Høgskolen i Østfold

From Hand-crafted to Self-attention

No-reference or Blind Image Quality Assessment

Master's Thesis

Sourav Paul Roman

School of Computer Sciences
Østfold University College
Halden
June 15, 2023

From Hand-crafted to Self-attention

No-reference or Blind Image Quality Assessment

Master's Thesis

Sourav Paul Roman

School of Computer Sciences
Østfold University College
Halden
June 15, 2023

Abstract

With the exponential growth of the production of digital images, human subjective perceptive quality measurement becomes infeasible when necessary. Maintaining perceptual image quality has also become essential with the overwhelming growth of digital platforms that uses user-generated content such as images, videos, audio, etc. Objective machine image quality measurement tools are the solution at scale. As most of the image content traffic is without any reference or pristine source, no-reference or blind quality measurement becomes the approach with robust applications. The image quality assessment research field has been growing with the challenges that come with the problem. A comparative study on a few state-of-the-art algorithms trained, validated, and tested across several datasets will be conducted in this research.

Source code `https://github.com/sourav-paul/hc2sa`

Keywords: Image Quality Assessment, No-reference Image Quality Assessment, Blind Image Quality Assessment, Machine Learning, Deep Learning, Transformer, Vision Transformer, Support Vector Regression

Acknowledgments

I would like to thank my thesis advisors Lars Vidar Magnusson and Roland Olsson for their help in the different stages of the thesis. Their invaluable help and sound guidance were essential to this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the rapid growth of information technology over the last couple of decades, almost every part of our lives started growing deep ties with some variation of a technological solution. Some of these solutions have disrupted their previous generation. The substantial gap between the generation of solutions made it clear the feasibility of using the new invention over the previous one. As a result, these solutions became the most common choice in our day-to-day lives.

Digital images and videos have been one of these solutions that replaced the previous generation of silver-halide(AgX) film photography. From filming to post-production processing is time-consuming and costly at every step with analog cameras. Digital cameras made the filming part way easier as well as post-production processing. With the growth of smartphone usage, digital camera in smartphones accelerated digital photography on a wide scale of the global population.

User-generated content has become an essential part of the digital age. From tiny tweets to ultra-high-resolution videos are being produced by users all over the world in almost all major content-sharing platforms. With the convenience of smartphone usage and social media booming, user-generated images have become one of the most used communication media. This resulted in image production to a scale that became beyond quality control by human subjects.

By nature, image acquisition, processing, and transmission are subject to distortions fairly easily. Depending on the production device, producer, and environment of images can be easily distorted at the stage of acquisition. Even with a pristine quality image production, processing stages such as compression can cause distortion as well. Depending on the transmission methods and bandwidth, the transmitted image data can have further distortions.

For the platforms that prefer having quality control of the provided images, human subjective quality measurement of the images has become infeasible with the volume of image production. And the majority of the image data that comes to the systems is without any reference to their pristine quality. The systems get whatever the user provided without any indication of quality measurement parameters. An objective image quality measurement system can resolve the issue of replacing human subjects and can deal with the scale of

production. In addition, this objective image quality measurement system does the task blindly, also known as no-reference or without reference to it's pristine quality source image.

There has been a decent amount of research conducted in the field of image quality assessment. From classical hand-crafted features extraction with support vector regressor (SVR) to recent day's self-attention with Transformer and/or Vision Transformer. The approaches have changed with time as newer, more cutting-edge methods were invented. Image quality assessment datasets have also evolved with time. Initially, most of the distortions were artificially created with a limited amount of distortion types with a smaller number of images, then larger, with more distortion types, in-the-wild datasets were introduced by the research community over time. Most of these researches in the field were done with one or more datasets for training and validation of the method in these researches.

In this research, training some of the state-of-the-art image quality assessment models across datasets, validating, and benchmarking the results will be done. The purpose of this research is to give the research community and application engineers a solid ground for extending further research and choosing the right models for applications respectively.

1.1 Motivation

In March 2023, smartphone users reached over 6.92 billion[1], and they are increasing every day. These increasing numbers indicate the major convenience and connected world with it. And the major part of these connections is user-generated images making them easy to express, and 1.83 trillion[2] images are being taken every year.

With the incremental volume of production of images, subjective quality control becomes infeasible even for a small portion of the samples. Platforms can overrun almost right after accepting user-generated images. Even though the most modern digital cameras, smartphone cameras come with out-of-the-box tools to take a good photo, put a filter on, and make it look great, there could be distortions that might easily be overseen by an individual producer of the source image. On the other hand, after the post-production of pristine-quality source images, they can be distorted over the transmission media and methods, and the recipient's media consumption specifications.

The impact of distorted images comes in a variety of outcomes. Let's consider a few cases where image and video perceptual quality are essential. On top of the list, are news portals, TV channels, digital advertising, and media content streaming services. News portals without image and video quality control may lose the credibility of the news itself. TV channels without media content visual quality control may lose the audience of the contents. Targeted digital advertising with a bad image and video perceptual quality may lose the targeted audience and create a negative user experience regarding the quality of the advertised product itself. Media content streaming services may lose their premium users due to the poor quality of the streamed content.

[1]https://www.bankmycell.com/blog/how-many-phones-are-in-the-world#part-1
[2]https://photutorial.com/photos-statistics/

As images tend to be subject to degradation in their acquisition, processing, and transmission stages, only one point of quality control might not resolve the issue at it's core. For pristine source image quality to be maintained throughout the life-cycle of the image, solutions must be placed at the production of the image in camera firmware, the processing stage of compressing at the serialization stage of the data, and both transmission and receiving point of the image data if it's subject to transport.

Solutions to these problems can be mitigated with objective perceptual quality assessment systems on the points of quality degradation. Objective image quality assessment can be achieved with a machine-learning solution and be placed on those points. Building the image quality measurement on machine learning tools makes it feasible to create, extend, update, and deploy at possible degradation points. A decent amount of research has been conducted in this field. The trends of related research have changed with time. Most of the initial research is related to hand-crafted feature extraction in regard to natural scene statistics(NSS) of images using some variety of support vector regression(SVR). Then related research took a turn in consideration with the powerful convolutional neural network(CNN) approaches. Then Vision-Transformer(ViT) came into the picture with the self-attention mechanism for images, and research trends in image quality assessment moved a bit toward that. All these approaches have high-performing machine-learning algorithms and available datasets at the time of invention.

With the evaluation of the research approaches and available datasets at the time of research, there is a lack of comparative studies of the permuted performance of cross-dataset cross-algorithm. This research approaches to train, validate, and benchmark high-performing algorithms across datasets with various distortion types.

1.2 Research Objective

As high-quality perceptual images are essential to the majority of digital platforms, and human subjective quality control's being infeasible at scale, objective machine solutions are the direction to begin with. Research fields on machine learning solutions that analyze the objective quality assessment of perceptual images are large enough pools. This research will narrow down the scope toward a comparative study of machine learning algorithms for objective perceptual image quality assessment across datasets.

Depending on the availability of the source or reference images or features, samples can be seperated in three categories. Full-reference (FR), no-reference or blind (NR/B), and reduced-reference (RR), and resulting image quality assessment (IQA) approaches are full-reference images quality assessment (FRIQA), no-reference or blind image quality assessment (NR/BIQA), and reduced-reference image quality assessment (RRIQA). Due to larger scope of NR/BIQA, this research stay in the search scope of NR/BIQA.

A comparative research on NR/BIQA will require a collection of existing algorithms and datasets. The existing machine learning algorithms that objectively do the NR/BIQA task can be found from the publications from the research community on the field. This research will choose from the variety of approaches such as hand-crafted feature extraction, convolution neural networks, and transformers implemented state-of-the-art algorithms.

While choosing from the publicly available datasets for IQA, the research will be scoped towards finding the image quality datasets with robust distortion types, in-the-wild/natural distortions, higher number of samples.

This research targets to make a comparison among the algorithms across the image datasets. This study of comparison will help the research community choose from the existing algorithms and datasets, as well as help study further when new algorithms and new datasets are created.

Research Objective

How do the state-of-the-art no reference or blind image quality assessment machine learning algorithms perform across various image quality datasets?

1.3 Method

There are several steps that need to be taken from the start to the end of the research. The method of comparing algorithms across datasets will need research on image quality datasets and their specifications, and algorithms and their specifications.

- Identifying the research objective
- Literature reviews and case studies
- Collect the existing implementations from the related works
- Take note of the relevant parameters from the results
- Analyze the results
- Comparison report with standard co-relation coefficients (Pearson, Spearman's, etc)

1.4 Deliverables

The tangible outcome of the research will include the thesis article, the machine learning algorithms in Python hosted on Github, and image quality dataset references.

1.5 Report Outline

Chapter 2 will deal with the analysis and study of related works. Chapter 3 will go through planning and designing methods. Chapter 4 will discuss implementations and Chapter 5 will deal with the results of the implementation and evaluation of the results. Chapter 6 will contain a technical discussion of the thesis overall. And chapter 7 will conclude the thesis with the thesis achievements and future scopes of the topics.

Chapter 2

Background

As images have become one of the most used representations and communications media of information with the growth and ease of digital image production [Bovik and Wang 2006]. Digital image processing and communication also evolved [Bovik 2005] to improve the appearance of images. By the nature of image data, they are subject to distortion during any of the stages from acquisition to transmission to display [Bovik and Wang 2006].

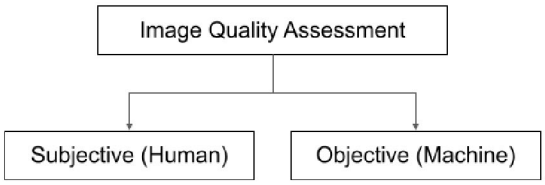


Figure 2.1: Image quality assessment by assessors

2.1 Subjective Image Quality Assessment

Subjective image quality assessment is involved human decisions on quality measurements of perceptual images. This becomes extremely complicated as a different subject might have a different opinion on the same image with the same distortions. Also, with the incremental amount of images in the world, it becomes infeasible soon to keep up with subjective quality measurements of images [Bovik and Wang 2006; Hussein AL-Qinani 2019].

2.2 Objective Image Quality Assessment

Objective quality measurement is the feasible way to ensure the quality of the visual representation of information when the data volume is high. Classification of objective image quality assessment relates to the availability of the original image [Bovik and Wang 2006].

2.2.1 Full-reference Image Quality Assessment

In the full-reference image quality assessment process, the machine is provided with degraded images with distortions and the subsequent pristine distortion-free images as reference.

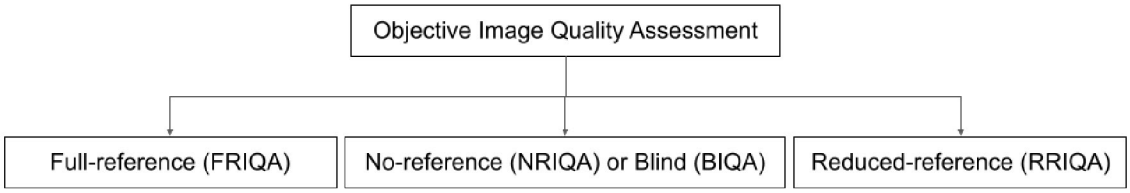


Figure 2.2: Classification of objective image quality assessment.

The system learns from the difference between the pair of images with distortion and distortion-free reference. A few top-down and bottom-up approaches such as implementing a human visual system with structural similarity index (SSIM), structural dissimilarity metric (DSSIM), mean structural similarity index metric (MSSIM), mean square error (MSE), peak signal to noise ratio (PSNR), peak mean square error (PMSE), maximum difference (MD), average difference (AD) [Hussein AL-Qinani 2019; Bovik and Wang 2006].

2.2.2 No-reference or Blind Image Quality Assessment

The availability of reference images might not always be expected. In this case, the machine learns from the image itself. Feature extraction from degraded images could be done in various ways. Such as learning natural scene statistics [Bovik and Wang 2006], saliency-guided and other CNN-based approaches, self-attention, etc 3.

2.2.3 Reduced-reference Image Quality Assessment

It could be also a case where there are original images but no way on the quality measurement side to get the whole image but some features of the original images can be transmitted. These are not the original image but are reduced to some part of the original image features that help measure the quality of the images [Bovik and Wang 2006].

2.3 Objective No-reference or Blind Image Quality Assessment

Due to the unavailability of the reference image in almost all the perceptual image-sharing platforms, no-reference image quality measurement involves learning about known distortion types such as white noise, Gaussian blur, fast-fading, JPEG compression, JPEG2000 compression, and many more. The image quality measurements could target one of these specific types of distortions or could be general-purpose [Bovik and Wang 2006].

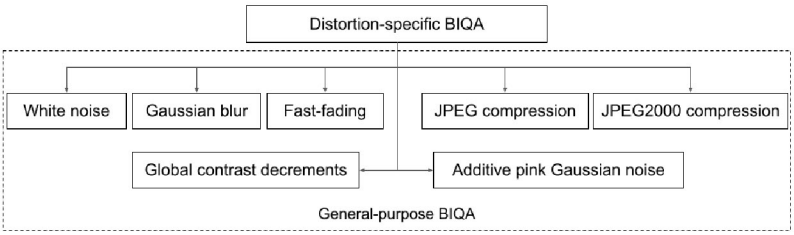


Figure 2.3: Distortion-specific and general-purpose BIQA

Figure 2.4: Different types of distortions. JPEG compression (top-left), Gaussian blur (top-right), Poisson/White noise (bottom-left), DeltaE Gamut Mapping (bottom-right)

2.4 Research Towards NR/B Image Quality Assessment

To solve this and similar problems, various research has been conducted in the machine learning domain of objective image quality assessment. To get more elaborate knowledge based on the domain, a thorough study of existing literature will be done in this research. This research intends to use the results of previous studies in the relevant domain. The research will go through studies done on image quality assessment and will do a result analysis of different datasets. Historical progress in research related to blind image quality assessment can be divided into a few categories [Ma et al. 2022] from an approach standpoint.

Handcrafted feature extraction usually studies the natural scene statistics, extracts feature, and learn image quality from a knowledge-driven process [Bovik and Wang 2006]. In CNN-based solutions, sometimes external knowledge injection is needed, in some other cases global average or maximum pooling to measure quality [Ma et al. 2022]. In Transformer

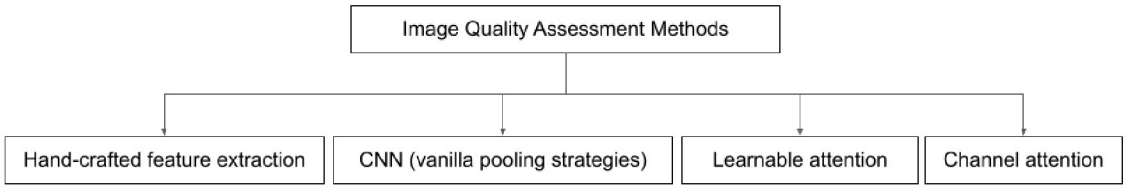


Figure 2.5: Blind image quality assessment methods

based solutions, spatial attention, spatial saliency prediction, channel attention, etc in the core of the approach [Yang et al. 2019; Zhu et al. 2021b; Zhang et al. 2019; Fu et al. 2019].

Support Vector Regression(SVR)

Support vector regression is an extended implementation of classification with support vector machines(SVMs). SVMs are built on supervised learning approaches for classification, regression, and outlier detection problems. SVMs are effective in high-dimensional spaces. It uses a subset(support vectors) of the training point in the decision function, resulting in low memory usage. Different kernel functions can be specified for the decision function, custom kernels can be specified. The core idea of SVMs is to find a maximum marginal hyperplane(MMH) that best divides the datasets into classes.

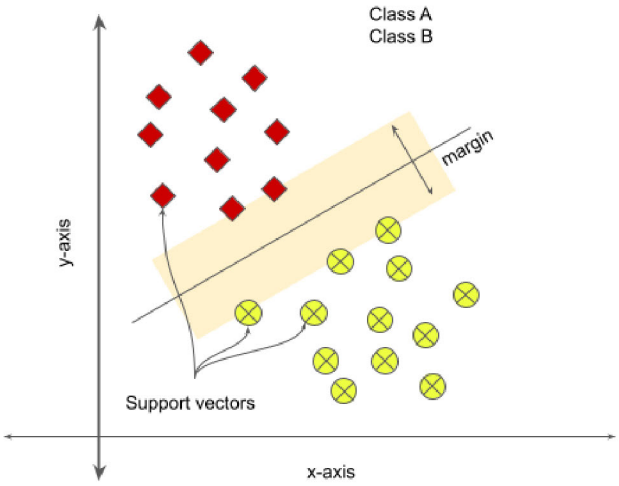


Figure 2.6: Support Vector Machines

The model produced by support vector classification relies only on a subset of the training data because of the cost function for building the model. It does not care about the training points that lie beyond the margin. So, the model generated by the SVR only relies on the subset of the training data because the cost function avoids the sample whose prediction is close to it's target. There are several implementations of SVR such as epsilon-SVR, nu-SVR, and linear-SVR.

SVMs are a powerful tool but their resource requirements such as computing and storage increase rapidly with the number of training vectors. The core of an SVM is a quadratic computation problem, splitting support vectors from the rest of the training data. The quadratic problem solver with LibSVM [Buitinck et al. 2013] scales between $O(n^2)$ $O(n^3)$, where n is the number of samples in the training dataset.

For linear use cases, linear implementation with LibLinear [Buitinck et al. 2013] is significantly more efficient than the LibSVM-based counterpart. And linear implementation can scale linearly to millions of samples and or features. The choices of penalties and loss functions make it more flexible toward supporting a larger number of samples. It's also possible to support both dense and sparse input [Buitinck et al. 2013].

Convolutional Neural Network(CNN)

Due to the nature of computational complexities of high-dimensional image data, traditional artificial neural networks(ANN) tend to struggle towards achieving optimal solutions. Simply increasing the number of neurons with larger weights results in overfitting [O'Shea and Nash 2015].

CNNs are designed primarily to solve image analysis problems, so the architecture of dealing with specific data types was essential. The neurons in the CNN layers are organized into three dimensions, the special dimensionality of the input height and weight, and the third dimension is activation volume. The neurons within any given layer will only be connected to a small portion of the layer ahead of it. The input volume of 64x64x3 will lead to a final output layer comprised of a dimensionality of 1x1xn, where n is the number of classes [O'Shea and Nash 2015].

CNNs are built on three types of layers. They are the convolutions layers, pooling layers, and fully-connected layers. All these layers stacked together form a CNN architecture.

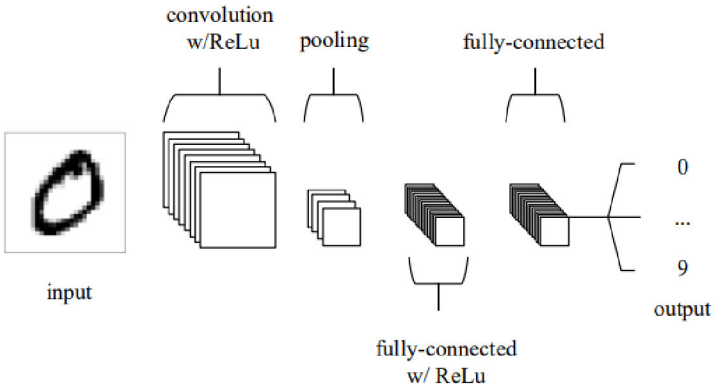


Figure 2.7: A simple 5-layer CNN architecture[O'Shea and Nash 2015]

The input layer holds the pixel values of the image. The convolutional layer determines the output of neurons that are connected to the local region of the input through the

calculation of the scalar product between their weights and the region connected to the input volume. The rectified linear unit(ReLu) is applied at each element as an activation function such as sigmoid to the result of the activation generated by the previous layer. After that the pooling layer simply performs a downsampling along the spatial dimensionality of the given input, reducing the number of parameters with the activation even further. In the next step, the fully-connected layers try to generate class scores from the activations and make use of them for classification. ReLu might be used among these layers to improve performance [Gua et al. 2017; O'Shea and Nash 2015].

A convolution is a mathematical operation that helps to derive the distribution of a sum of two random variables from the distributions of the two summands. The convolution is obtained by summing a series of products of the probability mass function of the two variables when the variables are discrete and random. If they are continuous, integration of the product of their probability density function is applied [Taboga 2021].

Convolutional layers are essential to CNNs. The layers' parameters focus on the use of learnable kernels. The kernels are smaller in spatial dimensionality compared to the whole image. They are spread along the entirety of the depth of the input. When input data hits a convolutional layer, the layer convolves each filter across the whole spatial dimensionality of the input to produce a 2D activation map. These activation maps can be visualized. Gliding through the input, the scalar product is calculated for each value in the kernel. The activations imply the network learning the kernels that trigger when they see a specific feature at a given spatial position on the input [Li et al. 2020; Gua et al. 2017; O'Shea and Nash 2015].

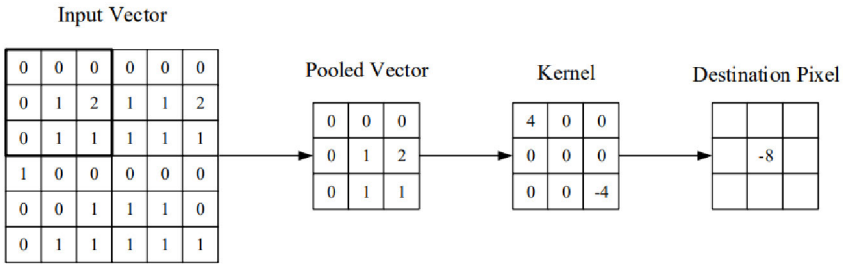


Figure 2.8: A convolutional layer [O'Shea and Nash 2015]

The pooling layers target to decrease the dimensionality of the representation gradually which decreases the amount of parameters and computational complexities of the model. They operate on each activation map and scale the dimensionality with a max function. Overlapping-pooling layers are utilized alongside max-pooling. The fully connected layers contain neurons that are connected directly to the neurons in the two nearest layers only [Li et al. 2020; Gua et al. 2017; O'Shea and Nash 2015].

Transformer and Vision Transformer

A transformer is a type of deep learning neural network model that approaches it's adoption of self-attention techniques differently. It's sequence-to-sequence architecture transforms a

given sequence of elements, such as words in a sentence into another sequence. It's excellent at natural language processing tasks such as translation, language modeling, etc overcoming the shortage of recurrent neural networks(RNNs), rated recurrent neural network(GR), and long short-term memory(LSTM) [Vaswani et al. 2017]. With the growth of self-attention mechanisms, machine learning tasks related to image processing are also introduced with transformers related to computer vision [Parmar et al.; Zhang et al. 2021; Arnab et al. 2021; Bao et al. 2022; Fan et al. 2021].

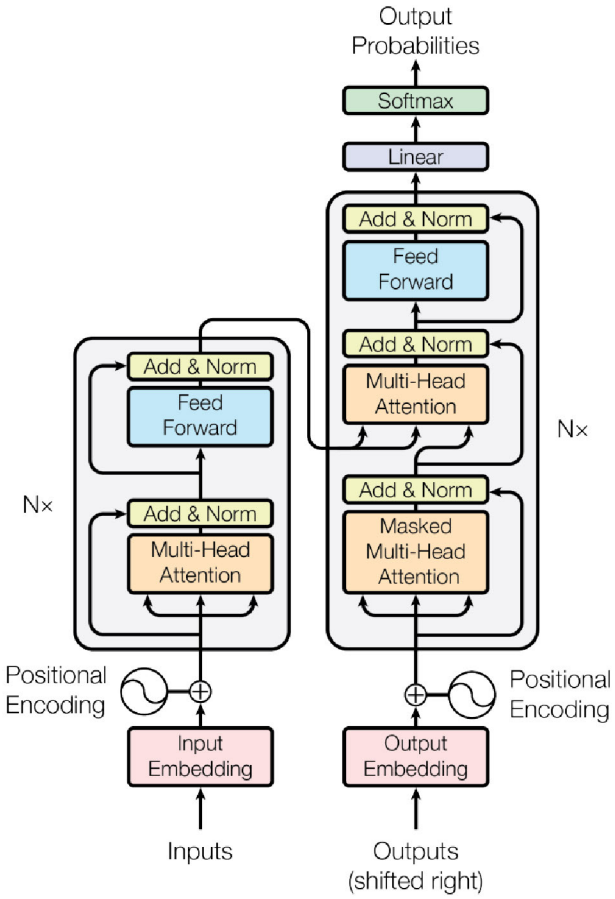


Figure 2.9: Building blocks of a Transformer [Vaswani et al. 2017]

With the goal of decreasing sequential computation, many models were introduced to utilize modern parallel processing systems for all input and output. Self-attention mechanism relates to different positions of a single sequence to compute the representation of a sequence. Encoder-decoder is an essential part of highly efficient neural sequence transducer models.

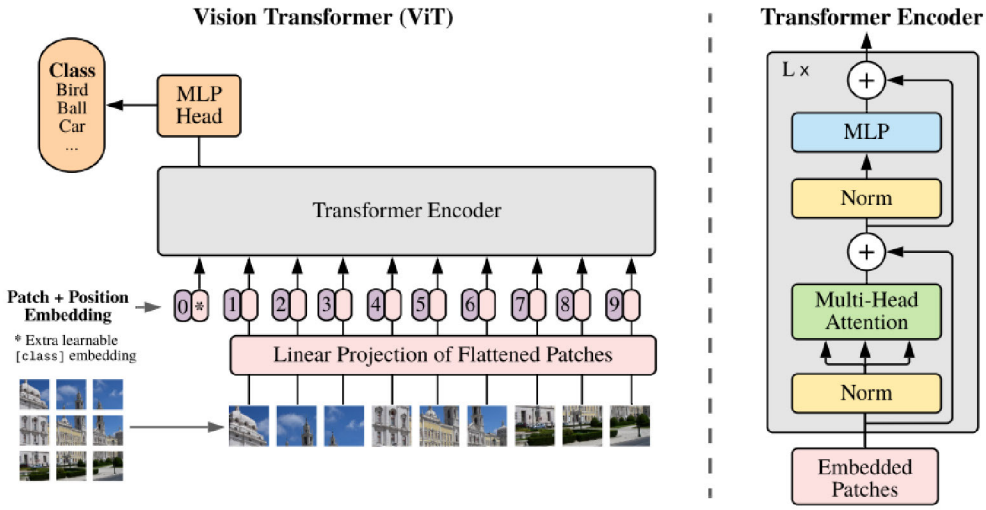


Figure 2.10: Vision Transformer [Dosovitskiy et al. 2021; Vaswani et al. 2017]

In the proposed Vision Transformer (ViT) [Dosovitskiy et al. 2021] implementation, images are split into fixed-size patches, linearly embed each of them, add position embeddings, and the feed the resulting sequence of vectors to standard Transformer encoder.

Chapter 3

Related Work

As introduced in the background 1 section, related work in the no-reference or blind image quality assessment could be divided in four [Ma et al. 2022] approaches with the evaluation of the research in the field.

Hand-crafted feature extraction

Traditional BIQA methods with handcrafted quality-aware feature extraction and learning the image quality in a knowledge-driven process were the beginning of modern objective image quality assessment-related research [Bovik and Wang 2006] [Ma et al. 2022]. Natural scene statistics are extracted from mean subtracted contrast normalized operation [Mittal et al. 2012b] [Ma et al. 2022] and DCT coefficient [Xu et al. 2016] [Ma et al. 2022], and perceptual quality is predicted with support vector regression (SVR).

Image compression algorithm categories such as block-based and wavelet create distinct distortions in digital images. Compression algorithms like JPEG, MPEG-1, MPEG-2, and H.26x are block-based. These algorithms usually use block partitioning of images before applying further processing steps. JPEG compression specifically partitions the target image in an 8x8 block and then applies a local discrete cosine transform (DCT) to each pixel of those blocks with further processing steps. The common distortions due to these processes are interblock blurring within blocks and blocking artifacts around block boundaries [Fig **??**]. Both of these distortions can be explained either in the spacial domain [Wang et al. 2002] [Bovik and Wang 2006] or in the frequency domain [Wang et al. 2000] [Bovik and Wang 2006].

To extract these distortion features with an objective model, blocking artifacts are measured in vertically and horizontally. The power spectrum of any of the directions can be measured with a Fourier transform method such as N-point discrete Fourier transform (DFT). In a real-world implementation, this can be computed with a fast Fourier Transform and the overall power spectrum can be estimated [Bovik and Wang 2006][Fig 3.2].

Handcrafted features are defined by the observation of natural images with the absence, and the presence of distortions. These shallow methods can't represent the complex human visual system. Even though these traditional methods accurately predict quality measurements for datasets with singular distortion and artificial distortions such as TID2013, LIVE [Ponomarenko et al. 2015b] [Sheikh et al. 2006a] [Ma et al. 2022] but they are not

Figure 3.1: Original image (left) and after applying JPEG compression to the original image (right)

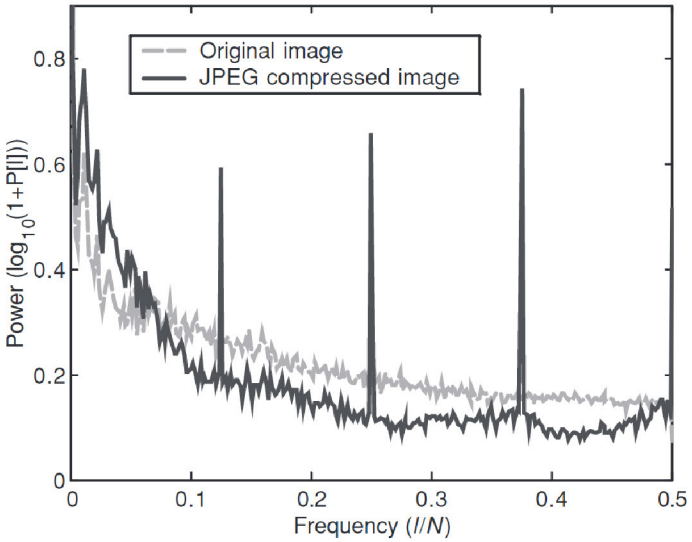


Figure 3.2: Power spectra of original and JPEG compressed image [Bovik and Wang 2006] that shows energy peaks on feature frequencies

efficient on the more wild datasets such as KonIQ, LIVEC, PaQ-2-PiQ [Hosu et al. 2020] [Ghadiyaram and and 2016] [Ying et al. 2020] [Ma et al. 2022].

Convolution neural network based

The powerful feature representation of convolutional neural networks became trendy due to the nature of high-level quality feature extraction. Adding plain CNN architecture to BIQA tools was the first feasibility test by researchers [Ma et al. 2022].

[Bosse et al. 2018] used CNN to extract features from distorted patches. Feature vector regression was applied to the patchwise quality estimate which is aggregated to a global image quality estimate. Optional regression of the feature vector is applied to patchwise weight estimates which allow for pooling by weighted average patch aggregation.

[Li et al. 2016] proposed a 31-layered deep CNN architecture to obtain a quality score for each patch and average them to obtain a global quality score. Multi-layer perceptron (MPL) layer, global average pooling (GAP) layer, and dropout layers are used in the proposed architecture. Rectified linear unit (ReLU) was used as an activation function instead of sigmoid or tanh neurons.

[Kang et al. 2014] proposed a framework that uses CNN to perform a contrast normalization and then sample the non-overlapping patches from it. It also estimates the quality score for each patch and averages the patch scores to estimate a quality score for the whole image. Similar to BRISQUE and CORNIA, it applies contrast normalization but in a simple and local manner. These locally normalized image patches are fed into a convolutional layer with 50 filters that generates a feature map and are applied pooling to reduce the filter responses to a lower dimension. NR-IQA/BIQA tasks were observed to be having image distortion which is most of the time locally homogeneous and the same level of distortion is distributed over the whole patch. ReLU was used in the fully connected layers, but linear neurons with identity transform are used in the convolutional and pooling layers to keep the negative outputs. The importance of the number of kernels, kernel size, patch size, and sampling stride is also part of the research.

Due to the unavailability of larger samples, these works tend to have over-fitting problems. To make the objective models learn better feature representation for NR-IQA/BIQA, later researchers started to change network architecture to incorporate external knowledge into the BIQA models [Ma et al. 2022].

[Lin and Wang 2018] proposed a CNN-based BIQA method with a hallucination-guided quality regression network. The model consists of three parts, the quality-aware generative network, the IQA-discriminative network, and the hallucination-guided quality regression network.

[Pan et al. 2018] used a BIQA model consisting of a fully convolutional neural network (FCCN) and a pooling network. The FCNN is responsible for generative quality mapping and the pooling network is responsible for quality scoring. U-Net, an extension of FCNN, was used as the base of the generative network. The hierarchical representation in subsampling layers with corresponding features in the upsampling layer which was brought from U-Net, can consider both high-level and low-level degradations for IQA. Batch normalization and leaky rectified linear unit (LReLU) are used after all convolutional layers. The pooling network is consisted of two fully connected layers with 50% dropout after each of them to prevent overfitting. It ends up with a squared Euclidean loss layer.

[Su et al. 2020] proposed a self-adaptive hyper network architecture for BIQA in the wild. The model has three major components (a) A foundational network that extracts logical features, (b) a target network that predicts image quality, and (c) a hyper network that produces a respective collection of self-adaptive parameters for the target network.

There were significant overall improvements with CNN-based methods over the traditional handcrafted methods. Global average pooling or global maximum pooling was the next logical step toward BIQA research with CNNs. But images tend to have local-variant

distortions and corresponding quality-aware responses throughout the arbitrary positions showing an unequal pattern, and the strategies above are incapable of dealing with such unequal distribution [Ma et al. 2022].

Learnable attention based

Dealing with the unequal pattern of quality-aware results along with the arbitrary positions on an image, learnable attention-based methods started to seem like the next best steps towards improving upon the existing BIQA research [Ma et al. 2022].

[Yang et al. 2019] proposed an SGDNet where visual saliency is learned and placed for measuring the weight of the quality-aware features in a more feasible way in a data-driven manner. SGDNet is built on an end-to-end multi-task learning framework with two sub-tasks (a) visual saliency prediction and (b) image quality prediction and both are connected with a shared feature extractor. The saliency prediction sub-task is more generic due to the nature of being independent of distortions. Saliency information is highly correlated with image quality according to the related works on the subject and it was fully utilized in this proposal with more informative labels with saliency maps and quality scores. The output of sub-task (a) is transparent to regression sub-task (b) by assisting with a spatial mask for a more perceptually-consistent feature combination.

[Gu et al. 2019] proposed a spatial attention module alongside the main branch, local quality, and local weight that is optimized collaboratively with the proposed attention-based pooling network(APNet). To generalize the purpose of this research, it focuses on the pooling stage and proposes an attention-based pooling network for BIQA. The goal of the learnable pooling is to represent human visual attention in a data-driven manner. Image quality prediction tasks can be conducted by fine-tuning a pre-trained classification network with global average pooling that appends the local quality estimation layer to obtain an overall quality score. The pooling can be improved with saliency prediction or generic object detection which is usually learned on natural distortion-free images. APNet attempts to learn an attention-based pooling strategy that assigns a positive weight to each location. The weight can be explained as the contribution or importance of location in combining the local quality estimations together. APNet consists of two branches (a) a local quality map, and (b) an attention weight map. A "soft" attention model [] is used to compute the attention weights.

[Dosovitskiy et al. 2021] [Fan et al. 2021] [Liu et al. 2021] papers related to vision transformer.

[You and Korhonen 2020] proposed a Transformer based BIQA method using ResNet50 [He et al. 2016] as a feature projector, and two transformer encoders to learn spatial dependencies in a global manner. The proposed method is called Transformer in Image Quality (TRIQ) assessment. Inspired by Vision Transformer(ViT), the proposed architecture uses a shallow Transformer encoder on top of a feature map extracted by a convolutional neural network. To handle the arbitrary resolution of images, an adaptive positional embedding is implemented in the Transformer encoder. Gaussian error linear unit (GELU) was used as the activation function as per the suggestions from ViT [] and BERT [].

[Ke et al. 2021] proposed a multi-scale image quality (MUSIQ) assessment method with

Transformer to process native resolution images with varying sizes and aspect ratios. The proposed method can capture image quality at different granularities due to muti-scale image representation. It consists of a novel hash-based 2D spatial embedding and a scale embedding in the multi-scale representation. MUSIQ is constructed to take input of images as multi-scale representation including native resolution image and aspect ratio preserved (ARP) resized variants. A patch encoding module embeds each image is split into fixed-sized patches. To handle images with varying aspect ratios and record the 2D construct of images, the spatial embedding is encoded by hashing the patch position in a grid of learnable embeddings. Scale embedding is used to record scale information. The Transformer encoder uses the input tokens and carries out a multi-head self-attention. And finally, to predict the image quality, MUSIQ follows a common approach in Transformers to add a token to the sequence to represent the whole multi-scale input and use the corresponding Transformer output as the final representation.

[Zhu et al. 2021a] proposed a saliency-guided Transformer network combined with local embedding (TranSLA) for BIQA. TranSLA combines divergent levels of information for a wider representation. HVS tends to focus more on the region of interest (ROI) when assessing image quality as the research says. Backing on that information, TranSLA integrates saliency prediction with Transformer to guide the model to highlight the ROI when grouping the global information. Local embeddings are imported with a gradient map for Transformer. Due to the gradient map concentrating on obtaining structured features in detail, it can be complemented to provide local information for Transformer, resulting in local and non-local information can be used. A boosting interaction module (BIM) is employed to speed up the aggregation of information of all tokens and enhance feature aggregation. Better interactions of patch tokens with the class tokens are forced by the BMI.

These Transformer based BIQA approaches have proven to effectively assign several types of attention to quality-aware responses in different positions and getting benefited from the quality prediction accuracy towards the heterogeneously distorted images. This results in such spatial attention learning approaches bound to limited improvement when it comes to homogeneous distorted images with pure CNN strategies [Ma et al. 2022].

Channel attention based

Recent research provided evidence that each channel of CNN-extracted features can contribute heterogeneously to the final result and learning channel-wise attention generally tend to boost the performance in DNN architecture in several computer vision tasks [Ma et al. 2022].

Squeeze-and-Excitation networks (SENet) architecture by [?] proposed a way for explicitly modeling interdependencies among channels to adaptively recalibrate channel-wise feature response. It is demonstrated that by stacking these SE blocks together, SENet architecture can be generalized pretty well even with challenging datasets. Dual graph CNN architecture in both spatial and channel domains by [Zhang et al. 2019] to boost the performance of semantic segmentation. It models the global content of the input feature by modeling two orthogonal graphs in a single setup. The first component models the spatial relationship between pixels in the image. The second component models interdependencies along the channel dimensions of the network's feature maps. And it's done by projecting the feature in

a new and lower-dimensional space where all the pairwise interactions can be modeled before reprojection into the original space. A Self-attention based scene segmentation approach by [Fu et al. 2019] records rich contextual dependencies with Dual Attention Network (DANet) to adaptively integrate local features with their global dependencies. Two types of attention modules are appended on top of dilated fully connected neural networks (FCN) which model the semantic interdependencies in spatial and channel dimensions respectively. The position attention module chooses and groups the feature at each position by a weighted sum of the features at all positions. Similar features would be related to each other regardless of their distances. In the meantime, the channel attention module selectively emphasizes interdependent channel maps by integrating associated features among all channel maps. Then the outputs of the two attention modules are summed to further improve feature representation which contributes to more precise segmentation results.

[Ma et al. 2022] explored the feasibility of incorporating an attention mechanism in a channel-wise manner for BIQA. With a systematic study of the interactions between channel-wise and spatial-wise attention, an adaptive spatial and channel attention merging transformer is developed for aggregating both spatial-wise and channel-wise attention information.

3.1 Image Quality Datasets

Due to image quality assessment as a machine learning task being critical and popular among the research communities, there have been several datasets introduced over time. Most of these image quality datasets come with pristine images, distorted images with different distortion types and levels, labeling of several kinds such as mean opinion scores (MOS) by human subjects, and different quality measurement parameters. Distortion types of the images in the datasets vary from 5 to 425 types and subtypes of distortions.

Table 3.1: State-of-the-art image quality datasets

Image Quality Assessment Datasets				
Dataset Name	Pristine Images	Distorted Images	Human Judgements	Distortion Types
LIVE 1 [Sheikh et al. 2003]	29	456	x	JPEG, JP2K
LIVE 2 [Sheikh et al. 2006b]	29	981	25,000	JPEG, JP2K, Gaussian Blur, White noise, Fast-fading
CLIVE [Ghadiyaram and Bovik 2015]	1,164		x	Natural
TID2013 [Ponomarenko et al. 2015a]	25	3,000	524,000	24 distortions [Ponomarenko et al. 2015a]

Continuation of Table 3.1				
Dataset Name	Pristine Images	Distorted Images	Human Judgements	Distortion Types
Waterloo [Ma et al. 2017]	4,744	94,880	x	JPEG, JP2K, White Gaussian noise, Gaussian blur at 5 levels of distortions
KonIQ 10K [Hosu et al. 2020]	10,073		x	Natural
KADID 10k [Lin et al. 2019]	81	10,125	x	25 distortions in 5 levels
KADIS 700k [at Universität Konstanz]	140,000	700,000	x	5 random distortions
CSIQ/CID-IQ [Liu et al.]	23	800	5,000	JPEG, JP2K, Poisson noise, Gaussian blur, SGCK gamut and DeltaE gamut mapping
PIPAL [Jinjin et al. 2020]	250 (288x288 patch)	29,000	1.13m	40 distortions [Jinjin et al. 2020]
PieAPP [Prashnani et al. 2018]	200 (256x256 patch)	20,280	2.3m	75 distortions [Prashnani et al. 2018]
BAPPS [Zhang et al.]	187,700 (64x64 patch)	375,400	484,300	425
End of Table Image Quality Assessment Datasets 3.1				

3.2 Image Quality Assessment Algorithms

With the growth of image quality assessment datasets, and machine learning techniques, objective image quality assessment has become popular among the communities of researchers. There have been several major breakthroughs in the field. From the analysis of related works **??** in this research, it's clear that blind image quality assessment has evolved over the last decades from hand-crafted feature extraction to convolutional neural networks to self-attention-based methods.

Table 3.2: State-of-the-art image quality algorithms

Image Quality Assessment Algorithms		
Algorithm Name	Type	Technique
BRISQUE [Mittal et al. 2012a]	Hand-crafted Feature Extraction	Support Vector Regression

Continuation of Table 3.2		
Algorithm Name	Type	Technique
SENet [Hu et al. 2018]	CNN	Squeeze-and-Excitation
KonCept512 [Hosu et al. 2020]	CNN	x
SGDNet [Yang et al. 2019]	CNN	x
PaQ-2-PiQ [Ying et al. 2020]	CNN	x
MANIQA [Yang et al. 2022]	Self-attention	ViT
HyperIQA [Su et al. 2020]	CNN	Self-adaptive
TRIQ [You and Korhonen 2020]	Transformer	ViT
MUSIQ [Ke et al. 2021]	Transformer	ViT
TranSLA [Zhu et al. 2021b]	Transformer	ViT
APNet [Gu et al. 2019]	Learnable Attention Transformer	ViT
ASCAMF [Ma et al. 2022]	Channel-attention Transformer	ViT
End of Table Image Quality Assessment Algorithms 3.2		

3.2.1 BRISQUE (NR/B-IQA in the Spatial Domain)

The most recent implementation of BRISQUE [Mittal et al. 2012a] uses the TID2008 image quality dataset and a Support Vector Machine library LibSVM to train a model to learn distortion features. The support vector regressor predicts the image quality score from 0 to 100. The higher the value the poorest the quality. Quality score prediction is run with KonIQ 10K [Hosu et al. 2020] dataset.

BRISQUE [Mittal et al. 2012a] seem to deviate from expected result **??** when tested with KonIQ-10K [Hosu et al. 2020]. Images with flat color surfaces seem to be evaluating inaccurately.

3.2.2 HyperIQA (Self-adaptive Hyper Network)

[Su et al. 2020] proposed a self-adaptive hyper network architecture to assess image quality without having a pristine reference image in the wild. The authors separated the quality assessment process into three stages, (1) content understanding, (2) perception rule learning, and (3) quality predicting. When the extraction of image semantics is done, rules of perception are established adaptively with a hyper network, and then adopted with a quality prediction network. In this model, image quality is supposed to be estimated in a self-adaptive manner and it generalizes in a robust collection of images captured in the wild.

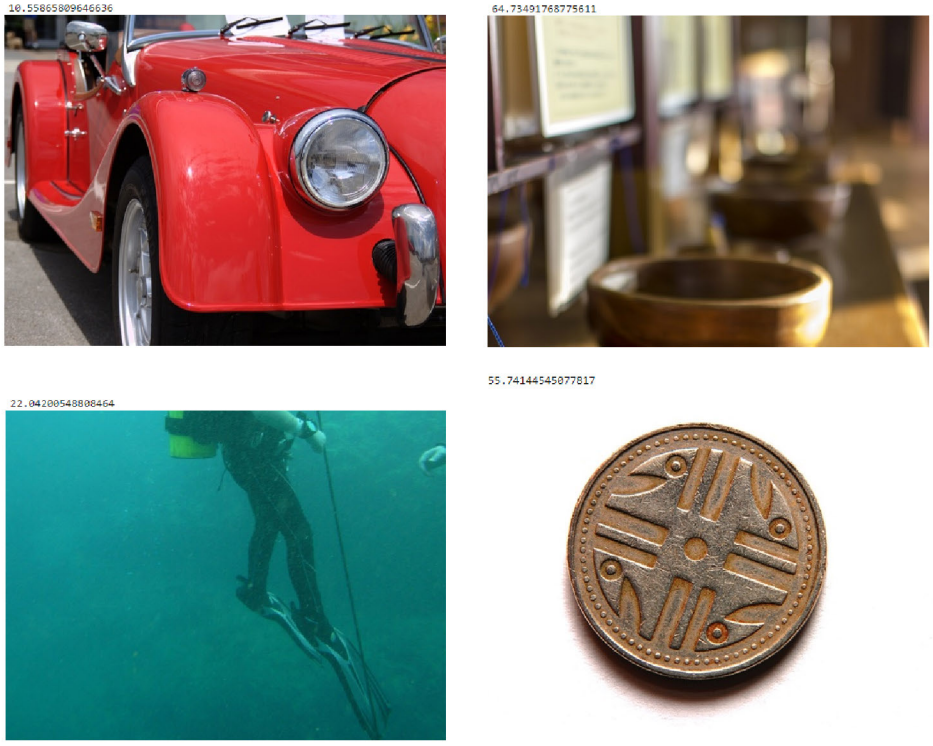


Figure 3.3: BRISQUE results in different scenarios

The proposed method aims to develop a quality assessment network that adaptively predicts image quality according to the image content without having knowledge of the pristine source image. The proposed network consists of three parts, (1) a backbone network that extracts semantic features, (2) a target network that predicts image quality, and (3) a hyper network that generates a series of self-adaptive parameters for the target network.
Image quality prediction approaches with convolutional neural networks usually receive an input image and directly map it to a quality score. These types of prediction models imply that the same kind of quality features are extracted for predicting diverse images. In practice, as image contents vary, using the same rule for predicting varied images, quality is not thorough to cover their different exhibited structures. As the image content varies, the way of perceived image quality varies accordingly. In this study [Su et al. 2020], the image quality assessment model becomes self-adaptive as it extracts different quality features in accordance with different image contents.

In the proposed hyper network, the authors defined the input of the hyper network as a function of semantic feature extraction from the input image. The function of the hyper network is to learn the mapping from the input image content and establish the rules on how to judge image quality. By introducing an intermediate variable, basic image quality steps are divided into three previously mentioned steps. The backbone network is used to extract semantic features from images, the hyper network is used for learning the image quality perception rule, and the quality prediction target network is used to obtain a final quality score. The designation makes the network more flexible toward extracting quality influential features when facing content-varying images with varying distortions.

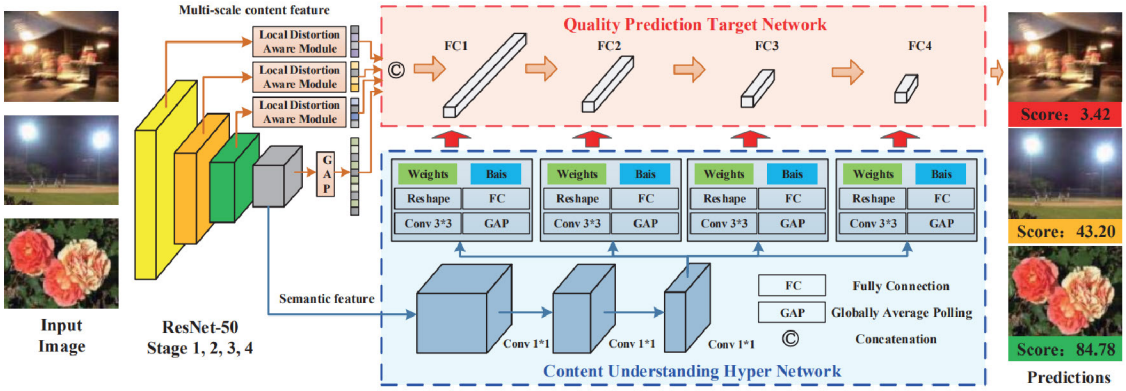


Figure 3.4: HyperIQA network architecture [Su et al. 2020]

The semantic feature extraction network in the study [Su et al. 2020] mainly focuses on understanding image content and outputs two different streams of features for quality measurement. The generated features are directly fed into the hyper network for generating weights and the multi-scale content feature stream is used as the input of the target network. The multi-scale content features are used due to the semantic features that are extracted from the last layer barely representing the holistic image content. The extraction of multi-scale features is done through a local distortion-aware module to capture the local distortions in the captured images in the wild. The designated local distortion-aware module is built with a series of operations including diving multi-scale feature maps into non-overlapping patches, stacking the patches along the channel dimension, running a 1x1 convolution, and globally pooling them into vectors. ResNet50 [?] which is pre-trained on ImageNet[] was used as the backbone network to initialize the network. Extracted multi-scale features conv2_10, conv3_12, and conv4_18 layers are the input of the local distortion-aware module.

The hyper network that is designated for learning perception rules that are built with three 1x1 convolutional layers and a few weight-generating branches. Fully connected layers are used as the basic target network components, two kinds of network parameters such as fully-connected layer weight, and fully-connected layer biases should be generated. Weights for the fully-connected layers are generated from the convolution followed by the reshaping operation of the extracted features, while fully connected later biases are generated by simple average pooling and full connection, as biases have significantly lower amounts of parameters. The weights that are generated in these layers are the representation of the learned rules of quality prediction and will be used as instructions for the target network.

The multi-scale features extracted with semantic extraction network being content-aware, the target network for the quality prediction can do the mapping of the learned features to a quality score. A simple network is used for the quality prediction. The target network is built with fore fully connected layers that receive the multi-scale content feature vectors as input and propagate through the specific weights determining layers to get final quality scores. A sigmoid function is used as an activation function.

3.2.3 KonCept (Pooling-based Convolutional Neural network)

At the time of crowd-sourcing and creating the image quality dataset KonIQ-10K [Hosu et al. 2020], the authors proposed a novel deep learning method, KonCept512 [Hosu et al. 2020] with the appropriate approaches in mind. The proposed method was built on top of a few considerations, (1) the input images size for the network, such as down-sized versions of the original image or crops, (2) the variety of choices for the base architecture, (3) minimized loss function, and (4) the aggregation strategy in case of multiple predictions. The architecture proposed by the authors is an end-to-end approach. An input image will

Figure 3.5: KonCept512 network architecture [Hosu et al. 2020]

pass through a few convolutional layers without the final fully-connected layers, followed by a global average pooling layer (GAP). The layers are connected to four fully-connected layers with 4,048, 1,024, and 256 units, and an output layer has either one output unit to predict MOS or five units to predict distributed MOS ratings. The three fully-connected layers use ReLU as an activation function and are connected to a dropout layer with rates of 0.25, 0.25, and 0.5 to avoid over-fitting. Soft-max activation is used in the final prediction layer.

The proposed method seeks to minimize the associated loss function which refers to the cost of a wrong prediction by the algorithms. The authors evaluated five loss functions. For MOS prediction mean absolute error (MAE) loss, and mean squared error (MSE) loss were used. For distributed rating prediction, cross-entropy loss, Huber loss, and earth mover's distance (EMD) loss were used.

While predicting MOS, the input image is fed into the proposed system. Then both the MAE and MSE are calculated. The MSE is differentiable at the origin that can generate smoother gradients in case of small errors than the MAE. And it also punishes the larger deviations from the ground truth efficiently.

To predict the distribution of ratings five-point ACR for subjective quality assessment of the dataset was used. For image classification tasks, a cross-entropy loss is standard. The same loss definition is for this regression task. Huber loss for a scalar prediction error controls the degree of influence given to larger prediction errors. The Earth Mover's Distance (EMD) loss enhanced results compared to cross-entropy. The used loss is established as the root mean squared difference between the predicted and ground truth better distributions of scores.

3.2.4 MANIQA (Multi-dimension Attention Network)

The multi-dimension attention network proposed by [Yang et al. 2022], tries to solve the existing image distortion as well as the GAN-based image distortions.

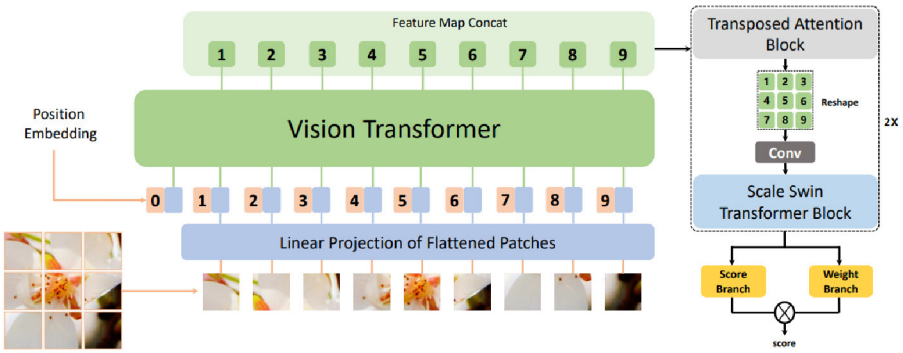


Figure 3.6: MANIQA network architecture [Yang et al. 2022]

The feature extraction is done with ViT (Vision Transformer) [Dosovitskiy et al. 2021]. The transposed attention block (TAB) and the scale Swin transformer block (SSTB) were proposed to enhance the global and local interactions. These two blocks apply attention mechanisms across the channel and spatial dimensions, resulting in increasing interactions among different regions of the images locally and globally. A two-branch structure for patch-weighted quality prediction is used to determine the final score depending on the weight of ht each patch's score.

The goal of the authors for the proposed method was to establish a model that can deal with multi-dimensional information from extracted image features. To utilize the information from the channel and the spatial dimensions, the three core components were proposed inside the main architecture, (1) transposed attention block (TAB), (2) scale Swin transformer block (SSTB), and (3) a two-branch structure for patch-weighted quality prediction.

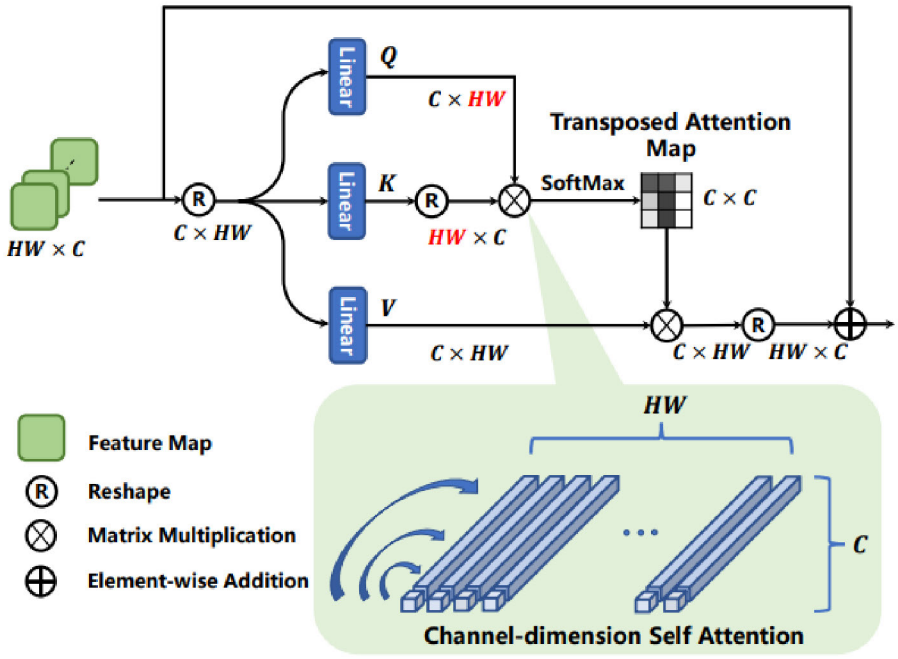


Figure 3.7: MANIQA transposed attention block [Yang et al. 2022]

Self-attention layer is an essential part of any Transformer[Vaswani et al. 2017] block. The key-query dot-product interaction establishes the global connection among the patches in spatial dimension in most of the self-attention implementations but they tend to ignore the valuable information among different channels. To counteract this issue, the proposed transposed attention block implements self-attention across channels instead of the spatial dimension to calculate cross-variance across channels to generate map encoding in the global context in a more implicit manner. From the joined features generated by the proposed transposed attention block query, key, and value projections are established with three independent linear projections to encode point-wise cross-channel context. The query and key projections are reshaped so that the consecutive dot-product interaction creates a transposed attention map. The layer normalization and multi-layer perception are removed from the original Transformer. The features from the initial four layers of ViT contain different information in channels about the input image. The transposed attention block restructures the channels' weights in regard to the importance of the perceptual quality score.
The scale Swin transformer block consists of Swin Transformer Layers (STL) [Liu et al. 2021]. The SSTB encodes the features through two layers of STL. A convolutional layer is applied before the residual connection. The convolutional layer with spatially invariant filters can enhance the translational equivariance. The scale factor makes the training stable with the residual connection.

For the final quality prediction block, a dual branch structure for patch-weighted was proposed by the authors. This module has a scoring and weighting branch. Those branches predict each patch's score and weight with 2 independent linear projections. Due to images

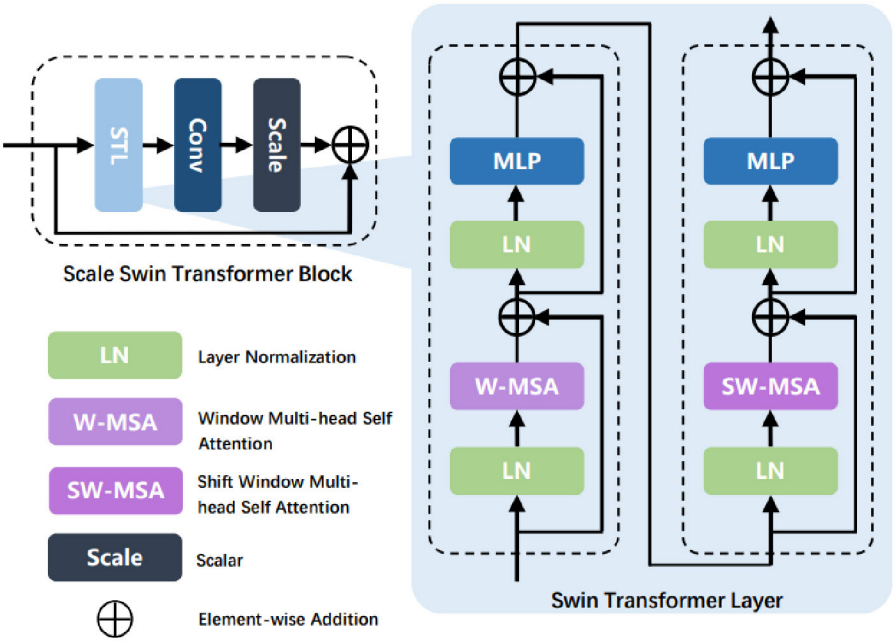


Figure 3.8: MANIQA scale swin transformer block [Yang et al. 2022]

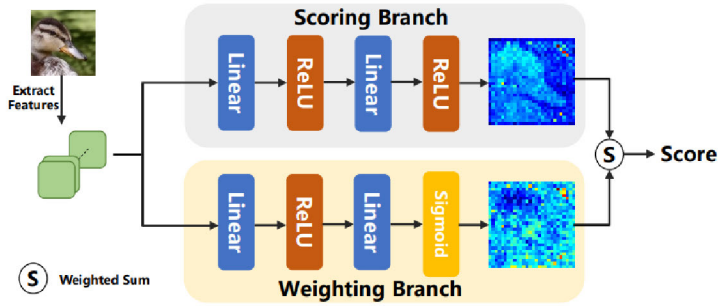


Figure 3.9: MANIQA dual branch patch-weighted quality prediction [Yang et al. 2022]

containing feature information in different regions, the final patch score of distorted images is calculated by multiplicating each patch's score and weight. And then the final score is calculated by summing up the final patch scores.

Chapter 4

Design

4.1 Design Planning Outline

The experiments will need several image-quality datasets and algorithms from published related literature. From the curated list in the related works section, the research will be scoped down to a few of the choices with appropriate reasoning. After establishing a curated list of datasets and algorithms, the curated algorithms will be trained, validated, and tested on the curated datasets. Standard performance matrices will be calculated in the experiment. Finally, these will be visually represented in various charts, graphs, and tables.

4.2 Curated Image Quality Datasets

While choosing from a pool of image quality assessment datasets, a few considerations might help reduce the number of choices. The number of samples for training, validating, and testing is the general criteria for any image analysis study, so it goes the same for image quality assessment.

Then domain-specific task like image quality assessment comes with their own requirements. As the image quality is assessed against the distortion present in the image, classifying image distortion types and detecting them seems the most intuitive to go. But in the wild, distorted images contain a combination of distortions with an arbitrary ratio of distortions across the image. The image datasets that contain images with natural distortions are most compatible for no-reference image quality assessment. The number of quality ratings and the number of human subjects to give those quality ratings are also important to reduce bias as much as possible. Keeping these factors in mind, this research is scoped down to two datasets, LIVE in the wild [Ghadiyaram and Bovik 2015], and KonIQ 10K [Hosu et al. 2020].

Table 4.1: Curated image quality datasets

Curated Image Quality Assessment Datasets					
Dataset Name	Pristine Images	Distorted Images	Quality Ratings	Human Subjects	Distortion Types
CLIVE [Ghadiyaram and Bovik 2015]	1,164		350,000	8,100	Natural
KonIQ 10K [Hosu et al. 2020]	10,073		1,200,000	1,459	Natural
End of Table Curated Image Quality Assessment Datasets 4.1					

4.3 Curated Image Quality Assessment Algorithms

While choosing from a collection of image quality assessment algorithms, a few considerations might be in place to reduce the choices. The available algorithms with well enough descriptive publications and source code availability are the primary curating criteria for this comparative study. The secondary specification is to explore the choices of approaches. The approaches include classical hand-crafted feature extraction, the most popular convolutional neural networks, and the most recent self-attention-based transformers.

Keeping these factors in mind, this research narrows down to four algorithms from the pool of choices. For the hand-crafted feature extraction approaches, BRISQUE [Mittal et al. 2012a] was chosen. For the convolutional neural network approach, HyperIQA [Su et al. 2020], and KonCept [Hosu et al. 2020] were chosen. For the self-attention-based transformers approach, MANIQA [Yang et al. 2022] was chosen.

Table 4.2: Curated image quality algorithms

Curated Image Quality Assessment Algorithms		
Algorithm Name	Type	Technique
HyperIQA [Su et al. 2020]	CNN	Self-adaptive
KonCept [Hosu et al. 2020]	CNN	x
MANIQA [Yang et al. 2022]	Self-attention	ViT
End of Table Curated Image Quality Assessment Algorithms 4.2		

4.4 Source Code

All the source code for this experimental setup is hosted on GitHub and is subject to be updated in the future. The source code is hosted at `https://github.com/sourav-paul/hc2sa`. The default branch is "main".

4.5 Train, Validation, Test Tools

All necessary datasets are locally hosted in the development environment setup. All the algorithms are customized to use the datasets hosted in the development environment. The used tools, software, hardware, etc are listed below.

1. Language, Tools, and Software
   (a) Python
   (b) Amazon Web Service
   (c) Jupyter Notebook
   (d) SageMaker
   (e) Google Slides
   (f) Google Sheets
2. Hardware
   (a) CPU, up to 64 cores
   (b) RAM, up to 112 Gigabytes
   (c) GPU, up to 4 x Nvidia Tesla T4 16 Gigabytes

4.6 Performance Matrices

In the field of quality of experience measurement, the mean opinion score is a standard way of predefining a scale of measurement. In image quality assessment, the mean opinion score plays the same role. Pearson's correlation coefficient and Spearman's rank correlation coefficient are the final tools to determine correlations between the ground truth MOS and predicted MOS.

4.6.1 Mean Opinion Score (MOS)

Mean opinion score (MOS) is a measurement tool used in the domain of quality of experience representing the overall quality of a stimulus or system. A predefined scale is used and assigned by a subject while their opinion of the performance of a system's quality is measured. Then the arithmetical mean is calculated over all the values assigned by the subjects. Usually, it's a subjective quality evaluation test but can also be estimated with an algorithm. MOS is usually used to measure video, audio, and audiovisual quality evaluation.

Table 4.3: MOS in Image Quality Assessment

MOS in Image Quality Assessment		
Image Name	Ground Truth MOS, x	Predicted MOS, y
100.bmp	32.5611	29.468
101.bmp	66.3595	69.112
102.bmp	44.695	42.53
103.bmp	39.2346	45.866

Continuation of Table 4.3		
Image Name	Ground Truth MOS, x	Predicted MOS, y
104.bmp	9.2291	15.762
105.bmp	41.3314	51.017
106.bmp	68.9221	67.1764
End of Table MOS in Image Quality Assessment 4.3		

4.6.2 Correlation Coefficient

Correlation coefficients are the numerical representation of some kind of statistical relationship between two variables and their strength. The variables might also be columns of two datasets of observation or samples, or two multivariate random variables with known distribution.

There are several types of correlation coefficients that exist in the field of statistical analysis with their own range of usability and characteristics. Pearson correlation coefficient and Spearman's rank correlation coefficient are two vastly used among them.

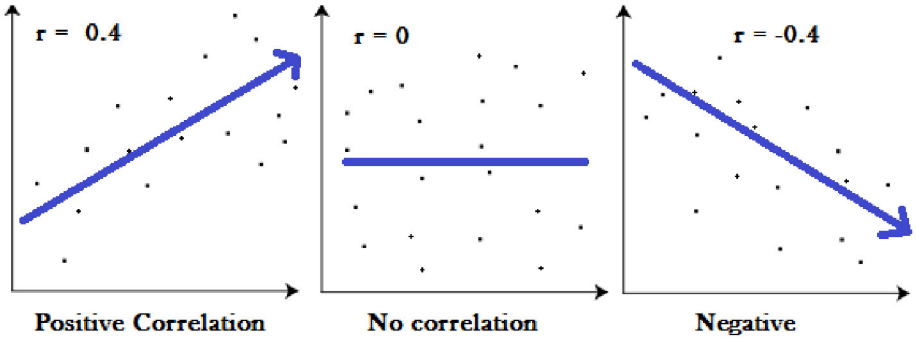


Figure 4.1: Different Types of Correlation

4.6.3 Pearson (Linear) Correlation Coefficient (PLCC)

Pearson correlation coefficient or Pearson's R is one of the most used correlation coefficients. Pearson's R is a measure of linear correlation between two sets of data. It represents the ratio between the covariance of the two variables and the product of their standard deviations. It ends up essentially being a normalized measurement of the covariance, and the result is always between -1 to 1. Due to covariance itself, the measurements only reflect a linear correlation of variables and ignore other types of relationships or correlations.

$$r = \frac{n(\sum(xy)) - (\sum(x))(\sum(y))}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{4.1}$$

4.6.4 Spearman's Rank (Order) Correlation Coefficient (SROCC)

Spearman's rank correlation coefficient is a non-parametric measure of rank correlation that implies the statistical dependence between the ranking of two variables. It justifies how well the relationship between two variables can be established with a monotonic

function. While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships but linear or not. The Spearman correlation between two variables will be high when observations have a similar rank between two variables, and low when observations have a dissimilar rank between the two variables. It's suitable for both continuous and discrete ordinal variables.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.2}$$

4.6.5 MOS, PLCC, SROCC in Image Quality Assessment

As the image quality assessment is a supervised regression task, the datasets come with a mean opinion score (MOS) assigned to each image in the dataset which is calculated from a collection of human subjects' individual scores. After the machine learning model is trained with the image and MOS pairs, it will predict a MOS as a result. The collection of these ground truth MOS and predicted MOS can be plotted and the correlation can be calculated with PLCC, and SROCC formulas.

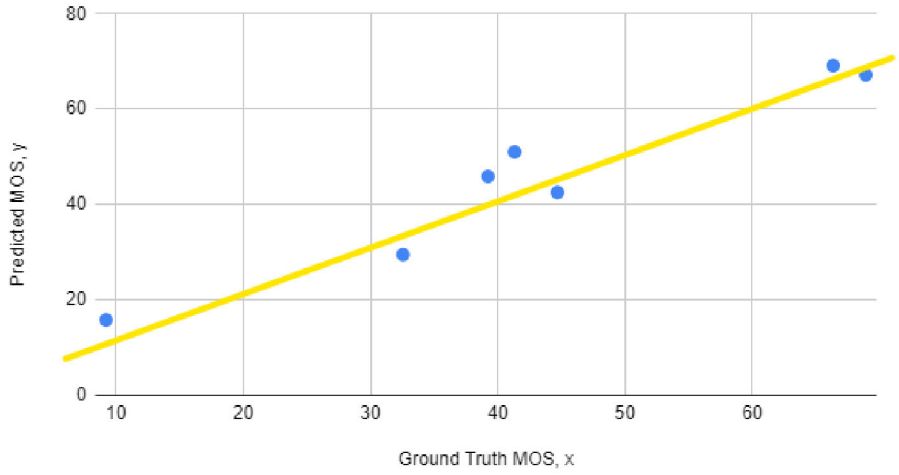


Figure 4.2: Positive Correlation in Image Quality Assessement

4.7 Implementation

The cross-analysis among state-of-the-art algorithms and datasets require both code and data ready to train, test, and validate.

HyperIQA Implemetation

The proposed method of image quality assessment by [Su et al. 2020] comes with open-source implementation in Python. PyTorch was the choice of machine learning framework by the authors along with other relevant libraries. While the original source code comes with pure Python implementation, the training interface code was converted to Jupyter Notebook format was brought in to keep the development environment reusable and easy to configure.

The original implementation comes prepared with the KonIQ-10K [Hosu et al. 2020] dataset. But additional redirection toward the local copy of the dataset was needed for this experiment. For the LIVE-in-the-wild [Ghadiyaram and Bovik 2015] dataset, some preparation was needed to make the acceptable to the implementation. A copy of the image and MOS pair for CLIVE [Ghadiyaram and Bovik 2015] needed to be placed And additional redirection toward the local copy of the dataset was needed to be placed in the original implementation code.

Training, validation, and test process are done two times using both of the datasets [Hosu et al. 2020; Ghadiyaram and Bovik 2015] separately. The training job in the original experiment was done with a high number of epochs (160), but due to time and resource constraints, this experiment uses 5% to 10% of the original number of epochs. And it applies to both of the training jobs of the separate datasets. In addition, the original training job runs multiple rounds of epochs.

Finally, PLCC, and SROCC results from every better-performing epoch are logged. The output represents the per-epoch correlation between the ground truth MOS and predicted MOS. These results will be summarized in the result section of this paper and will be discussed in regard to the research objective in the discussion section.

KonCept512 Implemetation

The proposed method of image quality assessment by [Hosu et al. 2020] comes with open-source implementation in Python. Keras-Utilities (Kutils) was the choice of machine learning framework by the authors along with other relevant libraries. While the original source code comes with Python implementation with Kutils library, some of the libraries were incompatible with this experiment's development environments. Instead, an open-source PyTorch implementation of KonCept512 with the training interface code was converted to Jupyter Notebook format were brought in to keep the development environment reusable and easy to configure. As the original implementation uses ResNet, InceptionResNetV2[] as backbone networks, the appropriate models were imported and used in the code.

The original implementation comes with the preparation of the KonIQ-10K [Hosu et al. 2020] dataset. But additional redirection toward the local copy of the dataset was needed for this experiment. On the other hand, for the LIVE-in-the-wild [Ghadiyaram and Bovik 2015] dataset, some preparation was needed to make the acceptable to the implementation. A copy of the image and MOS pair for CLIVE [Ghadiyaram and Bovik 2015] needed to be placed, and some extra code to trim them and add a new row defining as the type of data entry, such as training, validation, and test set separation to make the original easy to use. An 82% training, 9% validation, and 9% test data split ratio were maintained. And additional redirection toward the local copy of the dataset was needed to be placed in the original implementation code.

Training, validation, and test process are done two times using both of the datasets [Hosu et al. 2020; Ghadiyaram and Bovik 2015] separately. The training job in the original experiment was done with a high number of epochs (80), but due to time and resource constraints, this experiment uses 5% to 10% of the original number of epochs. And it applies to both of the training jobs of the separate datasets. In addition, due to using two

varieties of optimizers being used in the same training job, the number of epochs became twice as many.

In the testing stage, PLCC and SROCC results from every better-performing epoch are logged. The output represents the per-epoch correlation between the ground truth MOS and predicted MOS. These results will be summarized in the result section of this paper.

MANIQA Implementation

The proposed method of image quality assessment by [Yang et al. 2022] comes with open-source implementation in Python. PyTorch was the choice of machine learning framework by the authors along with other relevant libraries. While the original source code comes with pure Python implementation, the training interface code was converted to Jupyter Notebook format to keep the development environment reusable and easy to configure.

The original implementation comes with the preparation of the KonIQ-10K [Hosu et al. 2020] dataset. But additional redirection toward the local copy of the dataset was needed for this experiment. On the other hand, for the LIVE-in-the-wild [Ghadiyaram and Bovik 2015] dataset, some preparation was needed to make the acceptable to the original implementation. A copy of the image and MOS pair for CLIVE [Ghadiyaram and Bovik 2015] need to be placed, and some extra code to trim them and convert them to a regular text file from CSV to make the original easy to use.

Training, validation, and test process are done two times using both of the datasets [Hosu et al. 2020; Ghadiyaram and Bovik 2015] separately. The training job in the original experiment was done with a high number of epochs (300), but due to time and resource constraints, this experiment uses 5% to 10% of the original number of epochs. And it applies to both of the training jobs of the separate datasets.

In the testing stage, PLCC and SROCC results from every better-performing epoch are logged in an output file. The output file represents the per-epoch correlation between the ground truth MOS and predicted MOS. These results will be summarized in the result section of this paper.

Chapter 5

Results

There are some generalized steps in the experimental setup section for each algorithm and each dataset. All the curated algorithms are modified in a way to be able to train, validate, and test on both datasets.

5.1 HyperIQA Results

In the design section, the Python source code for HyperIQA [Su et al. 2020] is made ready to train, validate, and test on the curated datasets.

5.1.1 Trained, Validated, Tested On KonIQ-10K

KonIQ-10K [Hosu et al. 2020] dataset is a collection of images with natural distensions with a list of mean opinion scores. After the algorithm is trained, validated, and tested on the KonIQ-10K dataset, a few key performance indicators are collected.

Table 5.1: HyperIQA trained on KonIQ-10K result PLCC and SROCC

HyperIQA trained on KonIQ-10K			
Round/Epoch	Train SROCC	Test SROCC	Test PLCC
Round 1 - Epoch 1	0.5454	0.8804	0.9098
Round 1 - Epoch 2	0.9232	0.8885	0.9099
Round 1 - Epoch 3	0.9404	0.8885	0.9151
Round 2 - Epoch 1	0.5226	0.8750	0.8859
Round 2 - Epoch 2	0.9229	0.8827	0.8959
Round 2 - Epoch 3	0.9394	0.8926	0.9045
End of Table 5.2			

5.1.2 Trained, Validated, Tested On CLIVE

CLIVE [Ghadiyaram and Bovik 2015] dataset is a collection of images with natural distensions with a list of mean opinion scores. After the HyperIQA [Su et al. 2020]

algorithm is trained, validated, and tested on the CLIVE dataset, a few performances indicating results are collected.

Table 5.2: HyperIQA trained on CLIVE result PLCC and SROCC

HyperIQA trained on CLIVE			
Round/Epoch	Train SROCC	Test SROCC	Test PLCC
Round 1 - Epoch 1	0.7613	0.8134	0.8439
Round 1 - Epoch 2	0.9421	0.8245	0.8603
Round 1 - Epoch 3	0.9610	0.8208	0.8607
Round 2 - Epoch 1	0.7775	0.8334	0.8577
Round 2 - Epoch 2	0.9415	0.8281	0.8480
Round 2 - Epoch 3	0.9586	0.8223	0.8466
Round 3 - Epoch 1	0.7518	0.8386	0.8589
Round 3 - Epoch 2	0.9397	0.8474	0.8648
Round 3 - Epoch 3	0.9560	0.8491	0.8646
End of Table 5.2			

5.2 KonCept512 Results

In the design section of this article, the Python source code for KonCept512 [Hosu et al. 2020] is made ready to train, validate, and test on the curated datasets.

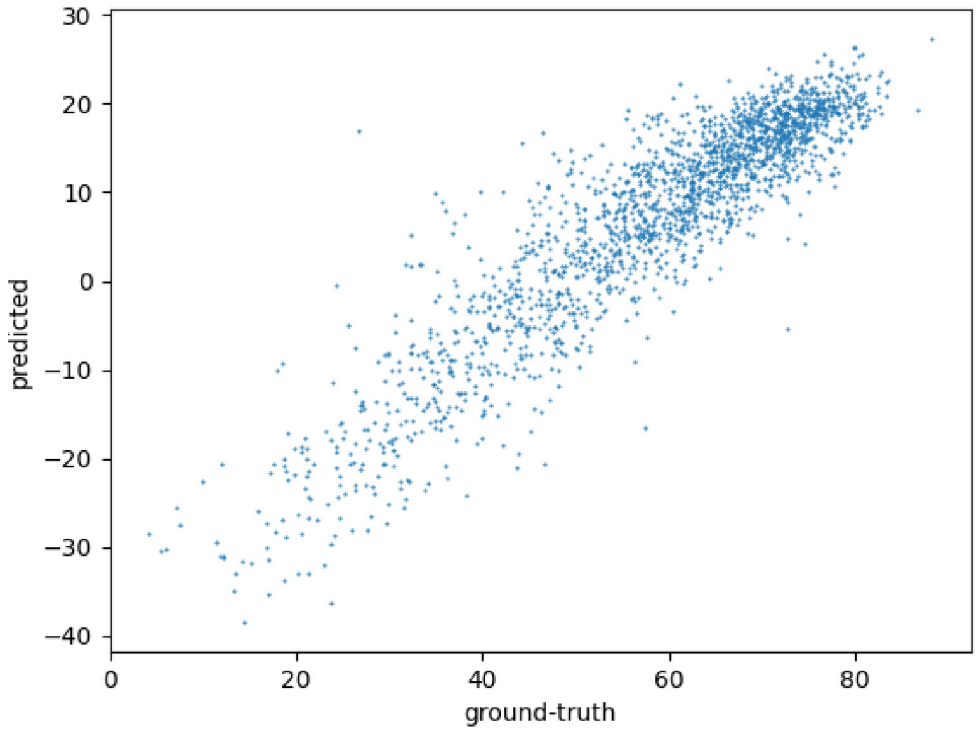
5.2.1 Trained, Validated, Tested On KonIQ-10K

The KonCept512 is ready in the experimental setup to utilize KonIQ-10K [Hosu et al. 2020] dataset is a collection of images with natural distensions with a list of mean opinion scores. After the algorithm is trained, validated, and tested on the KonIQ-10K dataset, a few key performance indicators are collected.

Table 5.3: KonCept512 trained on KonIQ-10K result PLCC

KonCept512 trained on KonIQ-10K	
Round/Epoch	PLCC
Epoch 1	0.8569
Epoch 2	0.8661
Epoch 3	0.8494
Epoch 4	0.8719
Epoch 5	0.8314
Epoch 6	0.8836
Epoch 7	0.8988
Epoch 8	0.9076
Epoch 9	0.8884
Epoch 10	0.9046
End of Table 5.3	

Figure 5.1: result of KonCept512 on koniq 10k



5.2.2 Trained, Validated, Tested On CLIVE

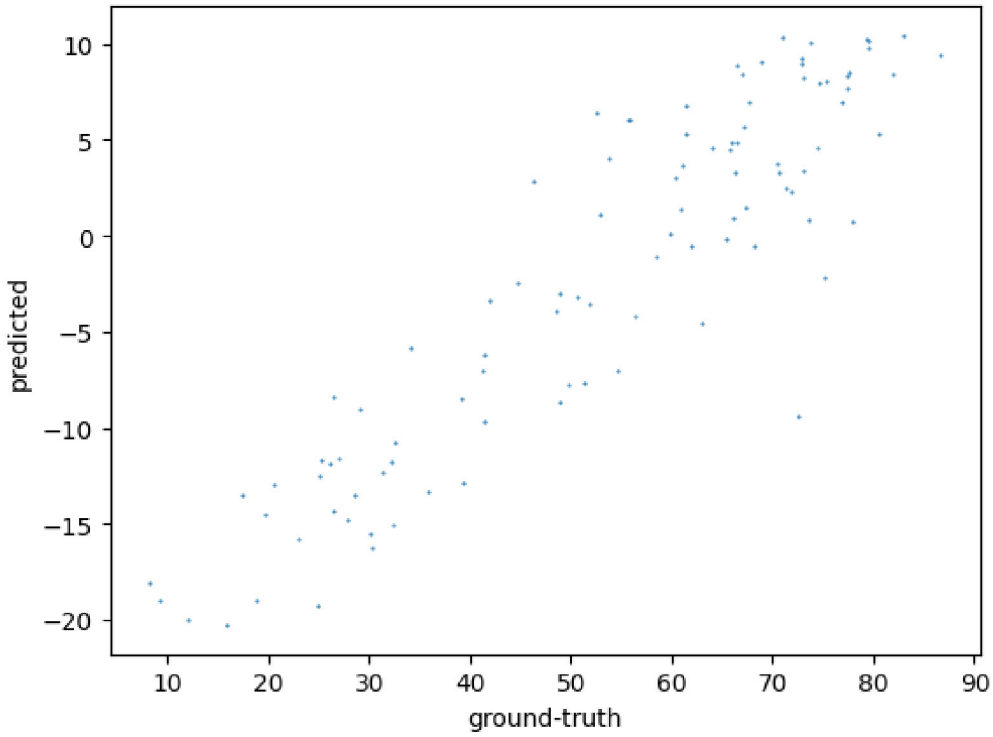
The KonCept512 is ready in the experimental setup to utilize CLIVE [Ghadiyaram and Bovik 2015] dataset is a collection of images with natural distensions with a list of mean opinion scores. After the algorithm is trained, validated, and tested on the CLIVE dataset, a few performance indicators are collected.

Table 5.4: KonCept512 trained on CLIVE result PLCC

KonCept512 trained on CLIVE	
Round/Epoch	PLCC
Epoch 1	0.5267
Epoch 2	0.7545
Epoch 3	0.8165
Epoch 4	0.8109
Epoch 5	0.8282
Epoch 6	0.8377
Epoch 7	0.8171
Epoch 8	0.8315
Epoch 9	0.8365
Epoch 10	0.8419
End of Table 5.4	

SRCC: 0.873 | PLCC: 0.91 | MAE: 55.519 | RMSE: 56.982

Figure 5.2: result of KonCept512 on clive



5.3 MANIQA Results

In the experimental setup section, the Python source code for MANIQA [Yang et al. 2022] is made ready to train, validate, and test on the curated datasets.

5.3.1 Trained, Validated, Tested On KonIQ-10K

The MANIQA is ready in the design to utilize KonIQ-10K [Hosu et al. 2020] dataset is a collection of images with natural distensions with a list of mean opinion scores. After the algorithm is trained, validated, and tested on the KonIQ-10K dataset, a few key performance indicators are collected as results.

Table 5.5: MANIQA trained on KONIQ-10K result PLCC and SROCC

MANIQA trained on KonIQ-10K				
Round/Epoch	Train PLCC	Test PLCC	Train SROCC	Test SROCC
Epoch 1	0.8224	0.8925	0.8436	0.9186
Epoch 2	0.9163	0.907	0.9369	0.9326
Epoch 3	0.9361	0.9148	0.9542	0.9368

Continuation of Table 5.5				
Round/Epoch	Train PLCC	Test PLCC	Train SROCC	Test SROCC
Epoch 4	0.9488	0.9187	0.9646	0.9381
Epoch 5	0.9572	0.9173	0.9699	0.9381
Epoch 6	0.964	0.9226	0.9745	0.9409
Epoch 7	0.9707	0.9176	0.9789	0.9381
Epoch 8	0.9749	0.9205	0.9818	0.9405
Epoch 9	0.9756	0.9197	0.9825	0.9412
Epoch 10	0.9775	0.9185	0.9834	0.9401
End of Table 5.5				

5.3.2 Trained, Validated, Tested On CLIVE

The MANIQA is ready in the design to utilize CLIVE [Ghadiyaram and Bovik 2015] dataset is a collection of images with natural distensions with a list of mean opinion scores. After the algorithm is trained, validated, and tested on the CLIVE dataset, a few key performance indicators are collected as results.

Table 5.6: MANIQA trained on CLIVE result PLCC and SROCC

MANIQA trained on CLIVE				
Round/Epoch	Train SROCC	Test SROCC	Train PLCC	Test PLCC
Epoch 1	0.3683	0.8249	0.324	0.847
Epoch 2	0.8108	0.8517	0.8581	0.8701
Epoch 3	0.8781	0.8771	0.9094	0.8881
Epoch 4	0.9236	0.8868	0.9403	0.8993
Epoch 5	0.9472	0.8913	0.9604	0.9045
Epoch 6	0.9437	0.8986	0.9589	0.9056
Epoch 7	0.9624	0.8977	0.9711	0.9099
Epoch 8	0.9711	0.894	0.9779	0.9023
Epoch 9	0.9786	0.9002	0.9833	0.9088
Epoch 10	0.9837	0.9029	0.9878	0.9116
End of Table 5.6				

Chapter 6

Discussion

In this chapter, the results collected from the experiment will be discussed in response to the research objective. The collected data are plotted and visualized as the experiment progresses.

6.1 HyperIQA Trained on KonIQ-10k

HyperIQA [Su et al. 2020] trained, validated, and tested on KonIQ-10K [Hosu et al. 2020] dataset. There were two rounds of training, validation, and testing were run. All three rounds had three epochs. From the graph, it's visible that with each round and epoch, the correlation coefficient gets close to 1 implying the predicted MOS has a positive correlation to ground truth MOS.

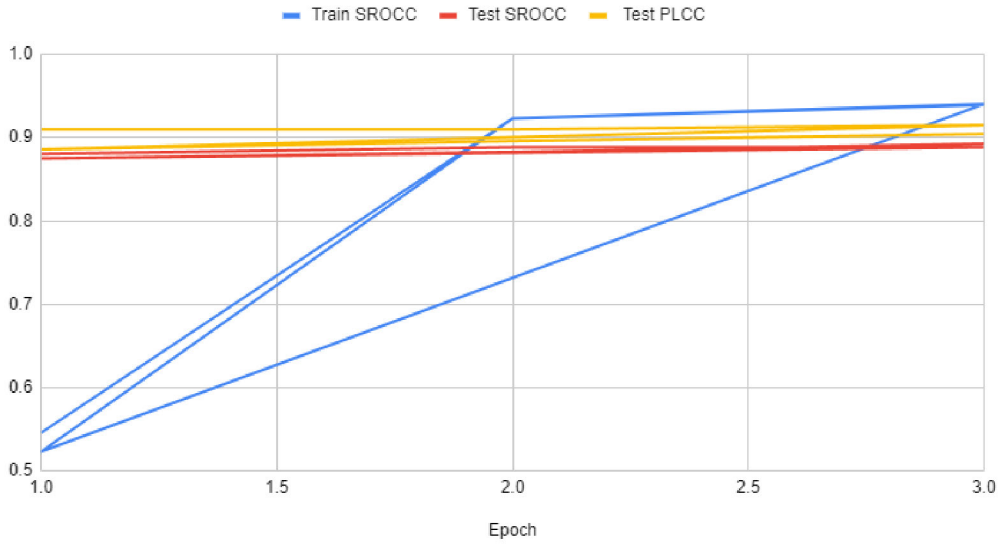


Figure 6.1: HyperIQA on KonIQ-10K

6.2 HyperIQA Trained on CLIVE

HyperIQA [Su et al. 2020] trained, validated, and tested on CLIVE [Ghadiyaram and Bovik 2015] dataset. There were three rounds of training, validation, and testing were run. All three rounds had three epochs. From the graph, it's visible that with each round and epoch, the correlation coefficient gets close to 1 implying the predicted MOS has a positive correlation to ground truth MOS.

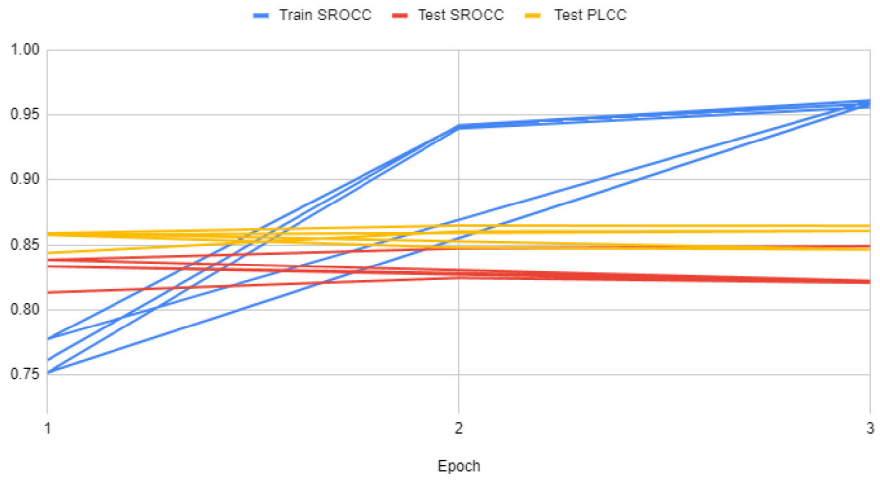


Figure 6.2: HyperIQA on CLIVE

6.3 KonCept512 Trained on KonIQ-10k

KonCept512 [Hosu et al. 2020] trained, validated, and tested on KonIQ-10K [Hosu et al. 2020] dataset. There was one round of training, validation, and testing run. They had ten epochs. From the graph, it's visible that with each epoch, the correlation coefficient gets close to 1 implying the predicted MOS has a positive correlation to ground truth MOS.

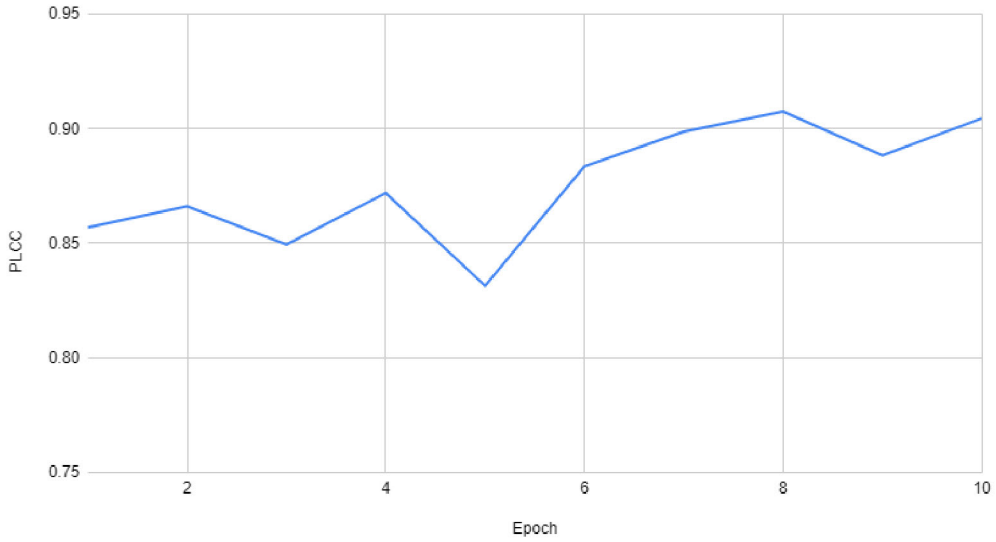


Figure 6.3: KonCept512 on KonIQ-10K

6.4 KonCept512 Trained on CLIVE

KonCept512 [Hosu et al. 2020] trained, validated, and tested on CLIVE [Ghadiyaram and Bovik 2015] dataset. There was one round of training, validation, and testing run. They had ten epochs. From the graph, it's visible that with each epoch, the correlation coefficient gets close to 1 implying the predicted MOS has a positive correlation to ground truth MOS.

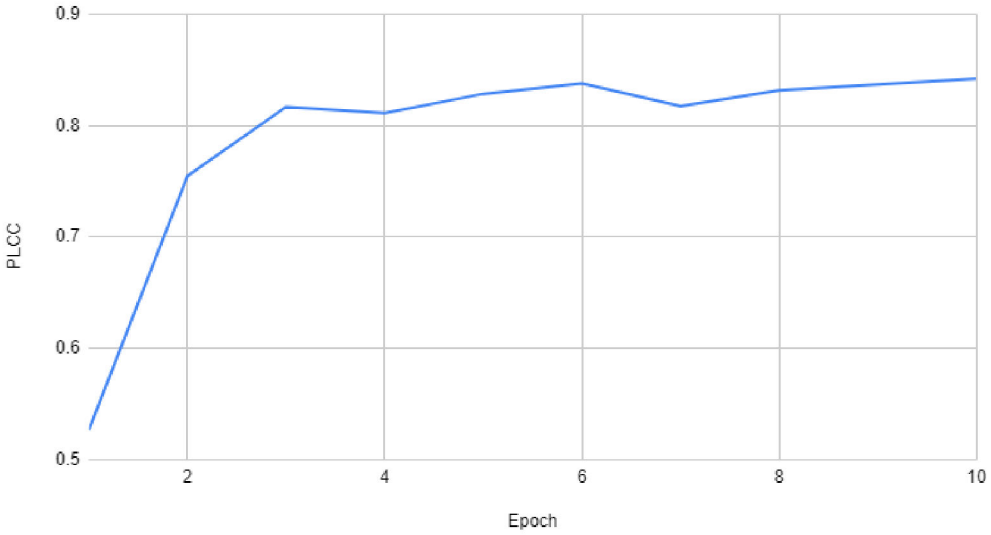


Figure 6.4: KonCept512 on CLIVE

6.5 MANIQA Trained on KonIQ-10K

MANIQA [Yang et al. 2022] trained, validated, and tested on KonIQ-10K [Hosu et al. 2020] dataset. There was one round of training, validation, and testing run. They had ten epochs. From the graph, it's visible that with each epoch, the correlation coefficient gets close to 1 implying the predicted MOS has a positive correlation to ground truth MOS.

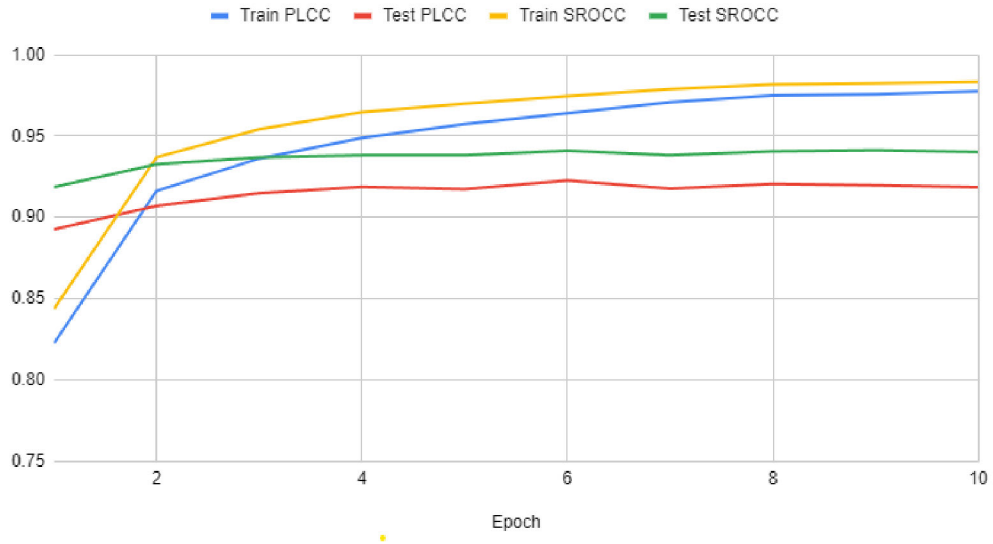


Figure 6.5: MANIQA on KonIQ-10K

6.6 MANIQA Trained on CLIVE

MANIQA [Yang et al. 2022] trained, validated, and tested on CLIVE [Ghadiyaram and Bovik 2015] dataset. There was one round of training, validation, and testing run. They had ten epochs. From the graph, it's visible that with each epoch, the correlation coefficient gets close to 1 implying the predicted MOS has a positive correlation to ground truth MOS.

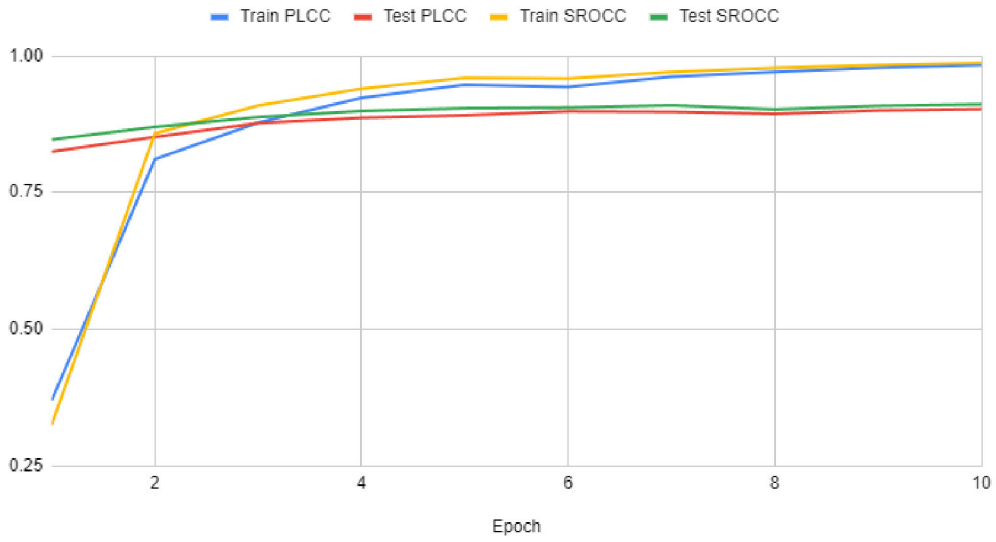


Figure 6.6: MANIQA on CLIVE

Chapter 7

Conclusion

A comparative study of objective perceptual no-reference or blind image quality assessment was conducted in this research. The purpose was to see how the algorithm performs against the datasets.

Three state-of-the-art machine learning algorithms were curated. Two image-quality datasets with natural distortion were curated. The algorithms were trained, validated, and tested across the datasets. The data were collected and plotted for visualization.

In each resulting case, it is visible that the correlation coefficients tend to reach near 1 implying that the predicted mean opinion score (MOS) has **a strong positive correlation** with the ground truth mean opinion score (MOS). The state-of-the-art algorithms perform well with different datasets with natural distortions.

7.1 Future Work

This study opens up the possibility of increasing the number of state-of-the-art algorithms used in the experiment. More datasets can absolutely be part of any future study on this topic. With the increased computational resources, more algorithms trained, validated, and tested across more diverse datasets will help researchers and application engineers choose among the algorithms and datasets with more reliable data.

Bibliography

A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid. ViViT: A Video Vision Transformer. 2021.

V. G. at Universität Konstanz. Konstanz artificially distorted image quality set (KADIS-700k). URL `http://database.mmsp-kn.de/kadid-10k-database.html`.

H. Bao, L. Dong, S. Piao, and F. Wei. BEIT: BERT Pre-Training of Image Transformers. 2022.

S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. Deep Neural Networks for No Reference and Full-Reference Image Quality Assessment. 2018.

A. C. Bovik. The Handbook of Image and Video Processing. 2005.

A. C. Bovik and Z. Wang. *Modern Image Quality Assessment*. Morgan Claypool, 2006.

L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale Vision Transformers. 2021.

J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual Attention Network for Scene Segmentation. 2019.

D. Ghadiyaram and A. C. B. and. Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. 2016.

D. Ghadiyaram and A. Bovik. Live in the wild image quality challenge database, 2015. URL `http://live.ece.utexas.edu/research/ChallengeDB/index.html`.

J. Gu, G. Meng, Shiming, and C. Pan. Blind image quality assessment via learnable attention-based pooling. 2019.

J. Gua, Z. Wangb, J. Kuenb, L. Mab, A. Shahroudyb, B. Shuaib, T. Liub, X. Wangb, L. Wangb, G. Wangb, J. Caic, and T. Chenc. Recent Advances in Convolutional Neural Networks. 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. 2016.

V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. 2020.

J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. 2018.

I. Hussein AL-Qinani. A review paper on image quality assessment techniques. 2019.

G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao. PIPAL: A Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. 2020.

L. Kang, P. Ye, Y. Li, and D. D. 1. Convolutional Neural Networks for No-Reference Image Quality Assessment. 2014.

J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. MUSIQ: Multi-scale Image Quality Transformer. 2021.

Y. Li, L.-M. Po, L. Feng, and F. Yuan. No-reference Image Quality Assessment with Deep Convolutional Neural Networks. 2016.

Z. Li, W. Yang, S. Peng, and F. Liu. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. 2020.

H. Lin, V. Hosu, and D. Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3, 2019. doi: 10.1109/QoMEX.2019.8743252.

K.-Y. Lin and G. Wang. Hallucinated-IQA: No-Reference Image Quality Assessment via Adversarial Learning. 2018.

X. Liu, M. Pedersen, and J. Y. Hardeberg. CID:IQ – A New Image Quality Database. URL `https://www.ntnu.edu/web/colourlab/software`.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021.

K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017. doi: 10.1109/TIP.2016.2631888.

X. Ma, S. Zhang, Y. Wang, R. Li, X. Chen, and D. Yu. ASCAM-Former: Blind image quality assessment based on adaptive spatial channel attention merging transformer and image to patch weights sharing. 2022.

A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012a. doi: 10.1109/TIP.2012.2214050.

A. Mittal, A. K. Moorthy, and A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. 2012b.

K. O'Shea and R. Nash. An Introduction to Convolutional Neural Networks. 2015.

D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang. Blind Predicting Similar Quality Map for Image Quality Assessment. 2018.

N. Parmar, A. Vaswani, J. Uszkoreit, Łukasz Kaiser, N. Shazeer, A. Ku, and D. Tran. Image Transformer.

N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. 2015a.

N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and J. Kuo. Image database TID2013: Peculiarities, results and perspectives. 2015b.

E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. PieAPP: Perceptual Image-Error Assessment through Pairwise Preference. 2018.

H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database release 1, 2003. URL `http://live.ece.utexas.edu/research/quality`.

H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. 2006a.

H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database release 2, 2006b. URL `http://live.ece.utexas.edu/research/quality`, `https://utexas.app.box.com/v/databaserelease2`.

S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang. Blindly Assess Image Quality in the Wild Guided by A Self-Adaptive Hyper Network. 2020.

M. Taboga. "Convolutions", Lectures on probability theory and mathematical statistics, 2021. URL `https://www.statlect.com/glossary/convolutions`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. 2017.

Z. Wang, A. Bovik, , and B. Evans. Blind measurement of blocking artifacts in images. 2000.

Z. Wang, H. R. Sheikh, and A. Bovik. No-reference perceptual quality assessment of JPEG compressed images. 2002.

J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann. Blind Image Quality Assessment Based on High Order Statistics Aggregation. 2016.

S. Yang, Q. Jiang, W. Lin, and Y. Wang. SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment. 2019.

S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang. MANIQA: Multi-dimension Attention Network for No Reference Image Quality Assessment. 2022.

Z. Ying, H. Niu1, P. Gupta1, D. Mahajan, D. Ghadiyaram, and A. Bovik. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. 2020.

J. You and J. Korhonen. Transformer for image quality assessment. 2020.

L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. Torr. Dual Graph Convolutional Network for Semantic Segmentation. 2019.

P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. 2021.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. Berkeley-Adobe Perceptual Patch Similarity (BAPPS) Dataset. URL `https://richzhang.github.io/PerceptualSimilarity/`.

M. Zhu, G. Hou, X. Chen, J. Xie, H. Lu, and J. Che. Saliency-Guided Transformer Network combined with Local Embedding for No-Reference Image Quality Assessment. 2021a.

M. Zhu, G. Hou, X. Chen, J. Xie, H. Lu, and J. Che. Saliency-Guided Transformer Network combined with Local Embedding for No-Reference Image Quality Assessment. 2021b.