa, T. and Varkur, K. L. (1993). Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 301–304. IEEE.

Moorthy, A. K. and Bovik, A. C. (2011). Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364.

Nah, S., Kim, T. H., and Lee, K. M. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3.

Narwaria, M. and Lin, W. (2011). Svd-based quality metric for image and video using machine learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):347–364.

Nemoto, H., Hanhart, P., Korshunov, P., and Ebrahimi, T. (2014). Ultra-eye: Uhd and hd images eye tracking dataset. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 39–40. IEEE.

Nguyen, T. and Marpe, D. (2012). Performance analysis of hevc-based intra coding for still image compression. In *2012 Picture Coding Symposium*, pages 233–236. IEEE.

Nill, N. (1985). A visual model weighted cosine transform for image compression and quality assessment. *IEEE Transactions on Communications*, 33(6):551–557.

Ohm, J.-R. and Sullivan, G. J. (2018). Versatile video coding–towards the next generation of video compression. In *Picture Coding Symposium 2018*.

Pei, S.-C. and Chen, L.-H. (2015). Image quality assessment using human visual dog model fused with random forest. *IEEE Transactions on Image Processing*, 24(11):3282–3292.

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77.

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F. (2009). Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45.

Prakash, A., Moran, N., Garber, S., DiLillo, A., and Storer, J. (2017). Semantic perceptual image compression using deep convolution networks. In *2017 Data Compression Conference (DCC)*, pages 250–259. IEEE.

Prechelt, L. (1998). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Richter, T., Artusi, A., and Ebrahimi, T. (2016). Jpeg xt: A new family of jpeg backward-compatible standards. *IEEE Multimedia*, 23(3):80–88.

Rippel, O. and Bourdev, L. (2017). Real-time adaptive image compression. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2922–2930. Journal of Machine Learning Research.

Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., and Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, 18(11):2385–2401.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Schaefer, G. and Stich, M. (2003). Ucid: An uncompressed color image database. In *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307, pages 472–480. International Society for Optics and Photonics.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Sheikh, H. (2005). Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*.

Sheikh, H. R. and Bovik, A. C. (2004). Image information and visual quality. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages iii–709. IEEE.

Sheikh, H. R., Bovik, A. C., and De Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Sullivan, G. J., Ohm, J., Han, W.-J., and Wiegand, T. (2012). Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and systems for Video Technology*, 22(12):1649–1668.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Tan, M. and Le, Q. V. (2019). Mixnet: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*.

Taubman, D. and Marcellin, M. (2012). *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice*, volume 642. Springer Science & Business Media.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. (2017). Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*.

Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., Covell, M., and Sukthankar, R. (2015). Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*.

Toderici, G., Vincent, D., Johnston, N., Jin Hwang, S., Minnen, D., Shor, J., and Covell, M. (2017). Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5306–5314.

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. (2019). Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*.

Tschannen, M., Agustsson, E., and Lucic, M. (2018). Deep generative models for distribution-preserving lossy compression. In *Advances in Neural Information Processing Systems*, pages 5929–5940.

Wallace, G. K. (1992). The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv.

Wang, J., Tao, X., Xu, M., and Lu, J. (2019). Semantic perceptual image compression with a laplacian pyramid of convolutional networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 699–703. IEEE.

Wang, Z., Bovik, A. C., and Lu, L. (2002). Why is image quality assessment so difficult? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 3313–3316.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Wang, Z. and Li, Q. (2011). Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198.

Wang, Z., Simoncelli, E., and Bovik, A. e. a. (2003a). Multi-scale structural similarity for image quality assessment. In *ASILOMAR Conference on Signals, Systems amd Computers*, volume 2, pages 1398–1402. Citeseer.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003b). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee.

Witt, K. (1995). Cie guidelines for coordinated future work on industrial colour-difference evaluation. *Color Research & Application*, 20(6):399–403.

Zhang, L., Zhang, L., Mou, X., Zhang, D., et al. (2011). Fsim: a feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386.

Zhao, H. and Liao, P. (2019). Cae-admm: Implicit bitrate optimization via admm-based pruning in compressive autoencoders. *arXiv preprint arXiv:1901.07196*.

Zhou, L., Cai, C., Gao, Y., Su, S., and Wu, J. (2018). Variational autoencoder for low bit-rate image compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2617–2620.

Zhou, L., Sun, Z., Wu, X., and Wu, J. (2019). End-to-end optimized image compression with attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Pinar Akyazi

Electrical and electronics engineer *specialized in*	Signal, image and video processing Machine Learning Multimedia Compression & Quality Assessment Biomedical Imaging Communication technologies

Personal Data

Place and Date of Birth:	Ankara, Turkey	30 May 1988
Address:	Avenue d'Ouchy 58, 1006 Lausanne, Switzerland
Phone:	+41 78 3019181
email:	akyazi.pinar@gmail.com

Academic Experience

Jan 2018 - *Present*

Research and Teaching Assistant in EPFL-STI-IEL-MMSPG

Research Interests: Learning-based image compression, learning-based objective quality assessment, subjective quality assessment, digital media

Currently building learning-based models for image compression and quality assessment of videos of different modalities, as part of an InnoSuisse project.

ISO member, taking part in SC29/WG1 as chairwoman of JPEG AI and participating in the development and evaluation of JPEG XL.

Sept 2014 - Dec 2017

Research and Teaching Assistant in EPFL-STI-IEL-LTS4

Research Interests: Signal processing on graphs, multiview systems, immersive communication, image processing

Developed graph based methods to improve free viewpoint navigation systems in 3D environments, as part of a SNSF funded project.

Sept 2010 - Aug 2013

Research and Teaching Assistant in Bogazici University, BUSIM/VAVLab

Successfully carried out joint research in departments of Electrical and Electronics Engineering and Biomedical Engineering, on revealing the heterogeneous characteristics of force distribution along skeletal muscle fibers on human leg.

Aug 2010

Summer project in Brno University of Technology

Project title: "Building an Open Source Speech Recognition Toolkit based on Subspace Gaussian Mixture Models (SGMMs) using Weighted Finite State Transducers"

Took part in the design of Kaldi, a free, open-source toolkit for speech recognition research.

Jun - July 2009

Summer workshop in Johns Hopkins University, CLSP

Project title: "Low Development Cost, High Quality Speech Recognition for New Languages and Domains"

Attended a two month workshop on building acoustic models for speech recognition and automatic lexicon learning.

Work Experience

Feb 2013

Intern at Vistek ISRA Vision, Istanbul

Worked on developing algorithms for artificial vision.

Jun - July 2008

Intern at iSEC-Siemens, Ankara

Worked on enhancing models for communication networks.

Education

Sept 2014 - *Present* PhD Candidate in Electrical and Electronics Engineering
Swiss Federal Institute of Technology, Lausanne
Thesis title: "Image Compression and Quality Assessment using Deep Convolutional Neural Networks" | Advisor: Prof. Touradj Ebrahimi
Our goal is to develop a learning-based codec that employs a learning-based objective quality metric in the training objective. We have built a convolutional autoencoder using PSNR and MS-SSIM in loss function, as well as a CNN-based objective metric that is able to predict image quality highly correlated with subjective ratings.

Sept 2010 - Aug 2013 Master of Science Degree in Electrical and Electronics Engineering
Bogazici University, Istanbul
Thesis: "Diffusion Tensor Field Deformations Under Active and Passive Stretching of Skeletal Muscles" | Advisors: Prof. Burak Acar, Prof. A. Can Yucesoy
By analyzing the diffusion statistics and strain distribution, we have confirmed the heterogeneous strain distibution along skeletal muscle fibers.
Gpa: 3.66/4.00

Sept 2006 - June 2010 Bachelor of Science Degree in Electrical and Electronics Engineering
Bogazici University, Istanbul
Senior project: "Optimization of Voxel Similarity Measures for Graph Theoretic Segmentation of Liver Lesions" | Advisor: Prof. Burak Acar
General training in electrical and electronics engineering with an emphasis on telecommunications, signal and image processing, biomedical image processing.
Gpa: 3.46/4.00

Sept 2005 - May 2006 Bachelor student in Electrical and Electronics Engineering
Bilkent University, Istanbul
Gpa: 3.79/4.00

Skills

Programming languages:	Python, C++, MATLAB, C, Java
Machine learning platforms:	PyTorch, Tensorflow, Keras
Cloud platforms:	Docker, Azure
Open access platforms:	GitHub, Authorea, Jupyter
Circuit design:	MicroSim, PSpice, LabView, ModelSim, Xilinx

Languages

Full working proficiency:	**English** (C2)
Limited working proficiency:	**French** (B1), **Italian** (B1)
Mother tongue:	**Turkish**

Interests and Activities

Passionate about open science and social entrepreneurship
Amateur pianist for twenty years
Love running, recently ran the 20km race in Lausanne

Publications

Journal articles

- Povey, Daniel, Lukás Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondrej Glembek et al. "The subspace Gaussian mixture model—A structured model for speech recognition." *Computer Speech & Language* 25, no. 2 (2011): 404-439.

Conference proceedings

- Akyazi, Pinar, and Touradj Ebrahimi. "A new end-to-end image compression system based on convolutional neural networks." Applications of Digital Image Processing XLII. Vol. 11137. International Society for Optics and Photonics, 2019.
- Akyazi, Pinar, and Touradj Ebrahimi. "Assessment of quality of JPEG XL proposals based on subjective methodologies and objective metrics." Applications of Digital Image Processing XLII. Vol. 11137. International Society for Optics and Photonics, 2019.
- Akyazi, Pinar, and Touradj Ebrahimi. "Learning-Based Image Compression using Convolutional Autoencoder and Wavelet Decomposition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
- Cheng, Zhengxue, Pinar Akyazi, Heming Sun, Jiro Katto, and Touradj Ebrahimi. "Perceptual Quality Study on Deep Learning based Image Compression." *IEEE International Conference on Image Processing (ICIP)*, 2019.
- Upenik, Evgeniy, Pinar Akyazi, Mehmet Tuzmen, and Touradj Ebrahimi. "Inpainting in omnidirectional images for privacy protection." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.
- Akyazi, Pinar, and Touradj Ebrahimi. "An Improved Objective Metric to Predict Image Quality using Deep Neural Networks." *IS&T Electronic Imaging Proceedings*, 2019.
- Akyazi, Pinar, and Touradj Ebrahimi. "A new objective metric to predict image quality using deep neural networks." *Applications of Digital Image Processing XLI*. Vol. 10752. International Society for Optics and Photonics, 2018.
- Akyazi, Pinar, and Touradj Ebrahimi. "Comparison of Comparison Efficiency between HEVC/H.264, VP9 and AV1 based on Subjective Quality Assessments." *10th International Conference on Quality of Multimedia Experience (QoMEX)*, 2018.
- Akyazi, Pinar, and Pascal Frossard. "Graph-Based Inpainiting for Zooming in 3D Scenes." In *26th European Signal Processing Conference (EUSIPCO)*, 2018.
- Tzamarias Olivier Eustathios, Dion, Pinar Akyazi, and Pascal Frossard. "A Novel Method for Sampling Bandlimited Graph Signals" In *26th European Signal Processing Conference (EUSIPCO)*, 2018.
- Akyazi, Pinar, and Pascal Frossard. "Graph-Based Interpolation for Zooming in 3D Scenes." In *25th European Signal Processing Conference (EUSIPCO)*, 2017.
- Povey, Daniel, Lukás Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek et al. "Subspace Gaussian mixture models for speech recognition." In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4330-4333. IEEE, 2010.
- Burget, Lukás, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek et al. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models." In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4334-4337. IEEE, 2010.
- Burget, Lukás, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek et al. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models." In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4334-4337. IEEE, 2010.
- Goel, Nagendra, Samuel Thomas, Mohit Agarwal, Pinar Akyazi, Lukás Burget, Kai Feng, Arnab Ghoshal et al. "Approaches to automatic lexicon learning with limited training examples." In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5094-5097. IEEE, 2010.