Master in Computational Colour and Spectral Imaging (COSI)

Explainable Artificial Intelligence for Image Quality Assessment

Master Thesis Report

Presented by

Ha Thu Nguyen

and defended at the

Norwegian University of Science and Technology

September 2023

Academic Supervisor(s): Dr. Steven Le Moan, and Dr. Seyed Ali Amirshahi

Jury Committee:

1. Dr. Joni Hyttinen, University of Eastern Finland, Finland
2. Dr. Basura Fernando, A*STAR, Singapore

Submission of the thesis: 10th August 2023
Day of the oral defense: 5th September 2023

Abstract

Image quality assessment has been an active research field for decades because of the high demand for images and video content in daily life. As visual information is processed in various steps from acquisition and storage to transmission, they are often degraded by multiple types of distortions. It is necessary to evaluate the quality of any imaging system to maintain the user's experience. Thus, objective image quality assessments were proposed to objectively evaluate the image quality as close to the perceptual quality rated by human users.

Among the three types of image quality assessment, No-Reference image quality assessment (NR-IQA) has the most potential to be used in various applications and is also the most challenging topic. The traditional NR-IQA metrics were proposed using domain knowledge of natural images to extract hand-crafted features that can indicate the degradation degree of the distorted image. Recently, many deep learning models have been used in NR-IQA and outperform the traditional method in predicting image quality. However, they are still data-driven models which contain numerous parameters and lack explainability. Therefore, it is challenging to understand how such deep NR-IQA models estimate the quality of images and why they do not work on some images. Moreover, although many different methods of explaining a deep learning model have been introduced, there is no work that targets to image quality assessment.

In this work, we address the research gap in the explanation for the deep NR-IQA model. Firstly, we defined a set of definitions and expectations for explainable artificial intelligence (XAI) in the field of image quality assessment. Then, we proposed a framework to provide explanations at different levels: from global to local prediction for the model. The global explanations were formed by analyzing the images that the model can not predict their quality accurately. To find such an image, we proposed to use objective detection methods for IQA models. We also used different existing XAI methods to obtain explanations for the model in different information domains from spatial, and frequency to color space.

Different explanation results are discussed in our project. We found out that the existing XAI methods can explain NR-IQA models to some extent. However, there is no current way to evaluate the effectiveness of those explanations for image quality assessment problems. Future work is needed to provide an objective evaluation of XAI for image quality assessment or to find an alternative method to better explain NR-IQA models.

Acknowledgment

First of all, I would like to express my gratitude to my supervisors, Associate Professor Steven Le Moan and Associate Professor Seyed Ali Amirshahi, for their guidance, understanding and patience and support during my thesis. I would not be able to complete this project without their advice and supervision.

I would also like to thank all of my COSI classmates, especially, Abhinav Reddy Nimma, Viktoriia Kuznietsova and Asma Alizadeh Mivehforoushi, who spent time and supported me during the last two semesters.

I would like to gratitude the EU committee and all the COSI coordinators for giving me the chance to participate in this program.

Finally, I want to take this opportunity to express my deepest gratitude to all my family and my friend, who always support and care for me no matter what. Your unconditional love and encouragement are the main strength for me to go through all the tough moments during the last two years.

Acronyms

IQA: Image Quality Assessment
FR-IQA: Full-Reference Image Quality Assessment
NR-IQA: No Reference Image Quality Assessment
BIQA: Blind Image Quality Assessment
HVS: Human Visual System
CSF: Contrast Sensitivity Function
MOS: Mean Opinion Score
DMOS: Difference Mean Opinion Score
PLCC: Pearson Linear Correlation Coefficient
SRCC: Spearman Rank Correlation Coefficient
KRCC: Kendall Rank Correlation Coefficient
XAI: Explainable Artificial Intelligence
CNN: Convolutional Neural Network

Contents

1 Introduction

Visual information has been known as one of the richest data representations for humans to acquire information. According to Sharma et al. (2012), almost 57% of information processing in the human brain is from visual communication, and around 90% of data received in our brain is visual. Nowadays, advances in technology have opened access to numerous resources of visual content through the Internet for various purposes such as communication, education, entertainment, and so on. For example, social media networks such as Facebook or Instagram have a millions number of images uploaded daily on their platform (Zhu et al., 2020). The pandemic times also witnessed the rise of the working remotely trend of employees in many companies, which resulted in the development of various videoconference tools such as Zoom, Skype, Microsoft Teams, etc. As the demand from the global population for producing and sharing visual information is still increasing, this type of content has played a more crucial role in our daily life.

Visual content such as images undergo many distortions during the process, at various stages, from acquisition, storage, and transmission to display. Because of the important role of such content, perceptual quality assessment of images and videos has become an essential problem for evaluating image processing algorithms and systems. Researchers have spent a lot of attention on this field of study, which is called image quality assessment.

1.1 Image Quality Assessment

Image Quality Assessment (IQA) can be classified into *subjective* and *objective* quality assessment. In subjective quality assessment, human observers are asked to rate the quality of images. Their judgments could be formed on the technical quality or the aesthetic feeling that they perceive from the images, depending on the design of the experiment. The former relates to perceptual distortions such as noise, blur, and compression artifacts while the latter emphasizes the beauty and artistic value of images. In this work, we focus on the technical quality aspect, which expresses the degree of distortions perceived in the image. The subjective

rating for each image is usually reported by taking the average of scores from all observers and is termed as Mean Opinion Score (MOS), which represents the opinion of a statistically average observer. Subjective quality assessment is the most reliable method of measuring the perceptual quality of images because it is calculated from human perception. However, this approach has many drawbacks: the requirement of a large number of observers to form a reliable MOS, the long time for preparation and recruiting participants, the inability to reproduce results, etc. Therefore, it is not convenient to conduct a subjective quality assessment in the image processing algorithms for evaluation purposes. To address this issue, objective quality assessment methods are designed with the goal of automatically predicting the quality of images as perceived by humans.

Objective quality assessment models, or image quality metrics (IQMs) are the solutions to objectively measure the quality of images. IQMs consist of three frameworks: Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA), and No-Reference IQA (NR-IQA). If a reference image, which is an undistorted version of the degraded image, is existed as accessible, FR-IQA methods are used to estimate the quality of the images in comparison with the reference. If only some information such as distortion type or histogram of the original image is available, RR-IQA methods are used. In most cases, the reference data is completely inaccessible or does not exist, NR-IQA methods are the solution to assess the quality of an image. Because of the unknown reference image, NR-IQA is also called Blind IQA (BIQA). The FR-IQA and RR-IQA methods usually achieve outstanding performance because of the usage of the reference image. However, there is a limited situation in which the reference data is available. Thus, these two types of IQA are not popularly used in practical applications. In contrast, the NR-IQA methods are more applicable to the image system. Therefore, in this work, we focus on the NR-IQA.

1.2 Motivation

For many years, image quality assessment has been a significant subject of research in various fields such as computer science, neuroscience, psychology, and so on. Among many directions of this topic, NR-IQA is a branch that has attractive potential for various applications. From an early start, many NR-IQA metrics were designed using domain knowledge of natural images to extract hand-crafted features that can distinguish distorted images from pristine ones. However, this approach shows poor performance on the images that undergo natural distortions, which are caused during the acquisition and subsequent processing. Recent advancements in image quality assessment have demonstrated the effectiveness of deep learning models in predicting the perceptual quality of such images. These models have

outperformed hand-crafted models (Athar and Wang, 2019). With the introduction of large-scale subjective datasets (Hosu et al., 2020; Ghadiyaram and Bovik, 2015; Lin et al., 2019), deep learning models have learned more generalized features to make better predictions about image quality. Still, they are data-driven models that typically rely on deep convolutional neural networks with a numerous number of parameters and limited explainability, often referred to as the *black box* effect (Zhou et al., 2019). Consequently, a challenge arises in understanding the underlying reasons behind the models' effectiveness and identifying the scenarios in which they may not perform well. In addition, the studies about both traditional and deep learning-based NR-IQA did not provide an objective method to find the image in which the metric fails to predict its quality.

Looking broadly, the lack of transparency within deep learning architectures has raised so much concern in Artificial Intelligence (AI) community. It restricts the deployment of deep learning models in critical sectors in which a bad decision could risk a human life such as healthcare. Researchers have published many papers that aim to make such models become more transparent. These approaches are commonly referred to as Explainable Artificial Intelligence (XAI), which nowadays is an important domain in AI.

Even though many different methods of explaining a deep learning model have been introduced, most of them are designed for classification or segmentation models, which are different from image quality assessment problems. XAI also has many forms: from the explanation of single prediction to the explanations in relationship with datasets. To the best of our knowledge, there is no prior work that targets providing explanations for deep NR-IQA models, especially at the dataset level. This lack of study limits our understanding of the NR-IQA model's performance. Consequently, it prevents the potential of developing more effective models in image quality assessment.

1.3 Research questions

Considering the research gap mentioned in Section 1.2, this thesis aims to investigate XAI for NR-IQA models. The research questions can be formulated as follows:

- Q1. What constitutes a good explanation for image quality assessment?
- Q2. How to leverage recent advances in XAI to create a framework that is specifically dedicated to IQA models?

1.4 Thesis outline

The thesis organization is as follows. In Chapter 1, we defined the main topic of the project and the research questions. In Chapter 2, the background related to the topics of research(No-reference image quality assessment and explainable artificial intelligence) are presented. Next, the proposed method of providing an explanation for IQA is described in Chapter 3. Chapter 4 represents the experiments and discussion. Finally, the thesis is closed by Chapter 5, where we discuss the conclusions and potential future works.

1.5 Contributions

The contribution of this thesis for XAI for NR-IQA is as follows:

- Through a few rounds of interviews with different researchers working in the field we have defined a set of different definitions and expectations for XAI in the field of image quality assessment. This contribution will address the research gap Q1 in section 1.3.
- We proposed a framework to explain NR-IQA models, which address the research gap Q2 in section 1.3. In this framework, we used different existing XAI methods. Our main contributions in this project with this regard are:
  - The proposing of objective outliers detection methods for IQA models.
  - The extension of the perturbation-based methods to the frequency domain and color space.
  - The investigation of the use of the existing XAI methods on NR-IQA problems.

** Disclaimer: Grammarly is the only software tool that was used in writing this report for grammatical correction purposes.*

2 Background and Literature Review

This chapter reviews the current state-of-the-art in No-Reference Image Quality Assessment and Explainable Artificial Intelligence.

2.1 Human Visual System

As image quality assessment aims to estimate the perceptual quality of an image, an *ideal* IQA model should behave similarly to the way humans perceive visual information in images. The study of Human Visual Systems (HVS) lies on the intersection of physiology and psychologic, which mainly focus on the eyes and brain. The eyes play the role of an image sensor or the camera to acquire visual information, while the brain is the component to process images. Although HVS is not fully discovered, many properties of it have been studied and accepted to help model the mechanism of human perception in machine and image processing algorithms.

The characteristics of the HVS have a significant impact on individual perception and evaluation. By taking into account certain limitations and features of the HVS, we can better understand how human perception works in the subjective assessment of image quality. Various IQA models (Wang et al., 2004; Wang and Bovik, 2002; Toet and Lucassen, 2003) were proposed with the principle of emulating the known property of HVS in the process of objective quality evaluation. We will discuss some of the factors of HVS that should be considered in IQA.

An important property of HVS was introduced by Legge (1981) which showed our ability to detect change in contrast depends on the spatial frequency of the visual stimulus. It is generally more sensitive to low-frequency distortions than the high-frequency ones. At each spatial frequency level, there is a visibility threshold of contrast that makes the visual stimulus become visible to HVS. The change of this threshold value at different frequency levels can be represented as *Contrast Sensitivity Function* (CSF), and is illustrated in Figure 2.1. The vertical acis

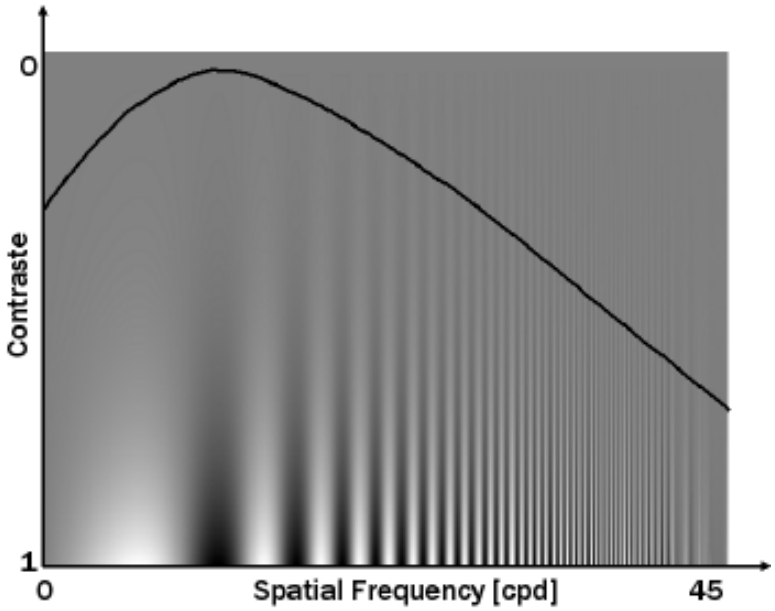


Figure 2.1: *Contrast Sensitivity function graph (image taken from (Hautière et al., 2007))*

represents the contrast amplitude and the horizontal xis indicates the spatial frequency.

The human visual system (HVS) has a poor response to color (chrominance) spatial detail compared to its response to luminance spatial detail (Poynton, 1997). It means that humans will notice the distortion happening in the achromatic area easier than the color ones. Taking advantage of this characteristic, in image compression, the data at the illumination channel is preserved more than those of the color channel. Thus, the size of the image after compression is reduced a lot, but the perceptual quality is not affected much.

We will consider these mentioned properties to provide an explanation of IQA models in Chapter 4.

2.2 No-Reference Image Quality Assessment

No-Reference image quality assessment which sometimes is referred as blind image quality assessment (Saad et al., 2012) is a group of objective image quality metrics that aims to predict the perceptual quality of an image without any inference image.

2.2.1 Traditional methods

NR-IQA is challenging because of its nature. Wang and Bovik (2011) suggested three types of knowledge that are necessary to build a successful IQM: 1) distortion types, such as compression or blur; 2) knowledge of the image source, which can be the reference image in the case of the FR-IQA, or the statistical information that can distinguish distorted images from good quality images in the case of NR-IQA; 3) knowledge is about the HVS, based on visual physiology studies about how humans perceive images. Many NR-IQA methods were proposed by using these three types of knowledge.

In the early era of NR-IQA studies, metrics were designed for specific distortions. For example, Marziliano et al. (2002) introduced a metric for blur images and video, based on the width of the edges in the spatial domain; Wang et al. (2002) proposed an NR-IQA metric for JPEG compression that considered blocking artifacts and blurring as the most significant reason for the quality degradation of the compressed images. In 2005, Sheikh et al. (2005) introduced a metric for JPEG2000 compression based on statistical properties of images on the wavelet domain.

Later on, with the availability of many subjective IQA databases, different studies have been proposed for various types of distorted images. Few NR-IQA methods that do not require training on human-rated scores are referred to as Opinion Unaware NR (OU NR). NIQE (Mittal et al., 2012b) is the pioneer in this direction. The NSS features are calculated from the image, then their distribution is captured by a Gaussian model. The quality of a given image is defined by the difference between the model fitted on the extracted features and those of the natural image. The ILIQUE (Zhang et al., 2015) extended NIQUE with three types of features: quality-aware gradient features, statistical features from the log-Gabor filters responses, and statistical features from color space. The results from these OU NR methods (QAC (Xue et al., 2013), LPSI (Wu et al., 2015)) are not competitive.

Opinion Aware NR are methods that were trained on distorted images whose human ratings are available. Most are Natural Scene Statistic (NSS) based methods, which were built based on the assumption that natural images hold certain statistical properties, while those are absent in the distorted images. For color images, the NSS based metrics are designed to work on the Y channel as distortion appears more in the illuminance than in color. Wang and Bovik (2011) proposed the DIIVINE metric, which uses the wavelet coefficients of the images to map the statistical feature to the quality score of each distortion category and perform final quality prediction. Saad et al. (2012) proposed an NSS-based metric, which is called BLIND-II, extracting image features on the frequency domain using the discrete cosine transform (DCT) coefficients. BRISQUE (Mittal et al., 2012a) uses scene statistics of neighboring (locally normalized) luminance coefficients to quantify

"naturalness" and the quality due to distortion presence. Some other metrics such as GWHGLBP (Li et al., 2016a) extract structural information as the handcrafted features or NRSL (Li et al., 2016b) uses both the structural and illuminance features. This group of metrics can predict well on one type of distortion or on a specific dataset, their performance is significantly lower on other types of distortion or other datasets.

The traditional NR-IQA methods achieve accurate predictions on synthetic distorted datasets but fail on authentic ones. Even so, they are explicit to the human user because we know which features in the image the metrics look for to make predictions about its perceptual quality. These methods are interpretable in the aspect of proving saliency maps, or the attribution of image features to the estimated score. However, they lack explainability in their relationship with the dataset, at global level.

2.2.2 Deep learning based methods

Recently, with the success of using deep learning in various computer vision tasks such as image classification, more attention has been paid to applying neural networks to NR-IQA. CNNIQA (Kang et al., 2014) was the first work that uses Convolutional Neural Networks (CNN) to predict perceptual image quality scores, the architecture of which is shown in Figure 2.2. In the original paper, the model was designed for gray scale images, but it can be extended to color images. Given an image, a local contrast normalization which is similar to BRIQUE is employed. Patches of size 32x32 are sampled from the normalized images in a way that they are not overlapped. The network was trained in a large number of image patches to predict the local quality score of image patches. The quality estimation for the image is defined as the average of the patches' score. Later on, they extended the networks into a multi-task CNN for simultaneously predicting image quality and identifying distortions (Kang et al., 2015). Although these methods exhibit better performance than the previous hand-crafted NR-IQA models, they lack transparency as there is no explanation of what features influence the model output.

The previous methods have a particular drawback: in the training process, the patches which are extracted from an image are associated with the same subjective quality scores. However, because of the variation in local image features, the local image quality should differ from the global image quality. Kim and Lee (2016) proposed a CNN-based NR-IQA model named BIECON, which predicts the local quality score at image patches before producing the global image quality. BIECON resolved the issue of lacking patch quality ground truth scores by employing local quality maps provided by a full-reference IQA metric, FSIM (Zhang et al., 2011). (Ma et al., 2017) followed the same approach in (Kang et al., 2015), in which two

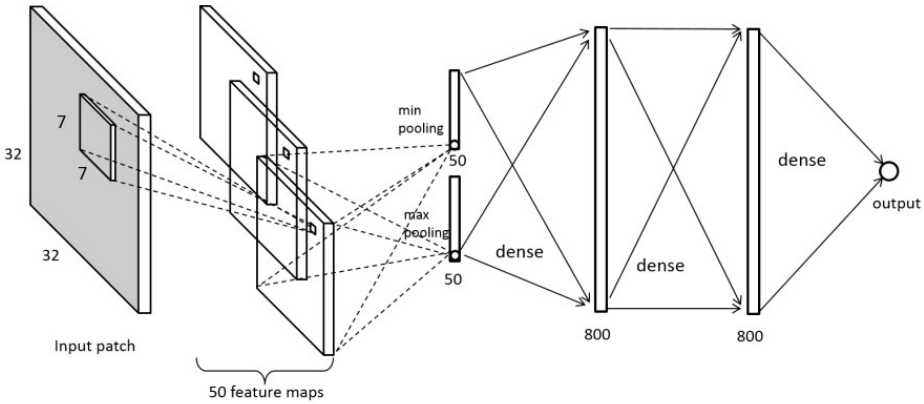


Figure 2.2: *The architecture of CNNIQA. (Image from (Kang et al., 2014))*

subtasks are performed: distortion identification and image quality estimation.

Bosse et al. (2017) proposed a deeper CNN network, which allows two different pooling strategies: a simple averaging of local patch qualities and weighted averaging quality aggregation. The former choice is referred to as DIQaM-NR, while the latter is called WaDIQaM-NR. They share the same architecture as in Figure 2.3, in which the feature extractor was inspired by the VGG model (Simonyan and Zisserman, 2014). The DIQaM-NR simply averages local estimated qualities to get the overall quality score of the image. Meanwhile, the WaDIQaM-NR uses a weighted pooling aggregation, which assigns each patch with a different estimated weight. The results from the models which were trained on TID-2013 (Ponomarenko et al., 2013) and tested on other datasets (CSIQ (Larson and Chandler, 2010), LIVE (Sheikh et al., 2006)) indicate that WaDIQaM-NR performs better than DIQaM-NR for most distortion types. This suggests that the relative importance of local quality is not uniformly distributed over an image. The local weight maps and quality maps from the WaDIQaM-NR can be considered as explanations for the global image quality. Still, how the model predicts quality for each patch is not explicit.

Yang et al. (2019) argue that image quality should be guided by visual attention, which is represented by a saliency map highlighting the regions in images people's eyes focus on. They proposed a SGDNet, which estimates saliency maps of the input image as a local weighting map for generating the global quality. The authors pointed out the limitation of the previous saliency-based methods, that the weighting maps are intermediate steps in the training process, thus they are not well optimized. Considering this issue, SGDNet is implemented as a multi-task network, which aims to predict both the visual saliency map and the image quality.

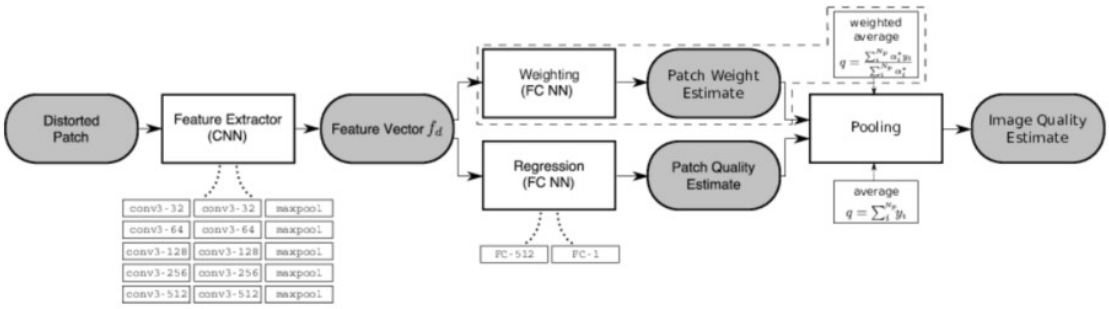


Figure 2.3: *The architecture of a deep network from (Bosse et al., 2017).*

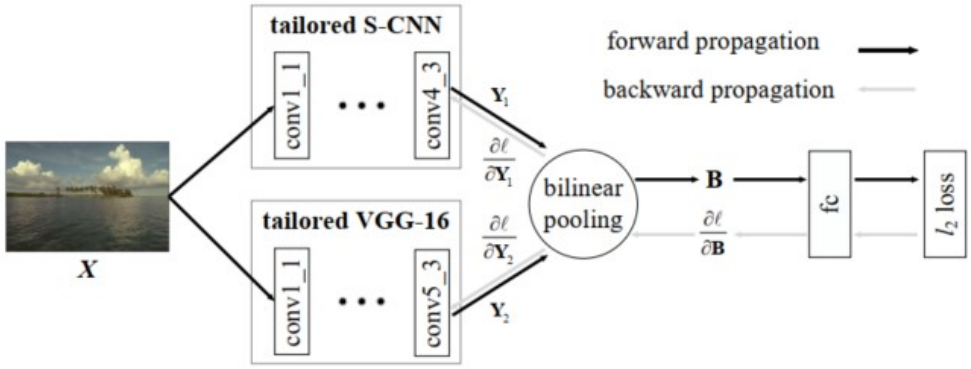


Figure 2.4: *The architecture of DBCNN (Image from (Zhang et al., 2018)).*

The saliency map, which represents the visual attention mask, is predicted by a supervised learning method. The ground truth for this process is provided by a guided saliency model.

Zhang et al. (2018) introduced a deep bilinear model, named DB-CNN, which consists of two feature extractors whose outputs are multiplied together to obtain image descriptors. Figure 2.4 shows the architecture of the network, which can handle both synthetic and authentic distortions. The CNN model for synthetic distortion (S-CNN) was trained on an IQA database for classification tasks, that outputs the class corresponding to the distortion type and level of distortion of the images. Meanwhile, the CNN for authentic distortion is adopted from the VGG-16 (Simonyan and Zisserman, 2014) that was trained for image classification task on ImageNet (Deng et al., 2009). The layers after the last convolutional layer of the pretrained S-CNN and VGG-16 are discarded. The features extracted from the two branches are pooled together into a representation for quality estimation. DB-CNN outperforms other competitors in predicting image quality scores of an in-the-wild dataset (Sheikh et al., 2006).

Su et al. (2020) proposed a self-adaptive hyper network, Hyper-IQA, that

follows the top-down flow of human perception, in which the quality prediction is mapped with content awareness. The network extracts both local features and global semantic features from a given image. The image quality is predicted by aggregating this representation at multi-scale levels. Ying et al. (2020) demonstrated that using both local patch quality and global image quality would provide a better image quality prediction. The authors introduced a network that uses the ResNet (He et al., 2016a) as the backbone for feature extraction. The networks predict the perceptual quality map (at patch level), to improve the global image quality estimation (at picture level), so it is called PaQ-2-PiQ. PaQ-2-PiQ is trained on a large database, which contains both image quality and patch quality labels of realistic distorted images. By selecting a suitable size of patches, this metric can be able to generate the spatial quality map for the whole image. The patch-wise map shows the regions in the image that have poor and high quality, estimated by the model. It can be considered as an explanation for the final prediction of the image quality score.

Fang et al. (2020) introduced an IQA database of smartphone photography, named SPAQ, which consists of authentically distorted images captured from various smartphone cameras, subjective quality judgments, and additional information such as image attributes, scene category labels, and EXIF tags. They also proposed three objective quality models constructed by baseline and multi-task deep networks. The baseline model adopted the ResNet-50 (He et al., 2016b), a residual network as the backbone feature extractor for image quality prediction. The multi-task models are modified from the baseline, to predict image attributes or the EXIF tags jointly with the quality score. Their experiments on the same dataset show that the multi-tasks model outperforms the baseline in predicting image quality. This result suggests additional information about the images such as image attributes, or the EXIF tags are useful in improving the prediction accuracy of the NR-IQA models.

Hosu et al. (2020) proposed an end-to-end BIQA architecture, which is illustrated in Figure 2.5. The network consists of a CNN backbone, followed by a global average pooling layer, and four fully connected layers. The authors tried different CNN base architectures: VGG16, RestNet101, InceptionResNetv2, and so on. The proposed models were trained and tested with each CNN feature extractor. Their experimental results suggest deeper base networks perform better in predicting image quality, and the best model is the one with the InceptionResNetv2 base.

Although CNN-based models are popular among NR-IQA, they have limitations of missing non-local information and producing strong local bias which leads to inefficiency when a complex combination of features is needed (Golestaneh et al., 2022). Inspired by Natural Language Processing (NLP), which uses Transformers to capture the dependencies of language sequences, recently NR-IQA models

Figure 2.5: *The architecture of the KonIQ model (Image taken from (Hosu et al., 2020)).*

((Zhu et al., 2021), (Golestaneh et al., 2022))adopted vision transformer (ViT) (Dosovitskiy et al., 2020) into their works. Golestaneh et al. (2022) introduced a hybrid network that combines CNNs and a Transformer block for image quality prediction. Local features are extracted by the CNNs at different spatial scales, while Transformer captures the interaction of those features. In addition, the authors applied relative ranking and self-consistency loss to improve the robustness of the proposed model. Using the similar approach of capturing local features at multi-scale of images, Ke et al. (2021) does not use the CNNs, but creates multiple-sized variants of the image (2.6). Each image is parted into patches of fixed size and fed into the model with a scheme to embed patch position and scale information to ViT. The transformer performs multi-head self-attention and produces a sequence as the final representation. In the end, a fully connected layer is applied to predict the image quality.

Yang et al. (2022) also used ViT as a feature extractor, but they applied self-attention across the channel of an image instead of the spatial dimension. They first extracted the features from four layers of the ViT. To assign different weights to each layer depending on their importance in image quality, an attention block is modified for capturing the global interaction between channels. Then, the intermediate representations are fed into the scale swin transformer block for boosting the local connection between image patches. Finally, a dual branch of weighting and scoring estimation is used to predict the quality score and the importance of each patch.

Although the deep learning-based NR-IQAs outperform their traditional counterparts such as NIQE or BRISQUE, they lack explainability. In traditional methods, knowledge about HVS or natural images is used to explain the selection of image properties that play an important role in the quality of an image. For example, the width of edges in the image, or the structural information is determined as key image quality factors by Marziliano et al. (2002) and Li et al. (2016a). Thus, they are more transparent to humans and we can understand how the metric predicts the quality of an image. In deep learning, these features are learned by a neural network, which consists of many layers via nonlinear relations. Even if the relation

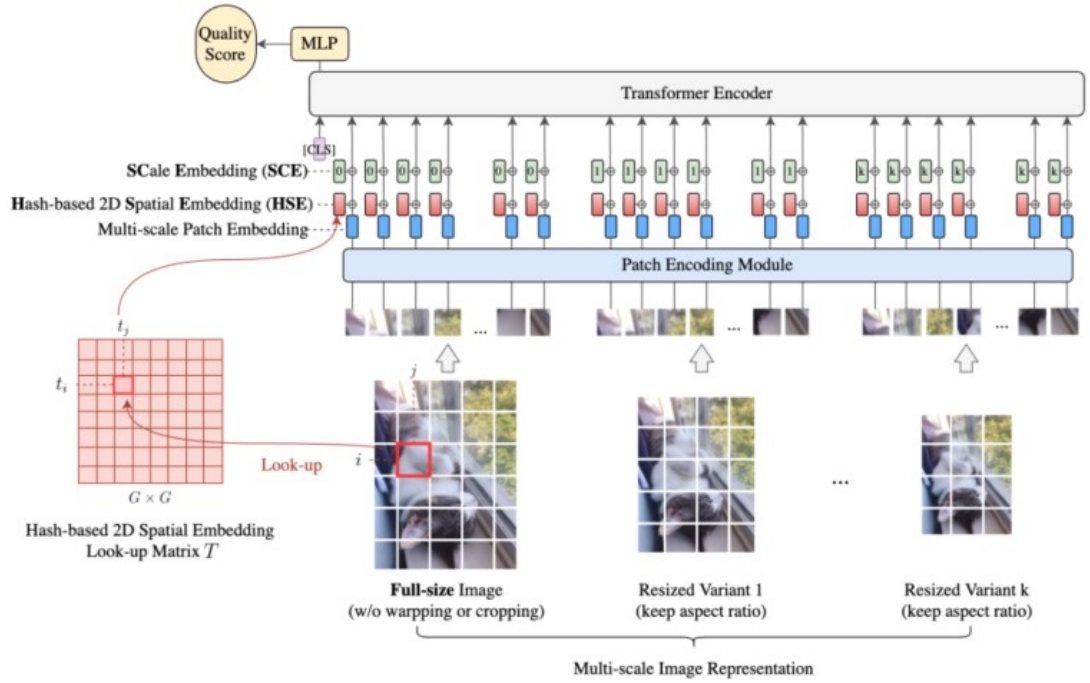


Figure 2.6: *Model overview of MUSIQ (Ke et al., 2021).*

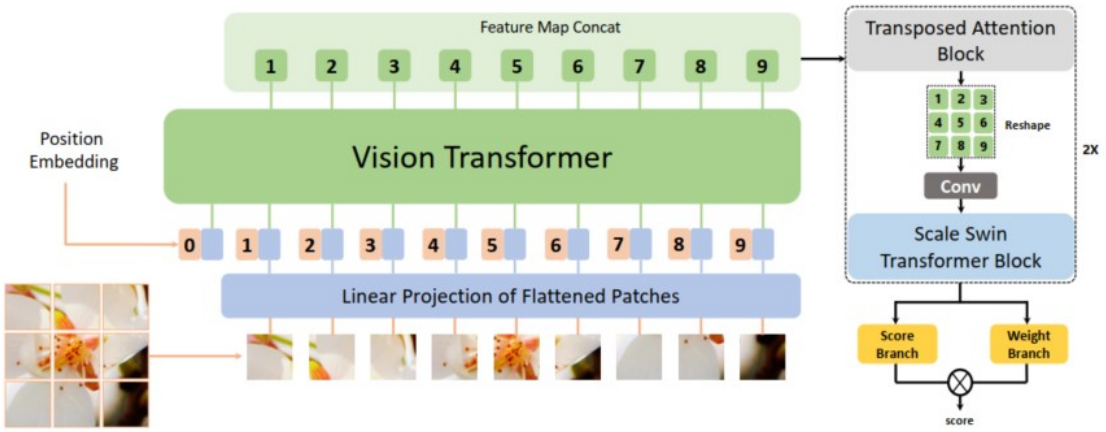


Figure 2.7: *The architecture of MANIQA from (Yang et al., 2022).*

between each layer is investigated, it is still unfeasible to fully understand how the model comes to a decision (Van der Velden et al., 2022). For example, given an image of a blue sky and an image of a cat, we do not know why an IQA model estimates one image has higher quality than the other. Another concern that such data-driven models may be biased in some way without notice. A NR-IQA model

may work well with one type of distortion but fail to predict the quality of images that are degraded by other distortions. Without understanding how the model comes to a quality estimation, one can not blindly trust them. Therefore, it is important to shed light on NR-IQA prediction.

2.2.3 Evaluation of the metrics

For evaluating the performance and training of image quality metrics, the most common way is by comparing the objective scores and the scores obtained by subjective quality assessment. The set of images and the score rated by human observers that are obtained by conducting psychophysical experiments is called an image quality assessment dataset.

2.2.3.1 Image Quality Assessment Databases

Over the years, a significant number of IQA datasets have been published. IQA datasets can be classified as *synthesized* and *authentic* databases depending on the characteristic of distortions appearing in the images.

Synthesized databases contain distorted images that are degraded from the reference images by multiple distortions of different levels. This type of dataset can be used in evaluation for both full-reference and no-reference IQAs. The LIVE dataset which was created by Sheikh et al. (2006) is one of the most popular IQA databases. This dataset contains 29 reference images of different resolutions and 750 degraded images. The distorted images were obtained by applying five types of distortions: JPEG200, JPEG, white noise, Gaussian blur, and fast Rayleigh decay on the pristine images. The subjective scores for each image were processed, and the Difference Mean Opinion Score (DMOS) of the range [0, 100] is provided. The lower value of DMOS indicates the higher quality of the image. The TID2013 dataset (Ponomarenko et al., 2013) includes 25 reference images and 3000 distorted images. The authors used 24 distortion types, at 5 distortion levels to simulate the image degradation. Along with the images, the corresponding DMOS is provided. The value range of DMOS is from [0, 9] with the larger value suggesting lower image quality. CSIQ is another simulated IQA dataset that was constructed by Larson and Chandler (2010). The dataset consists of 30 reference images and 866 degraded images of six distortion types, which include overall contrast reduction. In this dataset, the subjective quality assessment is also represented in DMOS of the value range from [0, 1]. The LIVE, TID2013, CSIQ have been used to evaluate the performance of the IQM. Many deep learning-based IQA models were proposed and trained in these databases. However, the small number of images can lead to an overfitting phenomenon of the models. Lately, Lin et al. (2019)

created two datasets, the KADID-10k and the KADIS-700k. The former contains 81 reference images and the corresponding degraded images of 25 distortions at five levels. The latter has 140,000 pristine images with five degradation versions distorted by random distortion types. The subjective judgments were conducted by crowdsourcing. The DMOS and the variance score for each image are published. The range of DMOS is [1, 5] with the higher value indicating better image quality. By far, these two databases are the largest simulated datasets.

Unlike the synthesized datasets, in authentic databases, the images are captured directly from the real-world environment. Thus, the distortions are natural. The CLIVE (Ghadiyaram and Bovik, 2015) and KonIQ-10k (Hosu et al., 2020) are the most common authentic datasets. The CLIVE or LIVE in the Wild was created by the same laboratory as the LIVE dataset. The database consists of 1162 images captured by mobile cameras. Thus, the distortions in the images are the results of the camera processing pipeline. The subjective ratings were collected from an online crowdsourcing system and processed to produce the Mean Opinion Score (MOS) for each image. The KonIQ-10k was introduced by the same authors of the KADID with the aim of a large in-the-wild dataset. 10,073 images were selected considering the diversity of content, the natural distortions, and the distribution of image quality indicators. Around 1,2 billion subjective ratings were obtained and processed to provide the overall opinion score. In addition to the MOS, detail of the image quality indicators and each category rating are also provided.

For the performance evaluation of an NR-IQA, all of the mentioned datasets can be used. However, there are disadvantages when using full-reference databases to assess the no-reference image quality assessment model. We will discuss these issues and a possible way to minimize their impacts on the evaluation process in Section 4.

2.2.3.2 Evaluation Criteria

The most common way to compare the performance of the image quality metrics is based on the correlation coefficient between the quality estimation scores predicted by the IQA methods and the subjective scores. The higher the correlation value is, the better the IQA model is.

Pearson Linear Correlation Coefficient

The Pearson Linear Correlation Coefficient (PLCC) was developed by Karl Pearson and is used as a measure of the linear relationship between two variables. Given two sets of subjective scores X and estimated quality scores Y of N images, the Pearson correlation coefficient can be computed using equation 2.1, in which X_i and Y_i are the subjective ratings and the estimated quality of the i^{th} image, $\bar{X}$ and $\bar{Y}$ is the mean values of subjective scores and all estimated quality score of all the images.

$$PLCC(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X}).(Y_i - \bar{Y}}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2.\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{2.1}$$

Because the scores estimated by IQA methods are usually not linear with the subjective opinion, a nonlinear fitting step is used before calculating of the PLCC. Any logistic function, which is monotonic can be used for this regression so that the order of the quality values is reserved. However, inherited from Sheikh et al. (2006), a five-parameter logistic function as in the equation 2.2 is commonly used.

$$f(x) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + \exp\beta_2(x - \beta_3)}\right] + \beta_4 x + \beta_5 \tag{2.2}$$

In the equation 2.2, $f(x)$ denotes the IQA score after the nonlinear mapping, x denotes the objective score produced by the IQA methods, β_1, β_2, β_3, β_4, β_5 are the model parameters that are found by using optimization function.

Spearman Correlation Coefficient

The Spearman Rank-order Correlation Coefficient (SRCC) is used to measure the monotonicity of the relationship between two vectors. SRCC is a non-parametric rank-order correlation and does not need the nonlinear fitting step. While PLCC assesses the linear relationship between the two variables, SRCC is equal to the SRCC between the rank values of these variables. The SRCC value between the subjective scores and the quality score estimated by an IQA method can be computed using the following equation:

$$SRCC(X,Y) = 1 - \frac{g\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{2.3}$$

In the equation 2.3, d_i denotes the difference between the ranks in subjective scores and estimated scores of the i^{th} image.

Kendall Correlation Coefficient

Kendall Rank Correlation Coefficient (KRCC) is another rank-order based method to measure the ordinal association between two variables. This correlation has a similar principle as the Spearman correlation but uses the difference between the probabilities instead of the difference between ranks. Let (x_i, y_i) is a set of subjective rating and an estimated score, a pair of concordant (x_i, y_i) and (x_j, y_j) is defined if the rank order of their variables agrees. In other words, if either both $x_i > x_j$ and $y_i > y_j$ or both $x_i < x_j$ and $y_i < y_j$, the pair of (x_i, x_j) and (y_i, y_j) is a consonant pair. Otherwise, it is a dis-concordant one. The Kendall correlation is computed as:

$$KRCC(X,Y) = 1 - \frac{2(numberofdiscordantpairs)}{\frac{N(N-1)}{2}} \tag{2.4}$$

In the equation 2.4, N is the number of images in the database. As KRCC is found highly consistent with SRCC and does not provide much more information, in the literature, KRCC is normally not reported.

Other evaluation metrics such as Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) report how accurate the predicted quality score is with regard to the subjective quality score. They are simple, computed by using the equation 2.5, 2.6, in respectively. However, these performance metrics require subjective scores and the objective scores are in the same range of value, with is not always true because the objective IQA methods are usually designed without the normalization of output score.

$$RMSE(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2} \tag{2.5}$$

$$MAE(X,Y) = \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i| \tag{2.6}$$

It has been shown that the deep neural network models which were trained on large-scale datasets achieve better results than the hand-crafted metrics. However, that high performance is highly dependent on the dataset that the models were trained on. Therefore, it is suggested to cross-test the models on many datasets. In addition, because the images which were collected in the same dataset share similar properties, for example, type of distortions; cross-dataset evaluation can be used to test the generality of the model. In this type of assessment, the model is trained on one dataset and evaluated on another by using correlation coefficients.

Image quality assessment in general, and NR-IQA in particular have been extensively studied for decades. The IQA models are usually assessed by using the correlation coefficients which were mentioned above with the subjective score on an IQA database. With recent advances in AI, many deep learning base models have been proposed and achieved excellent performance in predicting the quality of images. However, the explainability of them is still missing. Although correlation is good to evaluate how a model performs it does not give us any information about why a metric works well for some kind of an image and bad for another. It does also not provide us with information about why a model works better than another. This is why in the case of deep learning methods we are dealing with a *black box*, we would need look into the explainability of the models to have a better understanding of the metric and also be able to improve the performance of the model. This work aims to fill this gap by providing explanations for NR-IQA

models. In the next section, we will review the state-of-the-art explainable artificial intelligence methods and discuss their potential in interpreting IQA models.

2.3 Explainable Artificial Intelligence

Deep learning models are not only becoming popular in IQA, but they have been also immersed more in our daily life. While the models are achieving better predictive performance, their complexity is increasing. Thus, it is becoming more difficult to understand the underlying behind the models. There is usually a trade-off between the explainability and the accuracy of the models. (Figure 2.8). Simple models such as rule based or decision trees have low performance, while deep learning models exhibit high accuracy and little explainability. As the opacity of the AI models makes them a *black-box* to humans, it prevents the users from completely trusting and using the systems. For example, in the healthcare sector, if a doctor does not know if the model predicts a patient has a high chance of heart stroke because of their gender or their health record, they can not trust that prediction even though the model achieves high accuracy on a test set. Therefore, the need for explanations on how the AI-based system makes the decision is highly in demand. This led to the release of a new area of research, which focus on providing an understanding of AI models, called Explainable AI (XAI).

Recently, the number of studies in XAI has increased significantly. Many survey papers were published that provided an overview of the current research situation of XAI. Some of them ((Adadi and Berrada, 2018), (Angelov et al., 2021), (Belle and Papantonis, 2021), (Linardatos et al., 2020), (Samek et al., 2021)) reviewed XAI methods, Speith (2022) discussed the different approaches to construct taxonomies of XAI. Nauta et al. (2022) provided a systematic review of the evaluation of XAI. The term *interpretability* and *explainability* are usually used interchangeably with the meaning of providing ways to improve the understanding of users about an AI system. As the image is input data of IQA models, in the scope of this project, we will go through the XAI methods that can be applied to images. We follow the framework from Adadi and Berrada (2018) and Murdoch et al. (2019) to classify XAI techniques using three criteria: model-based versus post hoc, the scope of the explanation, the applicability of the XAI method. The following subsections will describe the methods in each category. In this report, we will use the term "model" to refer to the pretrained AI models and the term "methods" or "techniques" to indicate the XAI explanation.

Model-based versus post-hoc methods: The model-based explanation refers to the models, for example, a linear regression model, which is simple to be understood but can find well the relationship between the output and input (Murdoch et al., 2019). Thus, they are also referred to as *transparent models* or

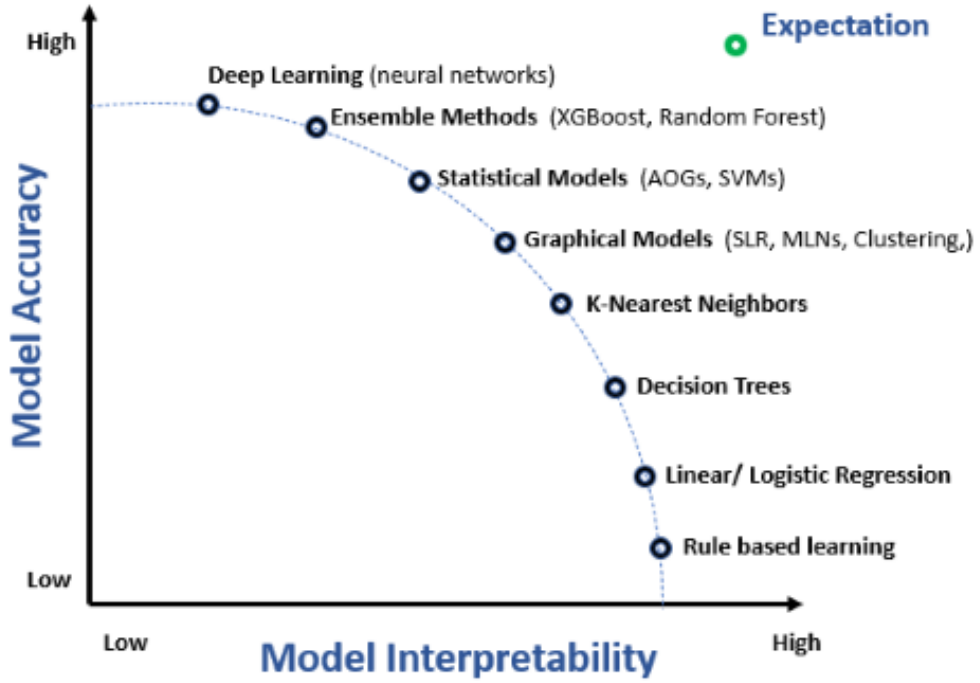


Figure 2.8: *Model interpretability vs. model accuracy for machine learning and deep learning algorithms (Joshi, 2021).*

white-box. They are usually the traditional machine learning methods such as linear regression or decision trees. The interpretation of these models comes directly from the algorithms. Such models that are intrinsically interpretable often have low accuracy. Meanwhile, more complicated models (for example neural networks) achieve better performance and usually do not provide an explanation themselves, which make them become *black-box* to human. The groups of XAI methods that aim to analyze the insight of such models are called post-hoc explanations. Most of the works on XAI belong to this category.

Scope of the explanation: As the name indicates, the XAI methods can be classified based on the scope of the explanation that it can provide: for the entire models or for a single prediction. The global explanations provide general reasons for all outcomes that the model makes. They are also called dataset-level explanations because this understanding of the model is obtained from investigating multiple input instances. For example, how much a feature contributes to the output in the entire dataset. Bau et al. (2017) is another global explanation that provides information about the semantic concept which are embedded latent spaces

of the convolutional neural network. Saleem et al. (2022) surveyed the global interpretation methods and found that most of them were constructed by using local explanation techniques and required computation cost. On the other hand, local explanation provides the reason for a specific decision of the models. For example, why an image is recognized as a dog image by a classification model. As the interpretation for a single output is usually readable to humans, the majority of XAI techniques are local explanation methods.

Applicability-based methods: This criteria distinguish if an XAI method can be applied for only specific types or any type of AI models. The former is referred to *model-specific* explanation while the latter is called *model-agnostic*. By definition, model-based methods are model-specific methods, but a model-specific method is not always model-based methods (Adadi and Berrada, 2018). While there is a limit in the choice of models that can be explained by model-specific XAI techniques, model-agnostic is model-independent. For example, XAI methods that aim to visualize the learned filter of CNNs can not be applied to other types of deep learning models. Meanwhile, the methods which find the importance of input features to the output by modifying input data can be used to explain any black box model.

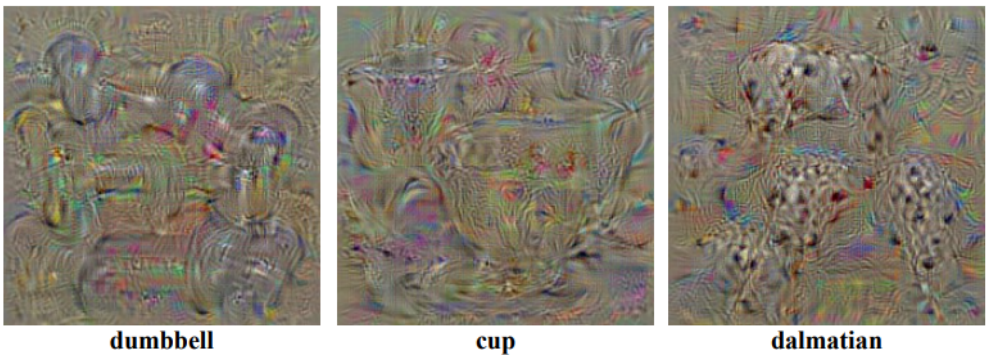
Because an XAI method can be classified into multiple groups based on different criteria, we provide here a brief review of the popular explanation techniques that can be applied to image data.

2.3.1 Visual explanation

The visual explanation provides the attribution map, which demonstrates the importance of regions in the input image, or in the intermediate representation of the networks that cause the model's decision. The common XAI visualization methods will be discussed in the following paragraphs.

The very first attempt from (Simonyan et al., 2013) visualized image classification models. The authors proposed two methods of visualization based on computing the gradient of the output corresponding to the change in the input image. The first technique generates an artificial image that represents the features of the class captured by the classification. The second provides a saliency map for a specific image and class, which highlight the region in the images that suggest it belongs to the given class. This method is referred to by the name Gradient in literature. Figure 2.9 shows the visualization of these techniques. In subfigure 2.9a, we can see that the images do demonstrate some properties of the corresponding class, i.e: the shape of the dumbbells, cups, and the color of the dog. However, these images look unrealistic and contain a lot of noise information. The saliency maps in the bottom row show the position of the main object in the

images with poor discrimination between the foreground and background. Zeiler and Fergus (2014) proposed a method - Deconvolution which improved Gradient with a modification of backpropagating only positive signal at the ReLU. Later on, Springenberg et al. (2014) introduced Guided Backpropagation which is similar to Deconvolution, but uses the ReLU at both the forward and backpropagation stages. As the computation of this method is based on gradient, it can be used for both classification and regression problems. We will describe it in detail in section 3.3 on NR-IQA models.



(a) *Generated images illustrate some class appearance model*

(b) *Image-specific saliency maps*

Figure 2.9: *Class model visualization and image-specific saliency maps of a CNN (Simonyan et al., 2013).*

Sundararajan et al. (2017) introduced two main axioms that an explanation should satisfy: *sensitivity* and *implementation invariance*, and states that most of the methods did not meet these requirements. Based on these axioms, a method called Integrated Gradients was proposed. Given a model, this method expects the existence of a baseline that corresponds to a 'neutral' output. For a specific

input, the method aggregates the gradient along the inputs that fail on a trajectory between the baseline and the given input. Although integrated gradients are simple in computation, the main challenge of this method is to select a good baseline (or root point). For example, with an input image of the IQA task, there is no exact definition of a baseline image. Thus, it is not applicable on our work.

SmoothGrad Smilkov et al. (2017) used a similar approach which uses gradients to produce an explanation. The main idea of SmoothGrad is to add noise to the given image, then take the average of the result of the saliency maps for each variant of the image. This method creates a smoother explanation and addresses the issue of scatter gradient in deep neural network (Samek et al., 2021). DeepLift Shrikumar et al. (2017) backpropagated the contribution to every feature of the input. The measurement of importance is based on the difference from a *reference* states. It compares the activation of each neuron to its 'reference activation' and associates neurons with contribution scores according to the difference. This method is different from most gradient-based methods, in a way that the discontinuities of the gradients are avoided, thus the importance of the signal is maintained from the target layer to the input space.

Grad-CAM (Selvaraju et al., 2017) is one of the most popular explanation methods for the classification deep neural networks. GradCAM is derived from Classification Activation Mapping (CAM) (Zhou et al., 2016), an visualization of features that discriminate an image to a target class. Given an input image and a target class, the explanation is created by computing the weighted sum of activation maps at the last convolutional layer which is located right before the final softmax layer. CAM has two particular drawbacks. Firstly, it can only be applied on the CNN networks that have a specific structure in the last layers and do not contain any fully connected layer. Secondly, it is not possible to visualize the layers before the last convolutional layer. GradCAM (Gradient-weighted CAM) generalizes CAM to apply to wider variants of CNN models. The authors analyzed the failure of the model by visualizing the failed examples, creating the adversarial images, and identifying the bias in the dataset. They also combine GradCAM with other visualization methods to provide high-resolution results. Guided Grad-CAM, the multiplication of the mentioned guided backpropagation and Grad-CAM, outperformed other explanation methods on both interpretability and faithfulness to the model. Figure 2.10 shows examples of using various techniques for finding relevant features supporting different classification predictions. While Guided Backpropagation results in fine-grained maps, Grad-CAM highlights an coarse area in the image that contributes toward a specific class (cat or dog).

Grad-CAM++ (Chattopadhay et al., 2018) is an extension of Grad-CAM that provides a better visual explanation of CNN models. Instead of using gradient in the computation of the feature's importance as in Grad-CAM, Grad-CAM++ uses

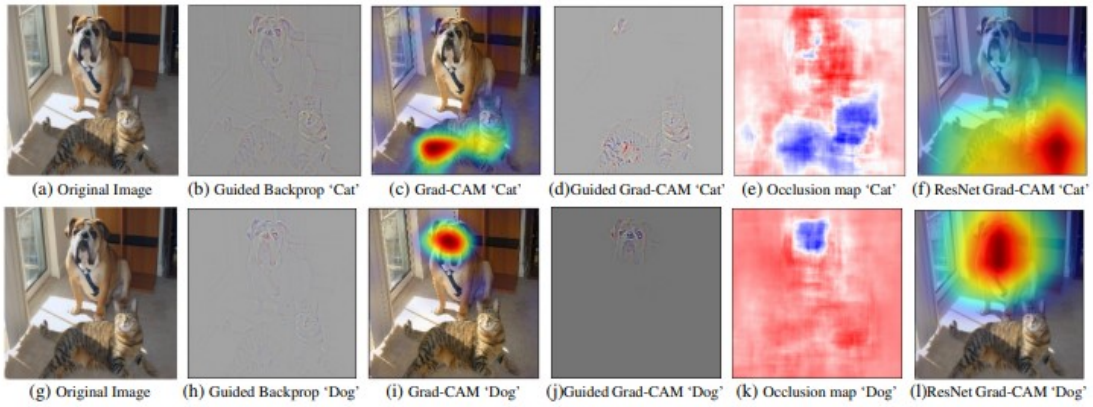


Figure 2.10: *Original images (a, g) and the supported evidence for the cat category by different visualization techniques for VGG16 (b-e) and for ResNet (f); Support for the dog category (h-l) Selvaraju et al. (2017).*

the positive partial derivatives of the last convolutional layers activation maps to generate the visual explanation for the target class. This method produces the saliency map for all the instances of the object in case there are multiple of them in an image. Therefore, the explanation from Grad-CAM++ gained more faithfulness to the classification model, in comparison with the result from Grad-CAM. For the NR-IQA models, as the number of objects in an image are not crucial to the quality of an image, Grad-CAMM++ show the similar result as Grad-CAM.

Layer-wise Relevance Propagation (LRP) Bach et al. (2015) introduced the explanation at the pixel level of the input image to the corresponding output. The method starts at the output of the network and propagates backward, layer by layer, until reaching the input. The main principle in LRP is the propagation rule, which aims to find the relevance value - the importance of each neuron in the models to the prediction output. From the last layers back to the input variable, the neuron in a layer redistributes the information that it received from the later layers to the neurons in the lower layer. In other words, the relevance value of a neuron in a deeper layer is equal to the summation of the connection-relevance values of the neurons in the previous consecutive layer. Subsequently, LRP satisfies the conservation property for all layers. Montavon et al. (2018) pointed out that the explanation produced by LRP is usually noisy, which could confuse human observers. While the original LRP propagation rule could only apply to certain types of neural network layers, Kohlbrenner et al. (2020) mentioned about using the combination of the LRP-rule can provide better results, and they quantify the fidelity of different methods by using a measurement based on the bounding box of object location.

Kindermans et al. (2017) mentioned that the previous methods such as De-

Convnet, Guided BackProp, and LRP work based on the assumption that it is possible to propagate the output signal back through all the layers until the input and get something that shows how the relevant signal was encoded and explained by the networks. Moreover, the theoretical analysis and the quantitative evaluation of the methods are lacking. The experiment that was implemented in this work shows that these methods are not able to distinguish *signal* from *distractor*, thus propagating the sub-optimal explanation of how the deep networks work. Based on the analysis which treats the input data as a combination of the signal (relevant information) and the distractor, the two explanation techniques PatternNet and PatternAttribution were proposed to provide an improved explanation for the deep network. A measurement was introduced to quantify the quality of the explanation methods and prove that the assumption used in DeConvNet and Guided BackProp is not right (the weights do not correspond to the detected stimuli).

Ribeiro et al. (2016) proposed LIME (Local Interpretable Model-Agnostic Explanation) that can provide the local explanation of a model by a simpler surrogate model. Firstly, a local distribution around the interested data point (for example, an image or a single scalar value) is defined. Then, the method finds the importance of each feature in data representation by minimizing the loss between the original output of the model and the output from the simplified model over the local distribution. As the name indicates, LIME can be applied to any black-box model, unlike the above methods, which are mostly applicable to deep-learning models.

Lundberg and Lee (2017) introduced SHapley Addictive exPlanation (SHAP), which is based on game theory to compute the explanation of model prediction. The authors defined a group of explanation methods which is called additive feature attribution methods. They are the techniques in which each feature in data representation is assigned a single importance value, and the sum of the important values approximately matches the prediction output of the original model. LIME, DeepLIFT and LRP are some additive feature attribution methods. SHAP is a unified framework for this group of explanation methods that can provide a unique solution for a prediction from a model. SHAP determines the importance of each feature for a particular output by the difference between the original output and the predictions when a feature is absent. As this method require multiple permutations of features in implementation, it is computation-consuming. Especially, if the input of the model is an image, the number of combination features, or pixels, to be hidden is numerous. Thus, it requires a strong computational resource to implement this XAI method on image data.

For computer vision tasks of images as input data, there is a group of explanation methods that share the same idea: computing the significance of image features by how much their perturbation changes the model's prediction. The using of local

data distribution in LIME is similar to this approach. In earlier work, Zeiler and Fergus (2014) occluded different regions in images to analyze which parts play the most important role in classification prediction. This method is referred to as occlusion analysis, in which the absence of an image region is represented by gray patches of the same shape. Fong and Vedaldi (2017) proposed meaningful perturbation for explaining classification models. Instead of using a grey box to occlude parts of the images, the authors suggested to simulates the interested region by naturalistic effects. The three types of perturbation were proposed: adding noise, replacing with a constant value, and blurring.

Zintgraf et al. (2017) presented the prediction difference analysis method for visualizing which pixels or regions in an image devote or again the prediction of the model. Based on two observations: a pixel mostly depends on small pixels around it, and the conditional probability of a pixel in its neighborhood regions does not depend on the position of the pixel in the image; the authors proposed a conditional sampling for perturbations of image features (pixels).

All the above methods provide saliency maps for data representation of a specific input image corresponding to a prediction of a classification model. The data representation could be the input image (in case of perturbation-based approaches), or a representation from intermediate layers of the model.

2.3.2 Explainable AI for regression models

Image quality assessment can be considered as a subset of regression problems because the output of IQA models are normally a continuous scalar. While explainable AI has been widely studied for image classification/recognition tasks, interpretation for regression models only got little attention. As a result, there are limited works that explored the interpretability of IQA metrics.

Letzgus et al. (2022) pointed out the main challenge in explaining the regression model: focus on the attribution of each feature at a specific sample (data point). The explanation should take into account the unit of measurement of each feature in the input samples. Moreover, it should be contextually sufficient, not only by being able to explain the reason for the output of each data point but also the relevant region around each specific point. For XAIR, the removal-based and gradient-based methods may require a reference data point for producing the explanation. On the other hand, the propagation-based methods do not require the reference point; however, this type of approach assumes that the model is disentangled. The paper proposed two approaches: retraining and reconstructing the models. In the former, a surrogate network is retrained to have the same accuracy as the target model. The training data for this approach need to be chosen carefully to avoid bias in explanation. The latter rewrites the last layers in a way that the representation

is alternated and has a fine-grained explanation. It assumes that the last two layers of the network are ReLU and linear. The reconstructing includes two steps: propagation in the linear layer and propagation in the ReLU layer. Because of this assumption, this approach is applicable to a limited set of models that have a specific architecture.

Tamaddon-Jahromi et al. (2020) and Papadopoulos and Kontokosta (2019) used Shapley values to get an understanding of a heat transfer model (DNN) and a building consumption computation (XGBoost model). Kratzert et al. (2019) applied Integrated Gradient to determine the importance of each natural factor to the rainfall-runoff which is modeled by a LSTM network. In these studies, the authors use the common XAI techniques of the classification task to interpret the model of regression. However, the input data of these models are in a table format, which is more flexible than image data.

By far, the most common approach for applying XAI to regression is to use the XAI methods that were originally developed for classification models directly on regression models. Another way is to approximate the regression task to a classification problem by clustering the regression output into some classes and applying the XAI techniques to get an understanding of the model.

In terms of the image quality assessment, the only prior work that aims to provide an explanation for any IQA model is from Prabhushankar et al. (2020). The authors proposed a contrastive explanation scheme that provides the answer to the question *"Why the quality score of an image is P, rather than Q?*. Consider a regression network $f()$, trained to predict a continuous output y. During the training, an empirical loss $L(y, y', \theta)$ is minimized with y is the predicted output, y' is the ground truth output, and θ is the network parameters. The process of minimizing the loss $L()$ is conducted by backpropagation using the gradients $\frac{\partial L}{\partial \theta}$. The authors define the contrast explanation as the difference between two predictions of the neural network. Each prediction belongs to a manifold that is spanned by the weight W_1 and W_2 of the neural network in the output space. An example is shown in Figure 2.11 in which the learned manifold is in blue, and the contrastive manifold is in purple. The difference between the learned manifold (of the predicted score P) and the contrastive manifold (of the contrastive score Q) is measured by using gradients backpropagating the loss between P and Q. With the loss function $L()$, the contrast is computed by $\frac{\partial L(P,Q,\theta)}{\partial \theta}$. In the paper, the authors choose mean square error for the regression networks and integrate with Grad-CAM to show the contrastive explanation. Firstly, the contrastive gradient $\frac{\partial L(P,Q,\theta)}{\partial \theta}$ is backpropagated to the last convolutional layer of the model, and K gradient maps are obtained. They are average pooled and then resized to the size of the input images as a contrastive map. The heatmap is overlaid on top of the

original image and shown.

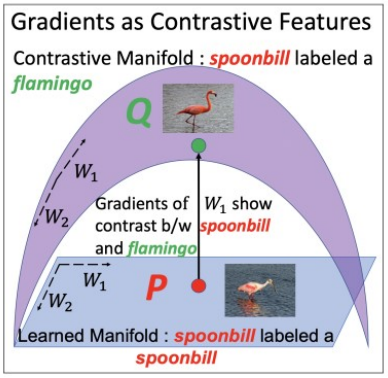
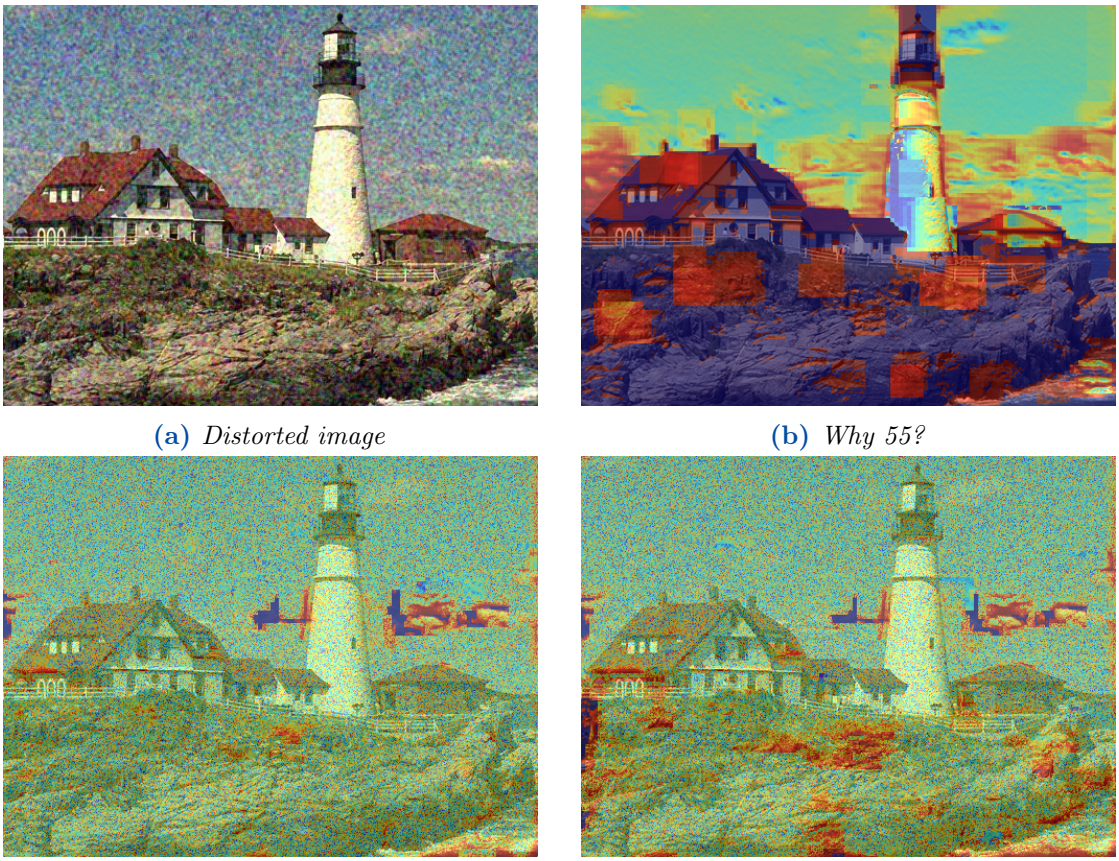


Figure 2.11: *Predicted manifold and contrastive manifold. Prabhushankar et al. (2020).*

Figure 2.12 shows the explanation for the predicted quality score 55 of an image and the contrastive explanations on why the NR-IQA model does not predict a higher score (75) or a lower score (45). The red pixels in the heatmaps represent the regions in the image that support the prediction, while the dark blue ones are the opposite. From Figure 2.12b, we can see that the model considers the areas around the middle horizontal of the images to estimate a quality score of 55. Comparing the top-down images, the dark blue region in Figure 2.12c, which suggests the quality should be higher than 55, overlaps with the dark blue region in Figure 2.12c, which indicates a lower quality than 55. In other words, the same features in the input images represent two opposite opinions, which is unexpected. This example shows the drawback of this method, in which the explanation results in conflict with each other.

Contrastive explanation brings a new idea to make an IQA model more interpretable by finding the features that can make a neural network change its prediction. However, their visualization results are inconsistent. It could confuse users instead of helping them to better understand the model. Therefore, future works are still needed to bring explainability to IQA neural networks.



(a) *Distorted image* (b) *Why 55?*

(c) *Why 55, rather than 75?* (d) *Why 55, rather than 45?*

Figure 2.12: *An image (a) and the explanations for each question shown below each image. Red pixels in the heatmap represent the regions that support the answer of the corresponding question.*

3 Methodology

In this Chapter, we will describe the general idea of our solution for the research questions of this work. Firstly, we conducted short interviews with students and some experts in the field of IQA to form an overall concept of good explanations for IQA models. The result of this interview is presented in section 3.1, which addresses question Q1. After that, an workflow to provide such explanations is proposed and described in section 3.2 and 3.3.

3.1 Definition of good explanations for IQA

IQA is a vast topic, which is an emerging multidisciplinary field based on social psychology, cognitive science and engineering science, focused on understanding overall human quality requirements. Thus, it is difficult to determine which factors will contribute to a good explanation of an objective IQA model. Therefore, we conducted a short interview with students and experts who have a background in this topic to find their expectations of good explanations. 15 participants were interviewed, among them, there are four experts. Our question was: "Given an objective quality model, which kinds of explanation will make you trust the model?". This question is equivalent to the question Q1 that was mentioned in section 1.3. The collected answers were slightly different between interviewees, and the common expectation are the followings:

- The explanation that can point out the limitations of the current model.
- The explanation that illustrates the features of the image the model focuses on to make the prediction of image quality.
- The explanation that can show how the model mimics the human visual system.
- The explanation provides the knowledge of the model, from the primary to higher layers, which can contribute to a better understanding of the HVS.

Ideally, it would be great to have explanations that can satisfy all four expectations. However, as it is difficult to achieve such ambition, in this work, we try to provide explanations that meet at least one of the above conditions. To address the first opinion - finding the drawback of the model, we aim to detect the images that the model fails to predict their subjective quality. Then, the explanation which corresponds to the second and third conditions is provided for selected predictions. The general process of the proposed workflow is shown in Figure 3.1. Given a pre-trained image quality model, our goal is to provide an explanation of how the model makes a prediction for the quality of images. Unlike other computer vision task such as image classification, in which the output of the model can be evaluated as right or wrong objectively; a prediction in the image quality assessment need to be evaluated in the relationship with the subjective ratings. Therefore, a set of images with the available human opinion score are necessary to check the model performance. Given a pre-trained IQA model and a subjective dataset. Firstly, we will find the outlier images, which are the images that the model fails to accurately estimate quality. Different outlier detection methods will be described in section 3.2. After this step, we will have a set of images, which represent the failure prediction of the model, while the other images in the input IQA subjective set are considered good predictions. By analyzing the sets of failure decisions of the model

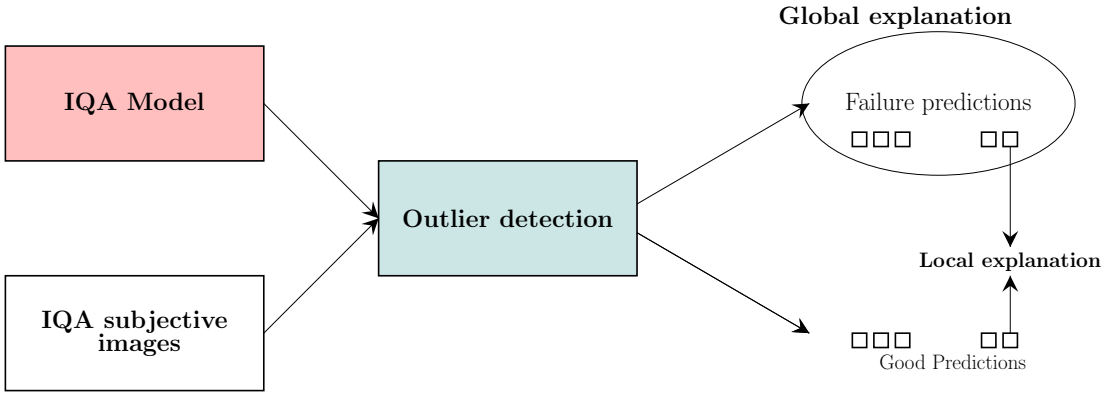


Figure 3.1: *The general workflow of our method to provide an explanation of an IQA model.*

on different IQA databases, we can form a global explanation of the model, which corresponds to the first point: limitation of the model. After that, an investigation into which features of an image contribute more to a prediction of the model is conducted. This explanation is called visualization explanation because it will provide the importance of corresponding to image feature on an interpretable map. The local explanation by visualization will be described in section 3.3. The visual explanation of the spatial domain provides the explanation of the second condition,

while the third condition is explained on the frequency and color domain.

3.2 Explanations through outliers

This section will address the issue of lacking objective methods to detect failure cases of the IQA models. In the literature, when evaluating the performance of an IQA model, the authors hardly showed the images that the estimated quality from the model does not match the subjective quality score. Even if this information is provided, these images are selected subjectively. To the best of our knowledge, there is no prior work that proposes objective methods to find these failure cases of IQA model. We believe that the set of these images can form a global explanation of the limitations of the model. In other words, the inaccurate predictions would provide more understanding of the model. In image quality assessment, a prediction is considered inaccurate if the subjective judgments and the estimated quality scores are poorly matched. We call the images that correspond to this type of prediction *outliers*.

In this work, we investigate different ways to detect outlier images given an image quality assessment database and an IQA method. It should be noted that these methods can be used for both traditional and deep learning based NR-IQA models.

3.2.1 Outlier detection based on correlation coefficient

In the literature, the Spearman correlation is commonly used for comparison between an IQA metric score and the perceived quality judgments. The coefficient value of this type of correlation indicates the degree of agreement between the estimations of the IQA model and the subjective quality assessment. The larger the magnitude of the coefficient is, the better the IQA model is. Meanwhile, the outlier images are the ones that represent the failure predictions of the metrics. Therefore, the appearance of outliers in a set of images will lead to a lower correlation coefficient. From this observation, we propose a method to detect outliers based on the change in the correlation coefficient. The outlier detection method consists of the following steps:

Input: an IQA model, a set of images I with the corresponding subjective quality rating MOS

1. For all image $I(i)$ in the image set, get the estimation score of the IQA model:

$$Prediction(i) = M(I(i)) \tag{3.1}$$

2. Calculate the overall correlation coefficient between the subjective scores and the predicted scores for the data set:

$$r_{all} = Correlation(MOS, Prediction) \tag{3.2}$$

3. Calculate the correlation coefficient between the subjective scores and the predicted scores when each pair of MOS and prediction score for one image is excluded:

$$r_{exl}(i) = Correlation(MOS_{/i}, Prediction_{/i}) \tag{3.3}$$

4. Calculate the change of correlation between the full data set and when each image is removed:

$$\Delta r(i) = r_{all} - r_{exl}(i) \tag{3.4}$$

5. Select the images with the biggest Δr as the outliers.

While this method can not classify the outliers automatically, it gives the freedom to define the number of outliers on the user side. Therefore, we carefully use this approach along with the visualization plot of data distribution for better reliability.

3.2.2 Outlier detection with RANSAC

RANdom SAmple Consensus (RANSAC) is an algorithm proposed by Fischler and Bolles (1981). This is an iterative method to estimate the parameters of a mathematical model that works with input data containing outliers. While other robust estimation methods such as least-mean squares initially came from statistic literature and then adopted to computer vision, RANSAC was designed for computer vision problems.

The RANSAC algorithm learns the parameter of the model by randomly resampling the observed data. While other popular techniques of model estimation use as much data as possible to find the solution of the model's parameters, RANSAC uses only a small set of data and then processes with large data. This principle of the algorithm is based on the assumption that the outliers do not contribute consistently to finding the optimal parameters of the model.

The RANSAC algorithm is implemented as follows:

1. Select a randomly minimum number of data points from the original set of data.

2. A model is fitted to the set of selected data points.

3. Test the model again with all the original data points. If the data point fits the model with a predefined threshold ϵ, it is called *consensus* or inlier point.

4. If the consensus set contains a reasonable number of data points, re-estimate the model using the inlier set. This is the solution for model estimation.

5. If the number of data points in the consensus set is not sufficient, repeat from step 1 to step 4 (maximum of N times).

The number of iterations, N is usually chosen high to ensure that at least one of the random sets does not contain outlier data.

We can apply RANSAC to find the linear model that best represents the relationship between the predicted scores from an IQA method and the subjective ratings for a set of images. When the relationship model is determined, the outlier images are also identified.

3.2.3 Outlier detection with logistic mapping and standard deviation of MOS

Many subjective IQA databases provide the standard deviation score along with the MOS for each image. This standard deviation score represents the uncertainty of observer while judging the quality of an image. This measurement was used in conventional studies (Krasula and Le Callet, 2018) of image quality assessment to classify if a data point is an outlier or not. According to Bull and Zhang (2021), if the MOS and the objective metric score have the same range of minimum and maximum values, an image whose subjective and predicted score are indicated by MOS_i and $Prediction_i$, respectively, is an outlier if:

$$|MOS_i - Prediction_i| > 2\sigma_i \tag{3.5}$$

In equation 3.5 σ_i represent the standard deviation of the i^{th} image. The outlier detection method in this section is proposed based on this knowledge. It should be mentioned that for many IQA databases, the number of observers who participated in rating the images is large. Thus, the standard deviation for each image is small, which leads to the classification of too many images as outliers. In our method, we add one more condition of the data point location to compensate for this issue.

Because the range of value in the MOS and the objective score can be different, and the correlation between them happens to be non-linear, a logistic mapping can be used to convert them into the same range.

1. Find a mathematical model (non-linear) $f_{map}()$ that represents the mapping from the objective scores $Prediction$ to the MOS. This model needs to be monotonic. so that the relative position of the data points is preserved.

2. Calculate the mapped scores

$$Prediction' = f_{map}(Prediction) \tag{3.6}$$

3. Calculate the distance d_i from the data point of each image to the mapping line.

4. If the mapped score $Prediction'_i$ of an image satisfies the condition in equation 3.5 and the distance from the data point is in top $p\%$ of all the distance, the image is classified as an outlier.

The set of detected outlier images can be considered as criticism examples, which fail the IQA methods. By analyzing the criticisms, we can have some insight into the IQA models.

3.3 Explanation through visualization

Visualization explanation shows the importance of features in an image for a model prediction. This is the most user-friendly approach to provide the model's understanding as we can see which part of the image the IQA model focuses on. In this section, we will describe the visualization methods that were used in this work.

3.3.1 Perturbation-based methods

Perturbation-based techniques aim to explain a black box by modifying the input of the model. If the input data is a text, words in that text could be replaced by other words; in the case of image data, pixels' values are perturbed. This group of understanding deep learning models assumes that the change in the model's output indicates which parts of input data are crucial in decision-making. The importance of the input features is measured by the difference between output prediction when the element is present and when it is absent. If an input feature is removed or replaced and the model output changes significantly, the corresponding feature is assigned a high value of importance. For image classification problems, a significant change in the models' output can be a different class of object from the original prediction of the unperturbed image. For image quality assessment problems, all changes in the quality estimation of the images are recorded.

The perturbation-based methods do not need access to the deep learning model, or modify any part of the models. Therefore, they are able to be applied to deep models, but also to any kind of black-box model. The traditional approach of perturbation-based methods is implemented on the spatial domain of the input

image. In other words, different regions in the images will be sequentially replaced by a new set of pixels of the same size as the original region to create a modified version of the input image.

3.3.1.1 Spatial domain

Following a process that was introduced by Zeiler and Fergus (2014) for the classic image classification problem, we propose a workflow of generating an importance map for an image in objective quality assessment with some modification. Figure 3.2 demonstrates the steps of our workflow. Given an image x_0 and a pre-trained IQA model f, the estimated quality of an image predicted by the IQA model is denoted as the Original score, $f(x_0)$. A perturbation rule is applied to the original image to obtain a set of perturbed images. These generated images are different from the original image at specific locations while keeping the pixels in other regions. Therefore, they can be considered as the neighborhood X_0 of the original input image. For each image x_{0i} in the perturbed set X_0, the quality score estimated by the IQA method is denoted as $f(x_{0i})$. The difference between the original score and the perturbed score $\Delta f(x_{0i}) = f(x_0) - f(x_{oi})$ indicates the importance of the corresponding location of pixel change. If the IQA model was designed in a way that the output estimated score $f(x)$ represents the better quality, the higher $absolute(\Delta f(x_{0i}))$ suggests more attribution of the modified region and vice versa. If $\Delta f(x_{0i})$ is greater than 0, the perturbation region contributes toward the prediction of higher quality. On the other hand, if $\Delta f(x_{0i})$ is negative, the corresponding location contributes toward the prediction of poorer quality. After calculating the importance values for all the regions in the image, we obtain a saliency map that represents the attribution distribution of spatial features of the original image to the objective quality score. This attribution map could provide the local explanation of the IQA model f at a single instance x_0.

It can be seen that the concept of image perturbation to identify the saliency map is simple and easy to understand. The main problem with this idea is how to generate the perturbed images, or how to absent different regions in the original image. We aim to find methods for simulating the absence of a set of pixels without creating many artifacts on the generated image. Thus, we consider four scenarios of image generation: 1) replacing the region in the image with a black patch, 2) replacing the region with a patch of the mean value of the original region, 3) replacing with a patch of the median value of the original patch and 4) blurring the patch region.

Figure 3.3 illustrates four types of perturbation. Given a patch (Figure 3.3a) from the original input image, in the first three perturbation types, the patch is replaced by a set of pixels of the same size and constant values. First, the new patch has all the pixel intensities equal to 0. This approach is taken from (Zeiler and

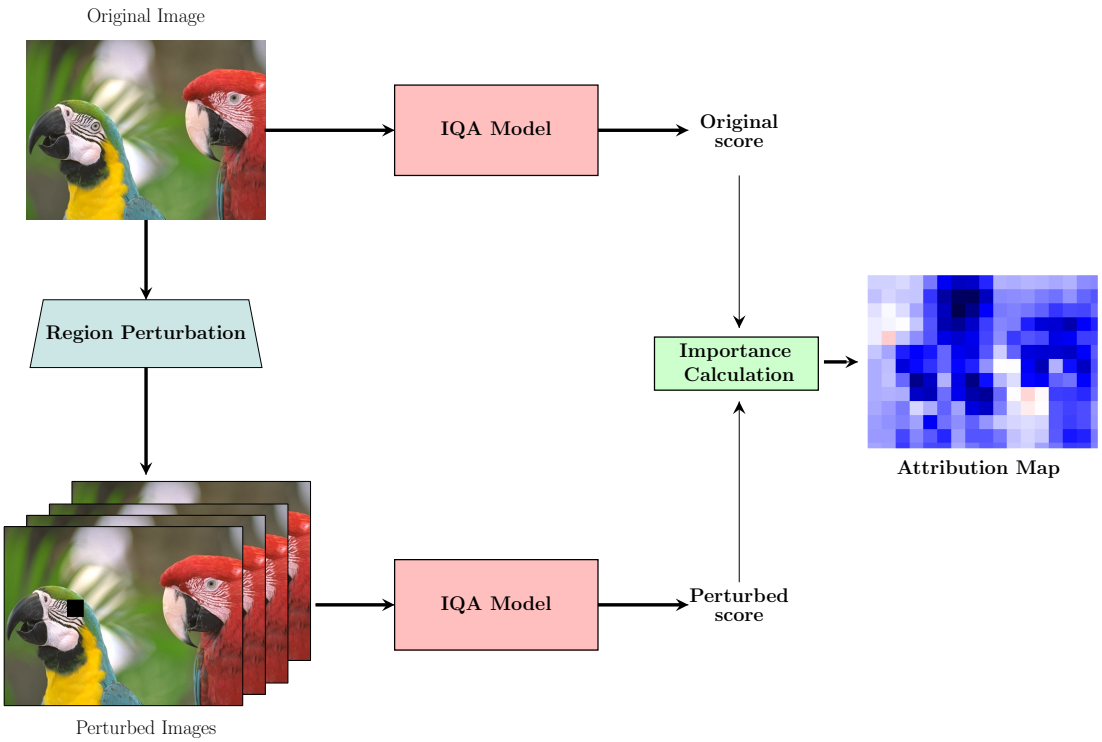
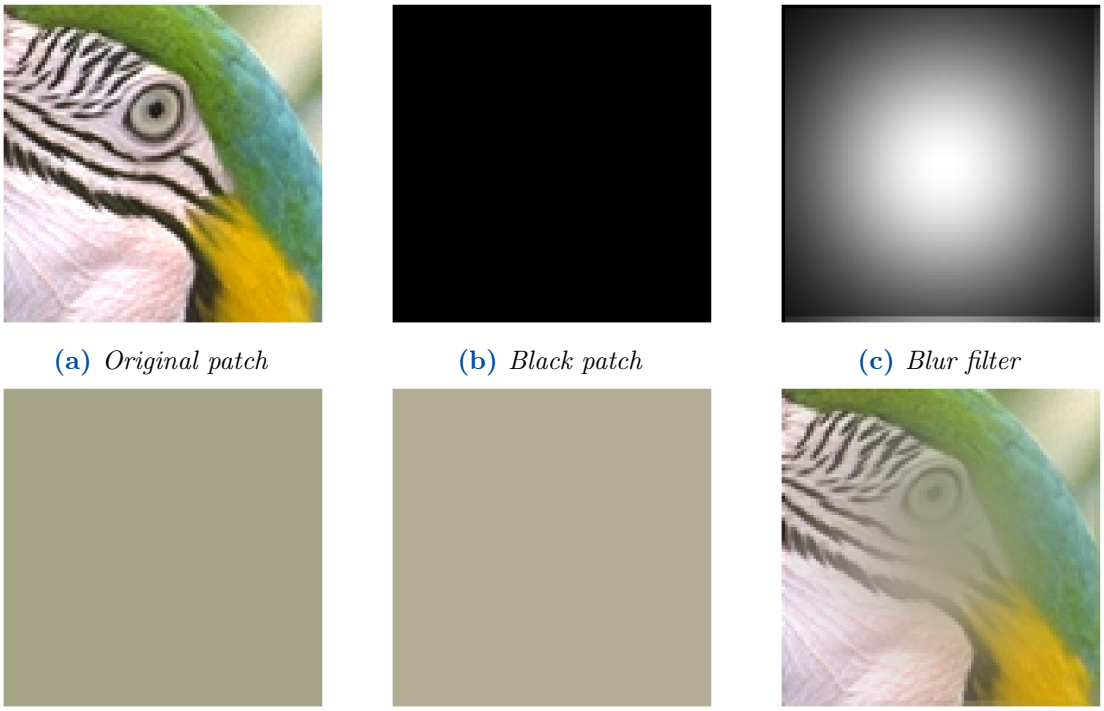


Figure 3.2: *Perturbation-based approach on the spatial domain. The Importance Calculation can be implemented by a simple subtraction.*

Fergus, 2014). As for any region in the image regardless of its location or information, the alternative patches are the same, which is a black area (Figure 3.3b); this type of perturbation *delete* all the features of the original input. Therefore, it is also called *occlusion* technique. However, when putting the black patch at the location of the original patches, it could lead to perceptible artifacts 3.4a. Perturbations by replacing with mean and median values of the target pixels' locations, as their name indicates, generate the alternative patches of constant intensity, which equal to mean and median of all pixels in the origin, respectively. The synthesized patches corresponding to the two techniques are represented in Figure 3.3d and Figure 3.3e. The two perturbed images are Figure 3.4b and Figure 3.4c. As we expected, they look to contain fewer artifacts than the black patch case, but at the border of the replaced patch, there is still discontinuation of the image pattern. These distortions could strongly affect the quality score that the IQA model predicts, which leads to an inaccurate attribution map. For minimizing the influence of this issue, the fourth perturbation type is considered: blurring the patches. We use the 2D Gaussian kernel as illustrated in Figure 3.3c and produce the alternative patch as in Figure 3.4d. The corresponding perturbation image is shown in Figure



(a) *Original patch* **(b)** *Black patch* **(c)** *Blur filter*

(d) *Mean patch* **(e)** *Median patch* **(f)** *Blurred patch*

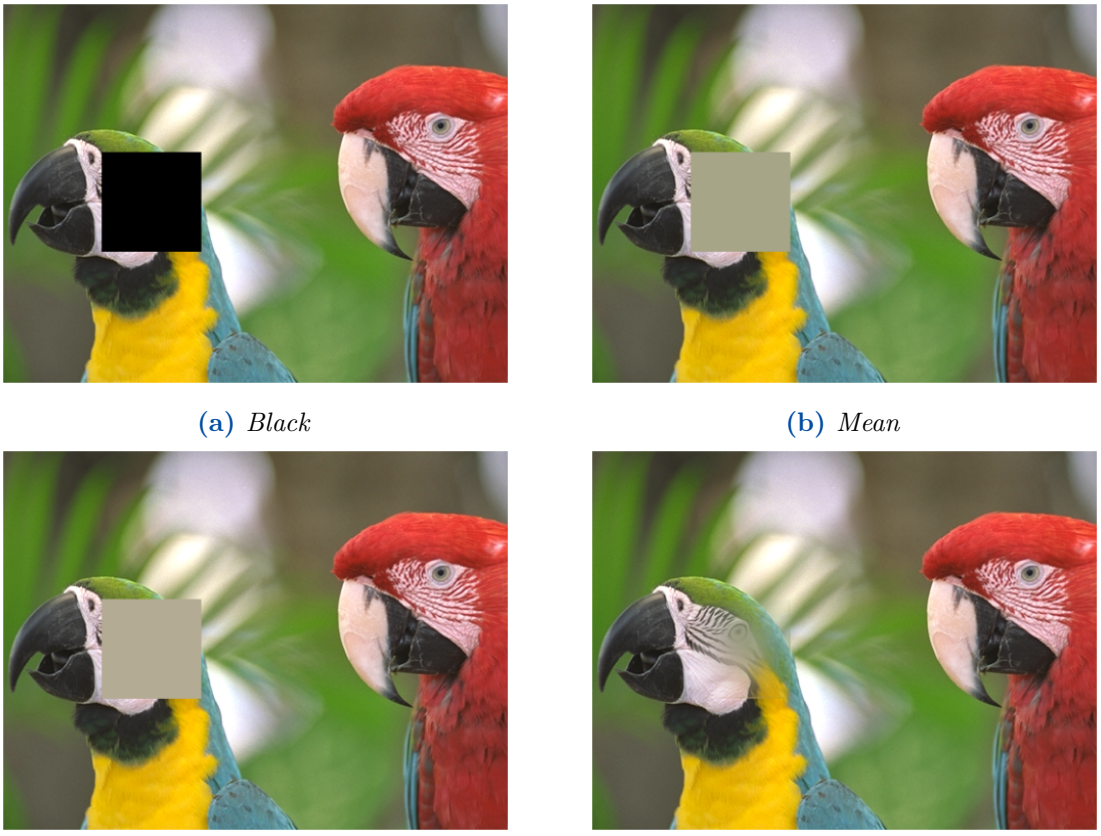
Figure 3.3: *Original patch (a) and difference types of perturbation (b,d,e,f).*

3.4d. We can see that the image patterns at the border of the replaced patches are smoother, which makes the images more naturalistic.

Further investigation about the effect of patch sizes and patch stride will be described in Section 4.3.

3.3.1.2 Frequency domain

While the mentioned process performs on the spatial domain of the image, we decided to extend the workflow to other domains of the image. This is a new contribution to the field as no prior studies have investigated the contribution of image features in other domains than the spatial. Because the majority of introduced IQA methods only accept *normal* images, which are in the spatial domain, as input, two additional steps of domain conversion are added as illustrated in Figure 3.5. The purple block indicates the transformation from the spatial space to the target space, while the other converts image information back to the spatial domain. The first problem we encounter is identifying which target information domains should be considered to bring meaningful explanations of the IQA methods. Based on the observation that the IQA models were designed to mimic how human visual



(a) *Black* **(b)** *Mean*

(c) *Median* **(d)** *Blur*

Figure 3.4: *Perturbed images generated using four perturbation types. The Importance Calculation can be implemented by a simple subtraction.*

system's perceptive quality of an image and the literature of HVS on different information domains (section 2.1), we select to investigate the frequency domain and color space (HSV). The next issue is how to perturb image features in these new domains. For each chosen space, different ways of hiding information were implemented.

Given an input image (Figure 3.6a), the information in the frequency domain is obtained by using the Discrete Cosine Transform (DCT) along each dimension of the image. Thus, the Domain Conversion block in the diagram (Figure 3.5) is the DCT in this case. After this step, the image information is represented in the form of matrices of coefficients (for a color image, there will be three matrices corresponding to three channels of the input image). A visualization of a transformed image is illustrated in Figure 3.6b, in which the brighter intensity indicates the greater value of the coefficient. We can notice that the upper left corner is brighter, which means

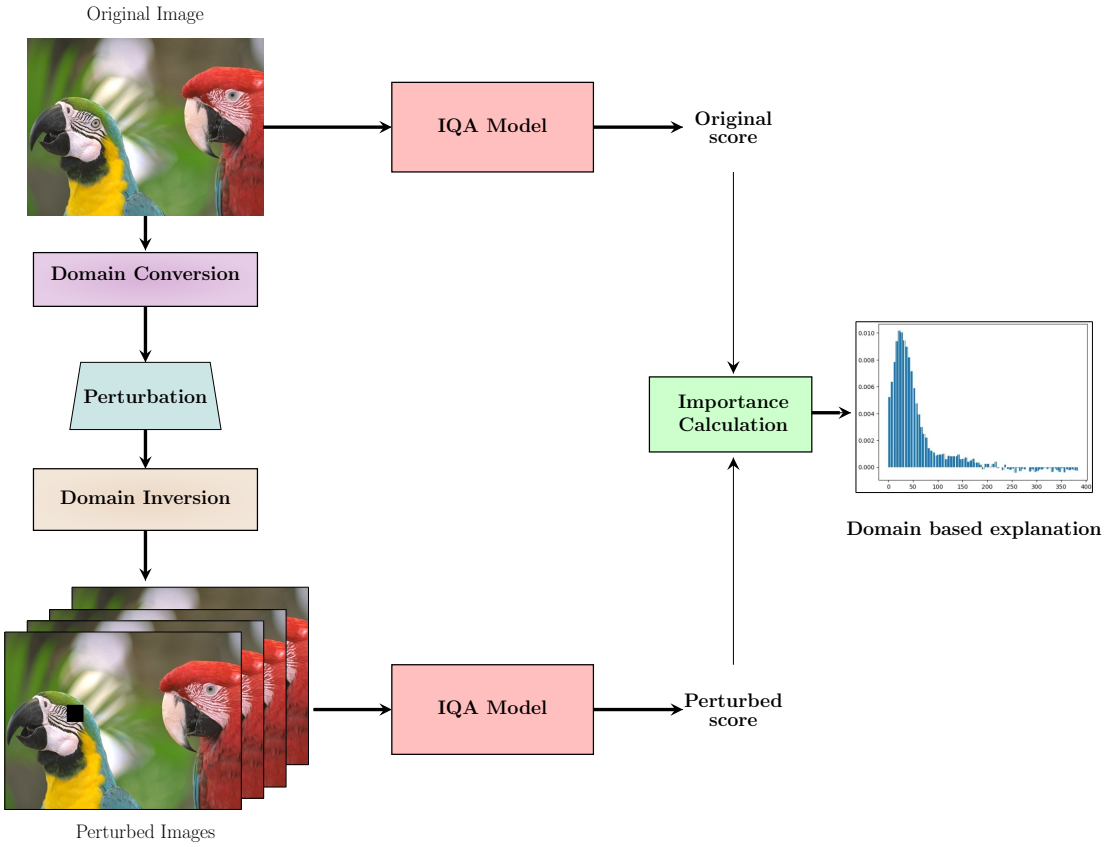


Figure 3.5: *Perturbation-based approach on other image domains.*

that the coefficients in this area are bigger than others.

As shown in Figure 3.7, the DCT can separate images into three sub-bands: low frequency, middle frequency, and high frequency. The highest energy of an image is concentrated on the low-frequency sub-band, which contains the most important visual part of the image. Therefore, the visual quality of an image would be degraded significantly if the information in the low-frequency DCT sub-band is distracted, while the image visual quality would not be affected if the distraction happens in the high-frequency sub-band. Many computer vision tasks take this knowledge to embed watermarks or compress the images without causing imperceptible artifacts. We wonder if the IQA models follow this property in their quality estimation process for images. Therefore, we design the perturbation methods to hide the information at each frequency sub-band. At one time, the information at one small sub-band is replaced by setting the magnitude to the value zero, while the other coefficients have remained the same. After that, the modified DCT matrix will be converted back to the spatial domain by using the

(a) (b)

Figure 3.6: *An image (left) and the DCT transformation (right).*

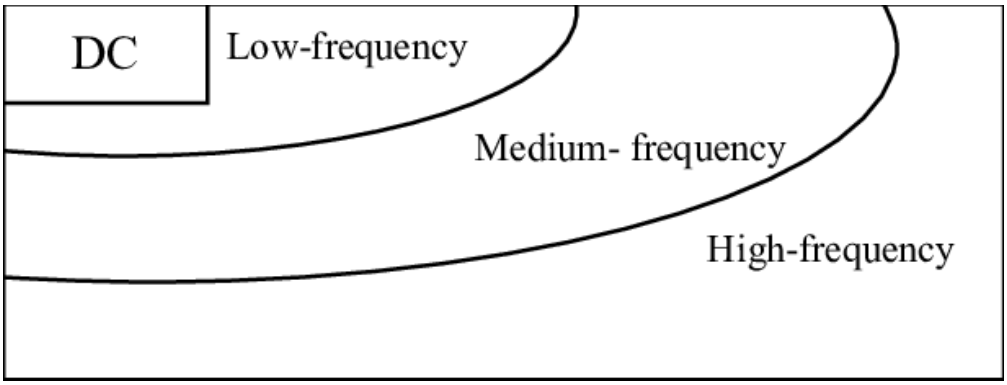


Figure 3.7: *Frequency distribution of DCT coefficients (Images taken from (Madhuri and Bindu, 2015)).*

Inverse Discrete Cosine Transform (IDCT). Thus, the Domain Inversion block is IDCT for the frequency domain. The output of the domain inversion step is perturbed images. These images are put through the IQA models to predict the new quality scores. The next steps are similar as described in the perturbation on the spatial domain: the importance of each frequency band is computed as the difference between the original score and the perturbed scores. Finally, we have the attribution maps which show the importance of frequency bins to the prediction of image quality produced by the IQA methods.

3.3.1.3 Color domain

We also applied the workflow illustrated in Figure 3.5 to the color space HSV. Our initial goal is to find which range of color (hue) will contribute the most to

(a) *Hue channel* **(b)** *Saturation channel* **(c)** *Value channel*

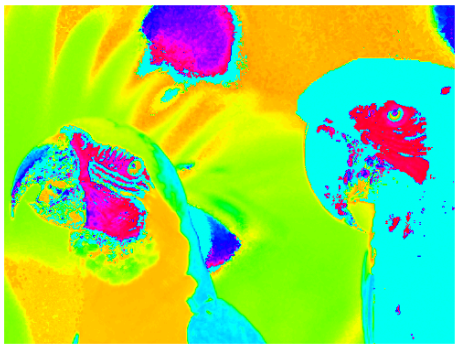
Figure 3.8: *Three components of image 3.6a in HSV color space.*

the objective quality score. The HSV color space has three components: Hue, Saturation, and Value. it describes colors (hue) in terms of shade (saturation) and brightness (value). Thus, the influence of image feature components such as color and brightness on the image quality can be modeled more separative in this color space.

Because normal images are usually stored in RBG format, we need to convert image pixel intensity from RGB to HSV. This is corresponding to the block *Domain Conversion* in Figure 3.5. Similarly, the block *Domain Inversion* represents the transform from HSV back to RGB color space.

Figure 3.8 represents three components of the image in HSV color space. The perturbation for the color channel is a bit tricky because setting the pixel intensity in this component to 0 will make their color change to 0. Therefore, we sequentially replace each pixel of value in a range $[p_{min}, p_{max}]$ by a new value p_{new} such as $p_{new} \notin [p_{min}, p_{max}]$. For example, by replacing all pixels whose values are in the range of [0, 0.1] in channel Hue by 0.5, we get the new hue channel as shown in Figure 3.9a. If other channels of saturation and brightness are unchanged, we will have a new image (Figure 3.9b) of color perturbation. We can see that the red parrot in the original image now is in blue. The new image is put into the IQA model to predict the perturbed quality score, and the difference between the new and the original quality score is recorded. The process is repeated until all the color range in the image is changed once.

Similar procedures are applied for saturation and brightness channels. We collected the score changes of all perturbations to evaluate the modification of which channel will significantly change the objective quality of the image. The result of this experiment is discussed later in section 4.5.



(a) *Perturbation in hue channel.*

(b) *The perturbed image.*

Figure 3.9: *An example of perturbation in color space.*

3.3.2 Grad-CAM and Guided Backpropagation

While the perturbation methods only make changes in input data to find what a deep model is looking for to make a prediction, we can access the network layers and visualize which are embedded in the latent space. In this section, we will describe the two well-known XAI methods: Grad-CAM and Guided Backpropagation, and apply them to explain the NR-IQA models.

3.3.2.1 Grad-CAM

Grad-CAM is proposed by Selvaraju et al. (2017) to visualize the feature maps which are learned by the classification deep learning model. In the classification model, the output is the predicted class of an input image. To get the class-discriminative localization map for the class c, we first need to compute the gradients of classification score y^c for the final convolutional layer feature map A^k, or $\frac{\partial y^c}{\partial A^k}$. This gradient is computed by using the backpropagation method. The importance weight of each feature map to the class score y^c is defined by taking the global average of all pixels of location (i, j) in the feature map:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k} \tag{3.7}$$

Because the gradients change when A^k change, α_k^c indicates how important each feature is to the class c. At each convolutional layer, there are multiple feature maps, they are combined by taking the sum of the importance weights and the maps. A Rectified Linear Unit (ReLU) is used to filter out the parts that have a negative effect on the class decision. The relevant features forwarding the class c are formulated as:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{3.8}$$

Because a ReLU is used in the formulation, the relevant features show parts of the image that positive influence on the class c. In the regression problem, there is only one output of infinite range, using Grad-CAM will result in visualizing the features that increase the output value. Additionally, for finding the part of images that decrease the output value, or the objective quality score, we will use a modification of equation 3.7 for computing the importance weight of each feature map toward this direction:

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j -\frac{\partial y^c}{\partial A^k} \tag{3.9}$$

The authors also suggest applying Grad-CAM in the last convolutional layer in the networks because the output of this layer is the features for the fully connected layers, which are responsible for the regression of the model output value.

3.3.2.2 Guided Backpropagation

Similar to Grad-CAM, Guided Backpropagation Springenberg et al. (2014) is a gradient-based XAI method. The important features of an output value are defined by the gradient of images when backpropagating through ReLU functions.

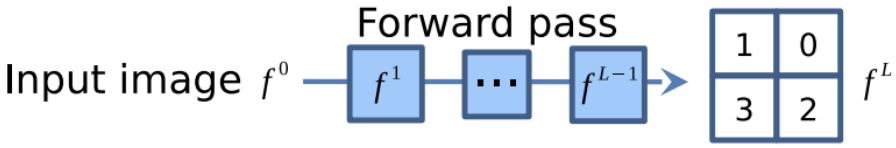


Figure 3.10: *Forward pass in a neural network. (Image taken from (Blog, 2020)).*

In the forward pass of a neural network, let's suppose there are ReLU activations between each two maps as shown in Figure 3.10. We will have:

$$f^{l+1} = ReLU(f^l) = max(f^l, 0) \tag{3.10}$$

The ReLU functions only allow the input values which are not negative. Similarly, at the backward pass, by applying ReLU, only the non-negative gradients are backpropagated to the previous layers. Let R^i denote the reconstructed feature maps at layer i, the reconstructed of the previous layer R^{i-1} can be formulated as:

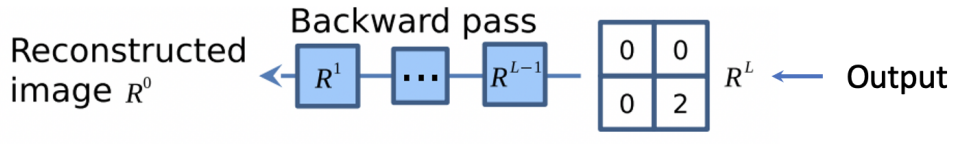


Figure 3.11: *Guided backpropagation. (Image taken from (Blog, 2020)).*

$$R^{i-1} = (R^i > 0).(f^{i-1} > 0).R^i \tag{3.11}$$

At the last layer, R is the gradient of the output value with respect to the learned featured map. using this technique until the input layer is reached, we get the visualization of relevant features in the input image to the output score. From the formulations of this method, we can see that it can be applied not only to the traditional classification problem but also to regression models.

4 Experiments and results

In the current Chapter, we will describe the experiments that we implemented using the proposed methods and evaluate their performance on some no-reference image quality models. The metric choice will be explained in Section 4.1, while the database selection for each experiment is clarified in Sections 4.2.1.

4.1 NR-IQA model selection

Although there are many no-reference image quality assessment methods that have been introduced, due to time limitations and the scope of this work we have defined the following criteria to select the IQMs we will use in our study.

- Popularity: The frequency that the models were mentioned in other studies about IQA.
- Availability: If the source code of the metric is publicly available for reproducing the process of estimating image quality and efficiency comparison.
- self-explainability: If the methods are patch-based or can provide the quality map of the input image, they are the explainable AI model to some extent. Therefore, we select the models which do not follow this approach.

Five deep learning NR-IQA models (Table 4.1) are investigated in our work, including Koncept512 (Hosu et al., 2020), DBCNN (Zhang et al., 2018), CNNIQA (Kang et al., 2014), SPAQ (Fang et al., 2020), and MUSIQ (Ke et al., 2021). Among them, the first four models are Convolutional Neural Networks while the last have transformer-based architectures.

4.2 Outliers detection

In this step, we will compare different methods to find *outlier images*, or the images for which the IQA metric fails to predict the quality score. It should be mentioned

Table 4.1: *The No-Reference Image Quality Assessment models that are investigated in this work*

NR IQA Method	Year	Trained on database	Type of model
CNNIQA	2014	KonIQ-10k	CNN
DBCNN	2020	KonIQ-10k	Bilinear CNN
Koncept512	2020	KonIQ-10k	InceptionResNetv2
SPAQ	2020	SPAQ	ResNet-50
MUSIQ	2021	KonIQ-10k	Transformer

that the absolute predicted score of an image made by a metric does not bring meaningful information, but the relative scores of different images do. This is the reason why in the previous studies that proposed a new image quality metric, correlation measurements such as PLCC and SRCC were used to evaluate the metric's performance. Firstly, the database choice is explained in Section 4.2.1, in which we pointed out a drawback of using the popular FR-IQA datasets in the evaluation of BIQA methods. After that, three different ways of detecting outlier images were implemented as will be described in Section 4.2.2 - 4.2.4. In addition to outlier detection, we also find the best-predicted images, whose quality score estimated by the IQA model well matches the subjective MOS.

4.2.1 Image quality databases selection

Initially, we expected to propose an outlier detection approach that can work for any image quality dataset. In many previous studies ((Zhang et al., 2018), (Kang et al., 2014), (Hosu et al., 2020)), full-reference image quality databases were used to assess the performance of NR-IQA metrics. The efficiency of a metric is defined as proportional to the correlation coefficient between the metric score and subjective judgments. However, as our work focuses on no-reference metrics, which predict the quality of an image without comparison to any reference, we figured out the limitation of using the full-reference image quality database to evaluate NR-IQA methods. While the NR-IQA metrics estimate the quality of an image independently, the subjective scores in the full-reference images quality however show the perceptual quality of an image in comparison with a reference image. In other words, the subjective rating and the objective scores do not measure the same property. For example, three images shown in Figure 4.1 are all associated with the MOS = 5 (highest in subjective opinion) because they are the reference

(a) *Predicted score: 59.80* (b) *Predicted score: 50.40* (c) *Predicted score: 44.86*

Figure 4.1: *Three reference images from (Larson and Chandler, 2010) with the same subjective quality judgment, but are predicted with very different quality scores by an IQA model.*

image in the psychological experiment. But their quality scores which are predicted by the same NR-IQA models are significantly different, for example in the case of using the DBCNN model, the objective score ranges from 44.86 to 59.80. Although the MOS, or DMOS obtained from human participants can be simulated using additional processes, we prefer to use the image quality assessment databases that were built specifically for blind image quality assessments. Only one legacy dataset is used in this part of the experiment. Table 4.2 summarizes the databases that will be used in the outlier detection experiment.

Table 4.2: *Image Quality Databases that were chosen for outlier detection experiment*

Database	Year	Type	MOS/DMOS	Score Range	# Images	# ratings per images
KonIQ-10k	2020	authentic	MOS	0 to 100	10,073	9-15
CLIVE	2014	authentic	MOS	0 to 100	1,162	137-213
SPAQ	2020	authentic	MOS	0 to 100	11,125	137-213
TID2013	2013	synthetic	MOS	0 to 9	3000	23

4.2.2 Outlier detection using correlation coefficient

The detail of this method was described in Section 3.2.1. In this section, we will describe the experiments that were conducted to select the suitable correlation type for outlier detection. We can choose any among Pearson correlation, Spearman rank, or Kendall rank correlation to measure the correlation coefficient between

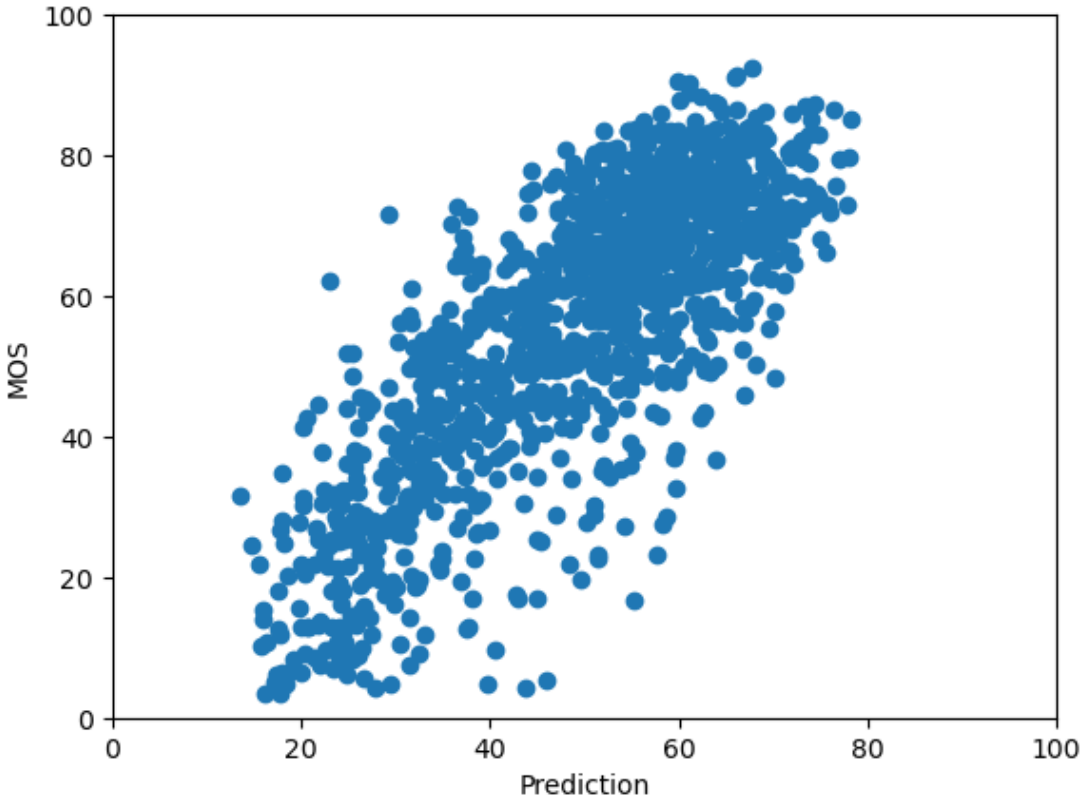
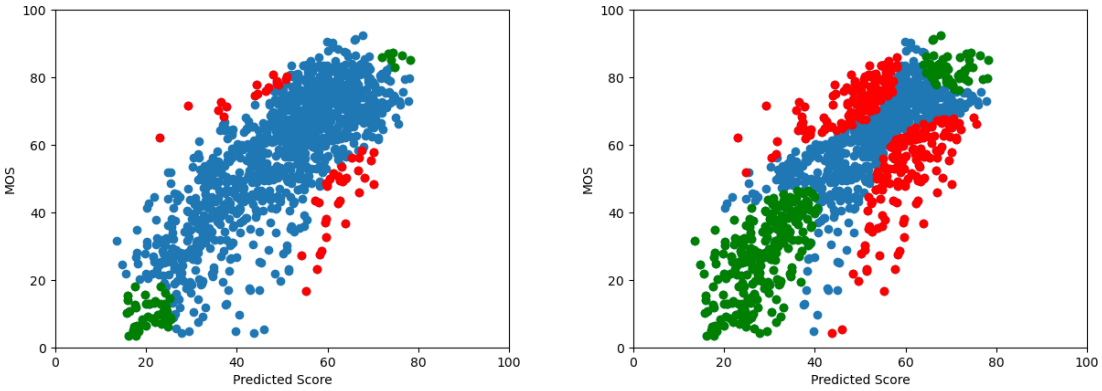


Figure 4.2: *MOS versus predicted scores by the DBCNN model on the CLIVE database.*

the subjective judgments and the objective quality estimation. As the SRCC and KRCC are similar, we only implemented the algorithm with PLCC and SRCC. It is worth noting that the PLCC indicates the accuracy of the metric, while the SRCC measures the monotone association between the subjective and objective quality scores.

An example of the relationship between the subjective and objective score produced by the DBCNN model on the CLIVE dataset is shown in Figure 4.2, in which the horizontal axis represents the estimated score, and the vertical axis represents the ground truth MOS. Each point corresponds to the location of the evaluation of an image by the IQA model and human observer respectively. The range of both subjective and objective quality scores is from 0 to 100, where 0 represents the worst quality and 100 is the best image quality. The outliers which correspond to the overestimated images are expected to locate in the right-bottom corner of the graph, while the underestimated ones are located in the top-left corner.

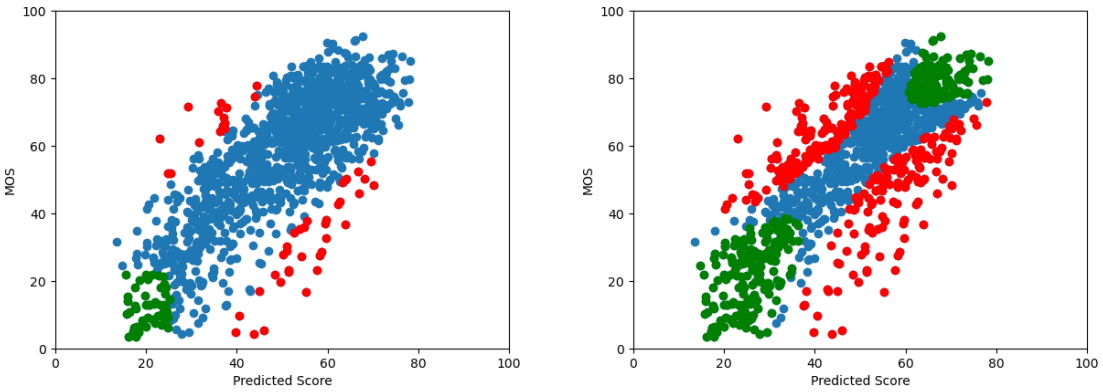
We use the algorithm mentioned in Section 3.2.1 using the Spearman rank correlation coefficients on the set of the MOS and prediction scores in Figure 4.2. Along with the outliers, we also plot the points whose removal will lead to the largest decrease in correlation values. This kind of point is expected to represent the images whose subjective and objective quality score well matches. Figure 4.3 shows the results when 5% and 30% of total images are defined as outliers. The result when using Pearson correlation with the same amount of outlier is represented in Figure 4.4.



(a) *Number of outlier = 5%* **(b)** *Number of outlier = 30%*

Figure 4.3: *Examples of outlier finding based on SRCC with different amounts of outliers (5% in the left and 30% in the right). In the plots, the red dots represent the outlier images, while the* ***green****green green dots indicate the best prediction, the blue dots are other data point.*

Comparing the results shown in Figure 4.3 and Figure 4.4, we can see that in both cases, the best predictions found by the algorithm are located in the tails of the data point cloud. The reason for this distribution is that the PLCC measures the linear relationship between two variables or the closeness of association of the points in a scatter plot to a linear regression line; meanwhile, the points which lie on the tail of the points cloud contribute toward shaping the cloud shape more linearity. Therefore, the removal of these points would lead to the largest declines in correlation coefficients between the two variables (the MOS and the predicted score). It should be noted that the "largest" correlation coefficient change can be very small, at around 0.001 for a dataset of 2000 images, because this value is inversely proportional to the number of images. In the case of the SRCC, as its coefficient is computed based on the difference between the rank of the subjective and objective score of each image (equation 2.3), the data points located in the tails of distribution have the smallest rank difference, dedicate to increasing the monotonic association of the two types of quality.



(a) *Number of outlier = 5%* **(b)** *Number of outlier = 30%*

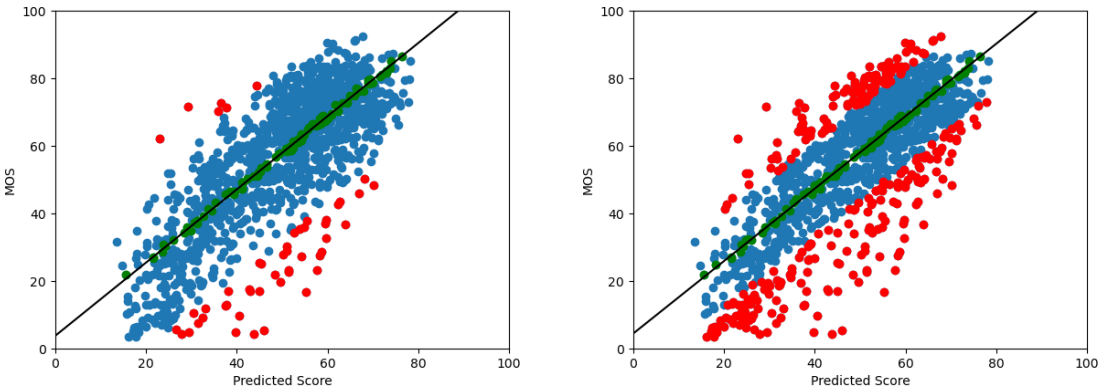
Figure 4.4: *Examples of outlier finding based on PLCC with different amounts of outliers (5% in the left and 30% in the right). In the plots, the red dots represent the outlier images, while the* ***green****green green dots indicate the best prediction, the blue dots are other data point.*

On one hand, the qualitative results of the outliers which are detected by using the SRCC show that this method is not effective for the detection task. When the number of outliers is small, the outliers' locations are appropriate, because they are the ones near the top-left and bottom-right of the graph. However, when the number of outliers is increased from 5% to 30% of total images, the distribution of outliers moves towards the center of the graph. Therefore, the Spearman rank correlation is not suitable for our algorithm. On the other hand, the outlier detection from using the PLCC shows promising results. When the defined amount of outliers increases, the outliers are still the nearest points from the top-left or bottom-right corner of the graph. Therefore, we choose the PLCC for the computation of correlation for this outlier detection method.

4.2.3 Outlier detection using RANSAC

The process of outlier detection by the RANSAC linear model is described in Section 3.2.2. As the number of outliers in this method can change according to the setting residual threshold, we select different numbers of outliers to evaluate the efficiency of the algorithm. The result of the implementation of this algorithm on two IQA datasets, the CLIVE and the TID2013, are shown in Figure 4.5 and Figure 4.6, respectively. The black lines in the graph indicate the regression line that are fitted by the RANSAC algorithm.

We can see that the finding best predictions are the points that locate on the regression line because their distance to the line is equal to 0. The outliers are the



(a) *Amount of outlier = 5%* **(b)** *Amount of outlier = 30%*

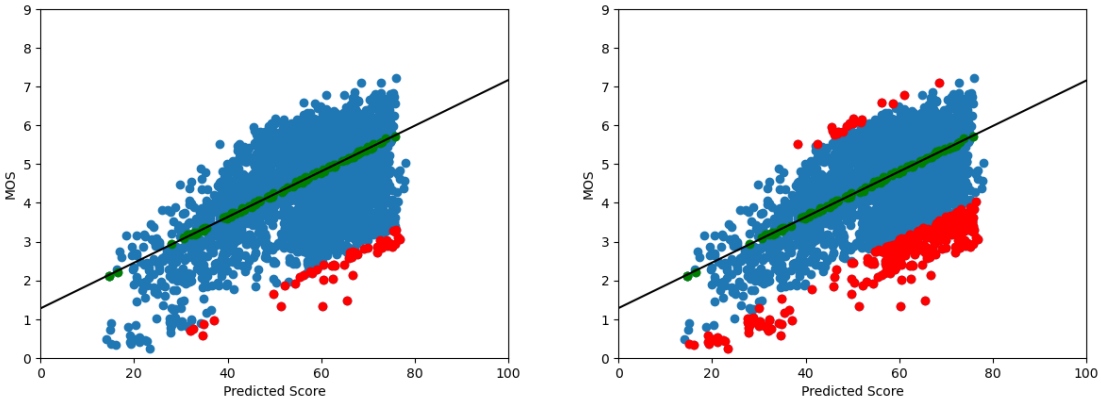
Figure 4.5: *Examples of outlier finding using RANSAC with different amounts of outliers (5% in the left and 30% in the right). In the graphs, red dots represent the outlier images, while green dots indicate the best prediction. The black line represents the linear model fitted by the RANSAC algorithm.*

farthest points from the line. The linear regression line captures the relationship between the subjective and objective quality score well in Figure 4.5. But when the data shows some nonlinear relation between the two variables, the line fails to reflect the relevance of the scores as in Figure 4.6. In both cases, we notice that when the amount of outliers is small, the detection result is more reliable as the outliers are more separated from the remaining data. If the number of outliers is set to a big value, the outlier and the good predictions can be dismissed, as their locations are close together. From our empirical experiments, determining 5% of total images as the outlier will give the most consistent results.

4.2.4 Outlier detection by logistic mapping

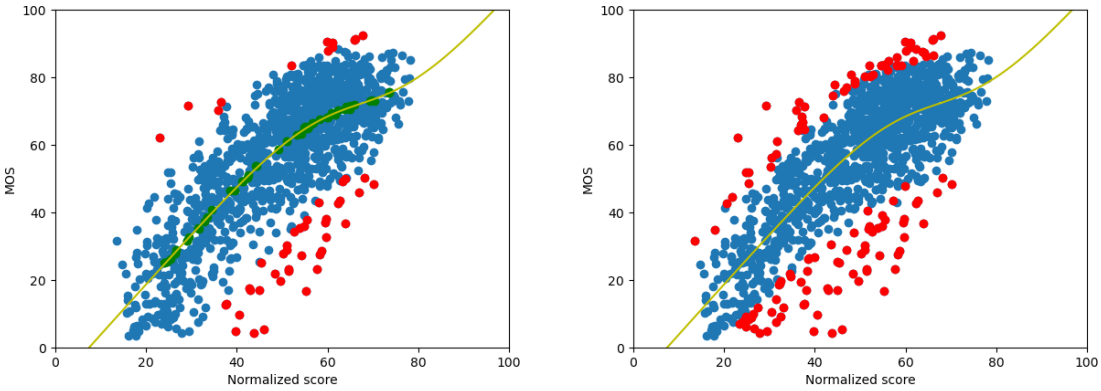
The process of outlier detection by logistic mapping and standard deviation of MOS is described in Section 3.2.3. We follow the same approach in the previous experiment to evaluate the reliability of this outlier detection technique.

The result of the implementation of this algorithm on two IQA datasets, the CLIVE, and the TID2013, are shown in Figure 4.7 and Figure 4.8, respectively. The yellow lines in the graph represent the logistic mapping from the normalized predicted score to the subjective MOS. When the setting amount of outliers is increased from 5% to 30% in Figure 4.7, thanks to the constrain of reliable range of MOS in equation 3.5, a number of data which are far from the non-linear mapping lines but has a small standard deviation of the MOS are excluded from outlier sets. Therefore, the detection result is more consistent with the location of outliers



(a) *Amount of outlier = 5%* (b) *Amount of outlier = 30%*

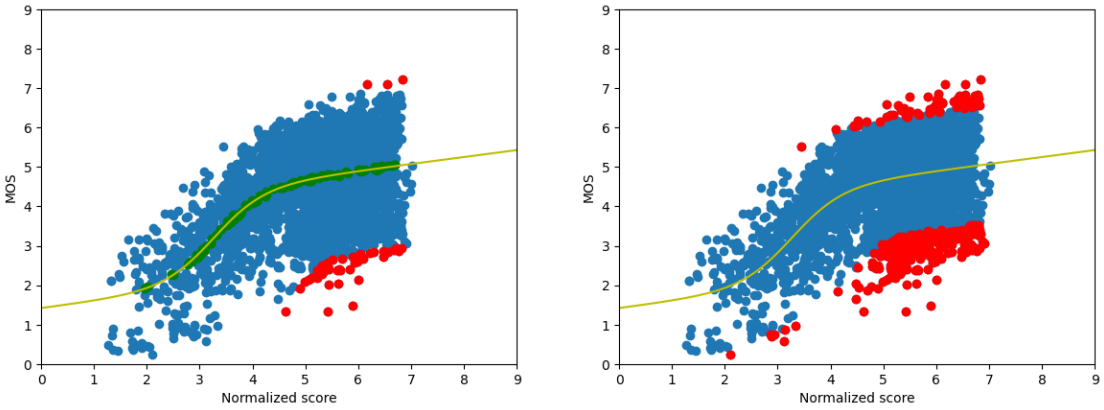
Figure 4.6: *Examples of outlier finding using RANSAC with different amounts of outliers (5% in the left and 30% in the right) on the TID2013 dataset. In the graphs, red dots represent the outlier images, while green dots indicate the best prediction. The black line represents the linear model fitted by the RANSAC algorithm.*



(a) *Amount of outlier = 5%* (b) *Amount of outlier = 30%*

Figure 4.7: *Examples of outlier finding using logistic mapping with different amounts of outliers (5% in the left and 30% in the right) on the CLIVE dataset. In the graphs, red dots represent the outlier images, while green dots indicate the best prediction. The yellow line represents linear nonlinear mapping.*

distinguish from other data. In addition, the mapping line also captures the relation between the MOS and predicted scores better than the result from the RANSAC method. For example, in Figure 4.8, the mapping line represents a nonlinear association of the two types of quality score, which also can be observed from the shape of the data cloud. Therefore, this method is more suitable for data that



(a) *Amount of outlier = 5%* **(b)** *Amount of outlier = 30%*

Figure 4.8: *Examples of outlier finding using logistic mapping with different amounts of outliers (5% in the left and 30% in the right) on the TID2013 dataset.In the graphs, red dots represent the outlier images, while green dots indicate the best prediction. The yellow line represents linear nonlinear mapping.*

contains a nonlinear relationship between the subjective and objective scores than the previous methods.

4.2.5 Results and discussion

In this section, we will find the answers to the following questions:

1. Which images are problematic for each IQA model?
2. What are the common features shared between the outlier images for each IQA metric?
3. Are there any common problematic images for all the IQA models, regardless of their architecture?

To answer the first question, we selected the outliers that are found by all three outlier detection methods. By combining the result from all the previous experiments, we have a robust result. Figure 4.9 shows the outlier finding on the data predicted by the DBCNN on the CLIVE database. In the graph, the outliers detected by the correlation coefficient, the RANSAC, and the logistic mapping are colored in yellow, green, and red, respectively.

Some representatives of the outliers from the DCBNN model on the CLIVE and the KonIQ-10k are shown in Figure 4.10 and Figure 4.11, respectively. The top row shows the images of underestimated predictions, in which the objective scores

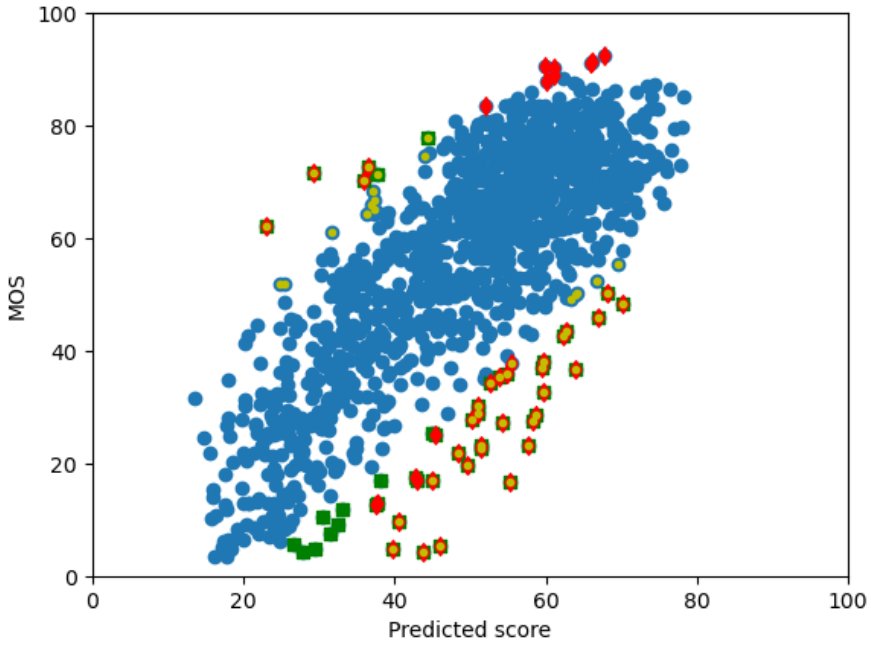


Figure 4.9: *The outlier detection result when combining the three methods performed on the data predicted by the DBCNN on the CLIVE database. In the graph, the outliers detected by the correlation coefficient, by the RANSAC and by the logistic mapping are represented by the colors yellow, green, and red, respectively. As shown in the figure different outlier detection approaches detect the same image as an outlier.*

are higher than the subjective score. The images of overestimated predictions are in the bottom rows. The number of underestimated predictions over the number of detected outliers for each IQA model on different datasets are reported in Table 4.2. We notice that the metrics tend to make more overestimated predictions.

Table 4.3: *Number of failure predictions of each NR-IQA model on each dataset. "Under" columns indicate the number of images whose quality is underestimated by the IQA metric, "Over" columns represent the number of overestimated cases.*

	DBCNN		CNNIQA		SPAQ		KonIQ		MUSIQ	
	Under	Over	Under	Over	Under	Over	Under	Over	Under	Over
KonIQ-10k	4	15	13	17	11	15	5	20	9	17
CLIVE	4	20	4	21	11	22	9	21	14	19
SPAQ	5	25	6	18	5	26	5	13	5	21
TID2013	0	33	1	11	0	29	0	25	4	27

Figure 4.10: *The images that the DBCNN model fails to estimate their perceptual quality in the CLIVE database. Top row: underestimated prediction, bottom row: overestimated quality.*

Question number two is answered by comparing the detected failures case of the same metric on each dataset. Due to the limit of space, we will not show all the outliers images in this report. For the DBCNN model, we notice that the majority of outliers in the authentic databases contain either blurred or dark backgrounds on a large part of the image. The results on the synthesized dataset, the TID2013 are the images that were distorted by the JPEG 2000 transmission, which causes a similar effect as blurring (Sun et al., 2020).

Figure 4.11: *The images that the DBCNN model fails to estimate their perceptual quality in the KonIQ-10k database. Top row: underestimated prediction, bottom row: overestimated quality.*

Figure 4.12 shows the common outliers for the four IQA models. The top-left image is taken from the CLIVE database, the bottom left is from the TID2013, and the rests are from the KonIQ-10k dataset. This is the answer for the question number four we listed at the beginning of this section. All the NR-IQA models predict that these images have a higher quality than the subjective opinion. By perceptual assessment, we notice the similar features between these images: they are all blurry. This point out popular failure scenarios in which NR-IQA methods do not work. This observation should be considered in the design of future IQA models for better performance by tackling the issue of predicting the perceptual scores for blurry images.

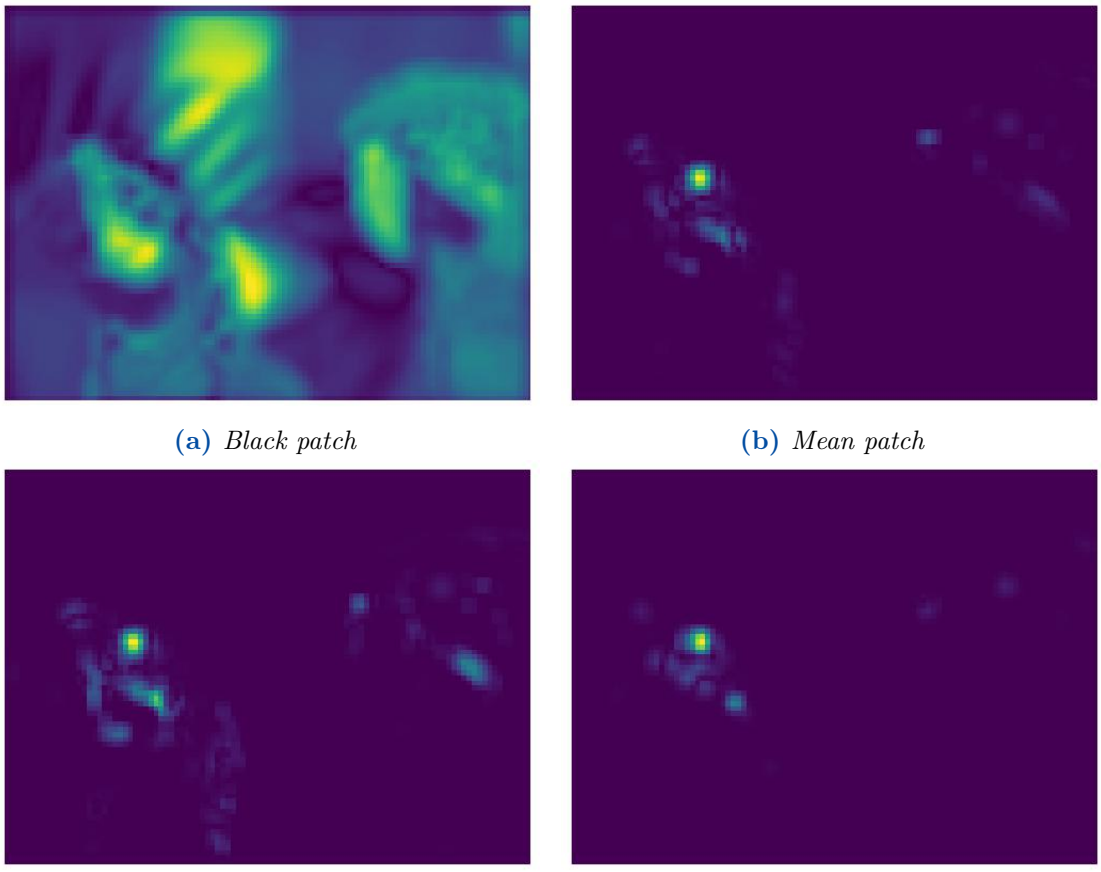
Figure 4.12: *Common outliers with all IQA models.*

In this section, we described the experiments to find the failure prediction of NR-IQA models. It can be seen that the models face difficulty in estimating the quality of blurry images. Some models of CNNs such as DBCNN and CNNIQA have another challenge with images that were captured in low illumination. In the

next experiment, we will explain which part of those outliers that are important in the decision of the NR-IQA model.

4.3 Spatial domain perturbation

In this part, we will have to find the answer to the following questions: In spatial space, which regions in the images play an important role in the prediction of an NR-IQA model? We will use the perturbation-based method, that was described in section 3.3.1 to generate the attribution map of each image area. Firstly, we will need to find which type of mask and which size of the hidden patches will produce a more effective visualization.



(a) *Black patch* **(b)** *Mean patch*

(c) *Median patch* **(d)** *Smooth blur patch*

Figure 4.13: *The attribution maps produced by using four types of patch perturbation with the input image in Figure 3.6a, and the CNNIQA model. The brighter color represents the more importance of the pixels to the predicted score.*

We implemented the algorithm in section 3.3, with the four perturbation types: black, mean, median, and smooth blur patches. Figure 4.13 shows the attribution maps of the original image in Figure 3.6a with the quality score estimated by the CNNIQA model. Similar results for other models are shown in Figure A.6 and A.7. The brighter color represents more importance of the regions to the overall quality score of the image. Apparently, the visual map corresponding to the black type is the most detailed, in which we can recognize the features from the original image; meanwhile, the results from other types of hiding patches are not meaningful. We argue that this result may be affected by the property of the input image, which has a large *out of focus* background. Thus, replacing the patches in this area with a mean, median, or blur alternative will actually not change the image. However, when the four types of patch replacement are tested on another texture image (similar to the bottom right image in Figure 4.12), we still have the same observation: using the black patches as the replacement for patches will provide more *meaning* attribution map.

Although the reason which is mentioned above for choosing black patches as the most suitable perturbation type seems appropriate. Our selection method is subjective, which is based on the observation of some images when we use the same normalization to visualize the attribution map. Therefore, it may not be the best choice for the general datasets. In particular, when different normalization techniques are applied to the attribution maps from each perturbation type, more image features can appear in the map and make them become more *meaningful*. We realize the limitations of our work. However, because there is no available way in the literature to validate the explanation for the IQA problem, this is the feasible way that can be used.

With regards to the size of the removed patch, we see that if it is small, the computation cost will increase, but more detail of the original image features will be captured in the attribution maps. On the other hand, if the size of the patches is large, the computational time is reduced, but the attribution maps will be more homogeneous as all the pixels within the patch size are aligned with the same importance value. From our empirical experiments, the removal of image patches of size 15x15 with a stride equal to 5 gives the best visualization result for the input image of resolution 384x512.

Figure 4.14 and Figure 4.15 show the corresponding attribution maps of the four common outlier images (Figure 4.12) to the prediction score estimated by the CNNIQA and the DBCNN model, in respectively. Attributions maps for other NR-IQA models can be found in Figure A.8 - A.10. Pixels in yellow colors indicate more influence of the feature on the output prediction score, while those in darker colors suggest less importance. We can notice that all the generated attribution maps are coarse, without fine-grained details. However, several interesting observations

Figure 4.14: *Attribution maps corresponding to four images in Figure 4.12 to the quality prediction by the CNNIQA model. The brighter color represents the more importance of the pixels to the predicted score.*

can be obtained when comparing the maps and the original images. The CNNIQA models focus more on the homogenous regions of bright color in the images to estimate the quality score. For example, the sky in the bridge image is associated with the color yellow in the attribution maps; and the wall in the image of people also corresponds to the bright color in the produced map. Meanwhile, the DBCNN focuses on the pixels representing the water in the bridge photo, and three people in the top-left photo. When the NR-IQA models have a deeper network of more hidden layers, it is difficult to interpret the generated attribution maps. For example, in the case of the KonIQ model with the InceptionRestNetv2 network which consists of 164 layers, the attribution maps do not highlight any particular regions in the image. Similarly, with the MUSIQ models of vision transformer architecture, four attribution maps are abstract.

Our experiment results suggest the perturbation method of explaining a single

prediction on the spatial domain can provide attribution maps, which represent where the NR-IQA model looks for in the image to predict its quality. This is equivalent to the first expectation which was mentioned in section 3.1 of the explanation for IQA. If the number of hidden layers in the networks is small, the attribution maps show patterns in the image. However, this method is not suitable for explaining the network of deep layers, as the produced attribution maps become more abstract.

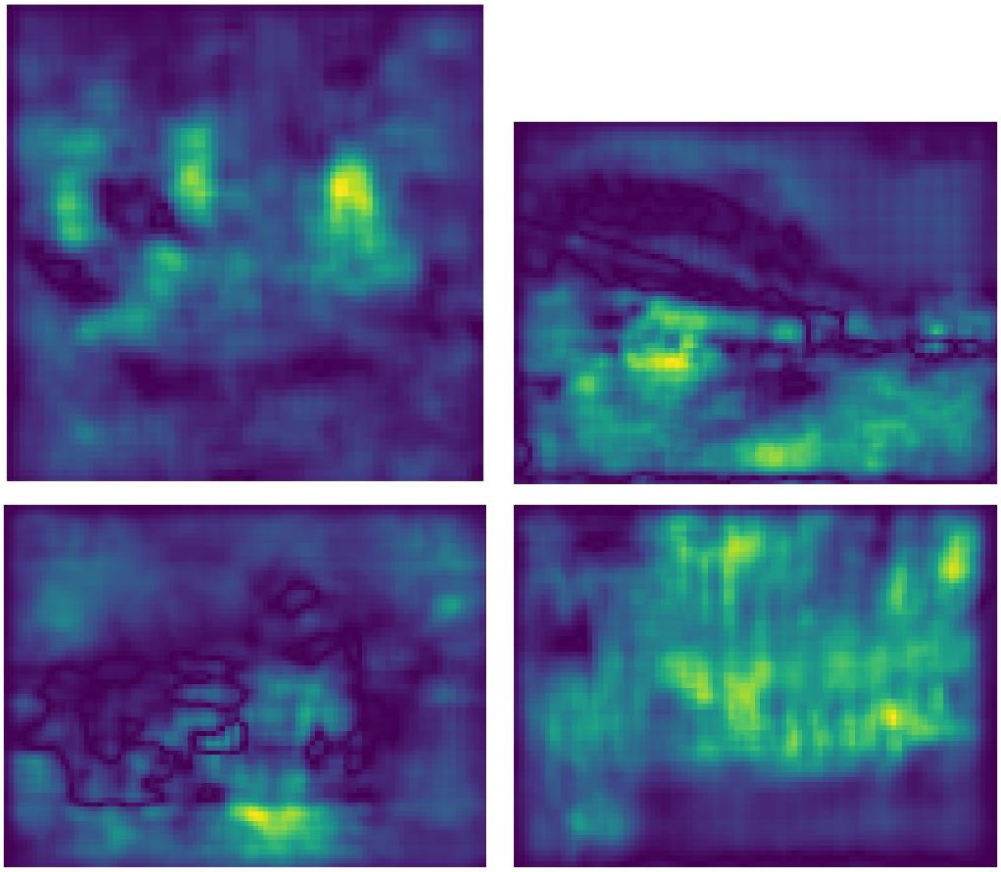


Figure 4.15: *Attribution maps corresponding to four images in Figure 4.12 to the quality prediction by the DBCNN model. The brighter color represents the more importance of the pixels to the predicted score.*

4.4 Frequency domain perturbation

This part of our study aims to explain an IQA model in terms of how well it mimics the human sensitivity to contrast in different frequency bands. The purpose of this

section is to address the second and third expectations of XAI for IQA which were mentioned in section 3.1. We will sequentially exclude the image information in different frequency bands and measure the change in the quality prediction from the models.

With each outlier image from Figure 4.12, the importance of image data in each frequency level is represented in Figure 4.16, the objectives scores were produced by the DBCNN model. The results for other models can be found in the Appendix in Figure A.11 - A.18. The vertical coordinate represents the contribution of image data at each frequency band to the estimated quality prediction, and the horizontal coordinate represents the frequency. The negative values indicate that the removal of the corresponding frequency band will result in an increase in the predicted quality score. It happens that the importance distributions are different among outlier images, however, we can notice a common similarity between the four plots: the data at low frequency bands (except the lowest band) are somewhat misleading the model. For all four outliers, removing the data at some low-frequency bins even increases the predicted quality scores.

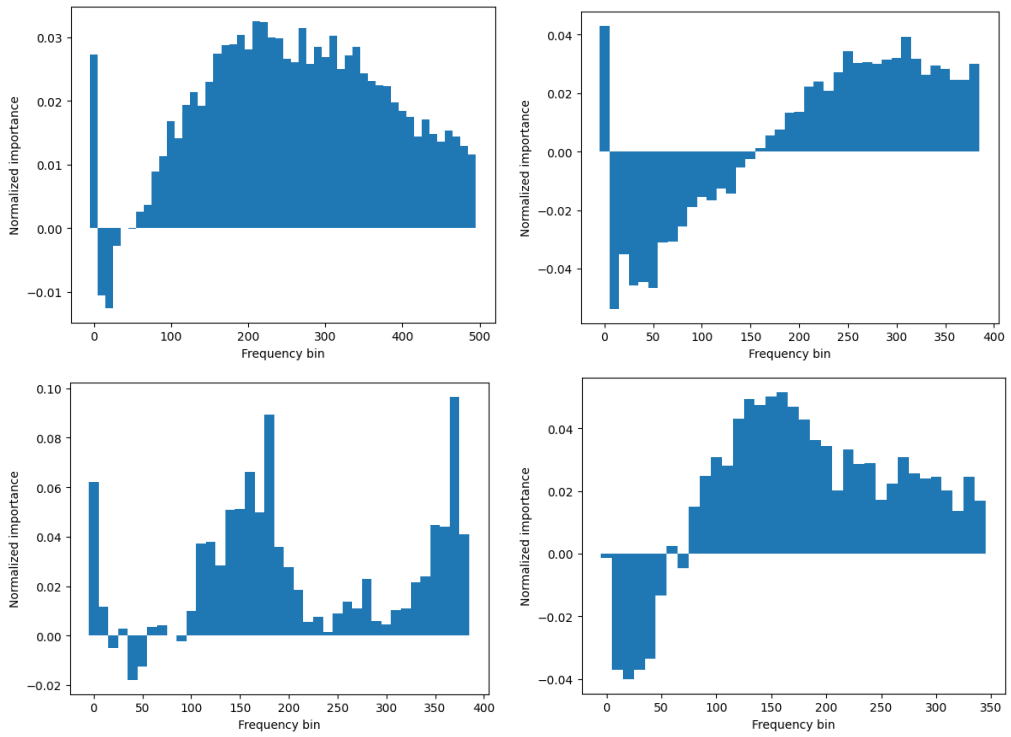
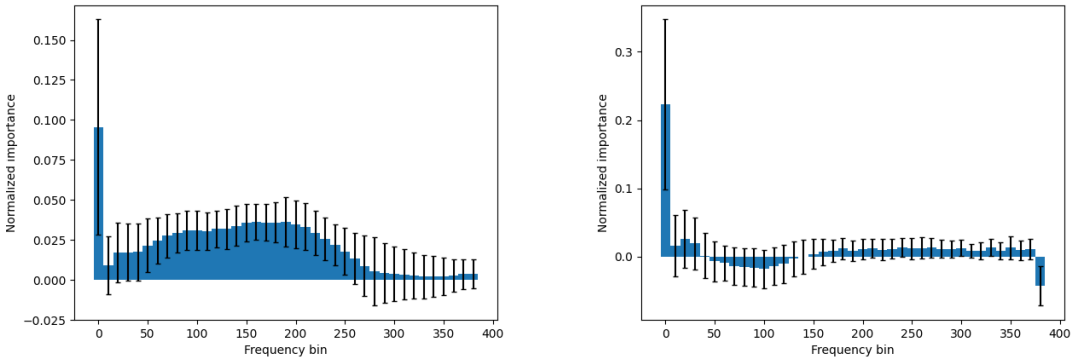


Figure 4.16: *Contribution of data in each frequency band to the estimated quality prediction by the DBCNN model of the four common outliers in Figure 4.12.*



(a) *Gaussian blur* **(b)** *High frequency noise*

Figure 4.17: *Contribution of data in each frequency band of two types of distorted images to the estimated quality prediction by the DBCNN model. The error bar indicates the standard deviation.*

We will compare these graphs with the results from the good prediction images. Because the different distortion types can result in different shapes of the importance distribution, we provide the result for two common types of distortions: Gaussian blur and high frequency noise. Figure 4.17 shows the average importance distribution for the images which are distorted by the two types of distortion to the prediction by the DBCNN model. The graph on the left suggests that the IQA model weighted more score on the lowest frequency bin, the middle bins are also important with lower importance values; meanwhile, the data at the high frequency bands seem to not contribute significantly to the final prediction result. Considering these good prediction images are distorted with the same visual appearance effect of the outlier images - blurring, we notice the difference: the model assigns more weight to the high-frequency information of the outliers. That might be a reason for the inaccurate predictions of the IQA model on the challenge images. Similar observation can be seen from the results produced by the CNNIQA model, which is shown in Figure A.11 - A.12. Thus, the difference in weight the models assign to each frequency in the outliers and the common images lead to the mismatch of the objective and subjective quality scores. The results for other models are shown in Figure A.13 - A.18.

Comparing the two graphs in Figure 4.17, we see that the importance values of the mid-range frequency in the images which were undergone high frequency noised are very small, and if the highest frequency band is removed, the overall quality will be better. This explanation shows that the model (DBCNN) does mimic HVS as the distortions are embedded at high-frequency levels. This provides insight into how image features at different frequency bands contribute to the estimated

quality of the image. This knowledge can forest the trust in the models, as an ideal model should predict the quality of an image the same way our HVS perceives it. However, if a model does not show a similar property as HVS, it does not mean that it can not be trusted at all. In that case, the analysis of what image features in the spatial domain, or which information is encoded in models during the process of making a prediction should be considered.

4.5 Color domain perturbation

In this section, we investigate how the NR-IQA models react to changes in achromatic and chromatic dimensions in the HSV color space. The conversion from the RBG to HSV and vice versa was implemented by using the built-in function of *skimage* library. The intensity of pixels in HSV is in the range [0.0, 1.0]. The score changes were collected for all the combinations of replacing pixel values in hue, saturation, and value channels with a new pixel value.

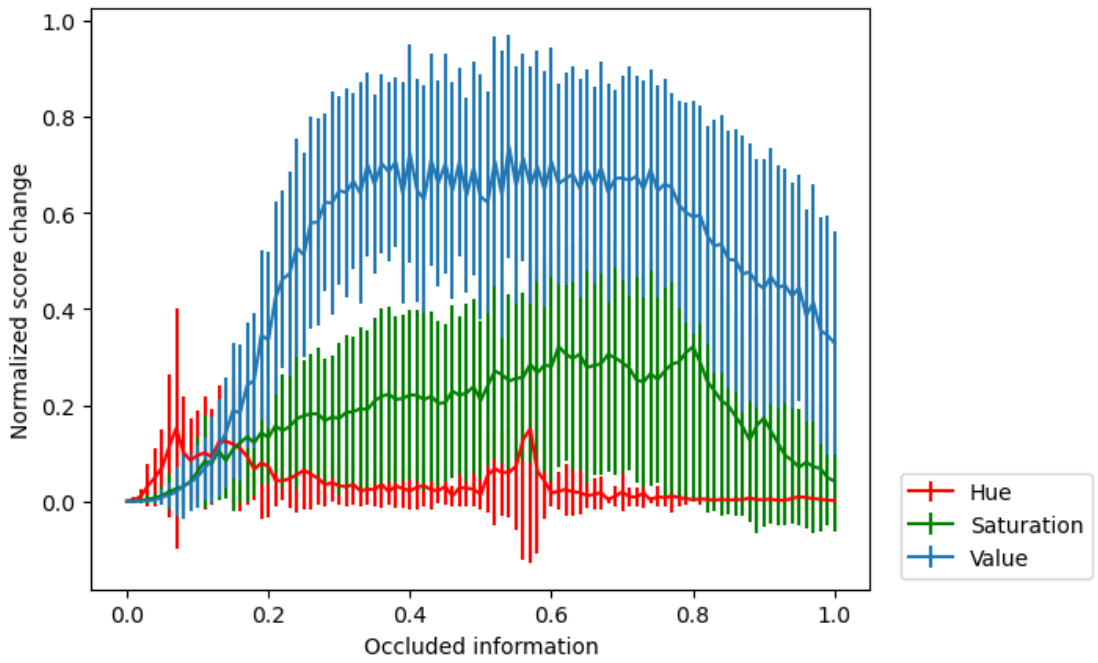


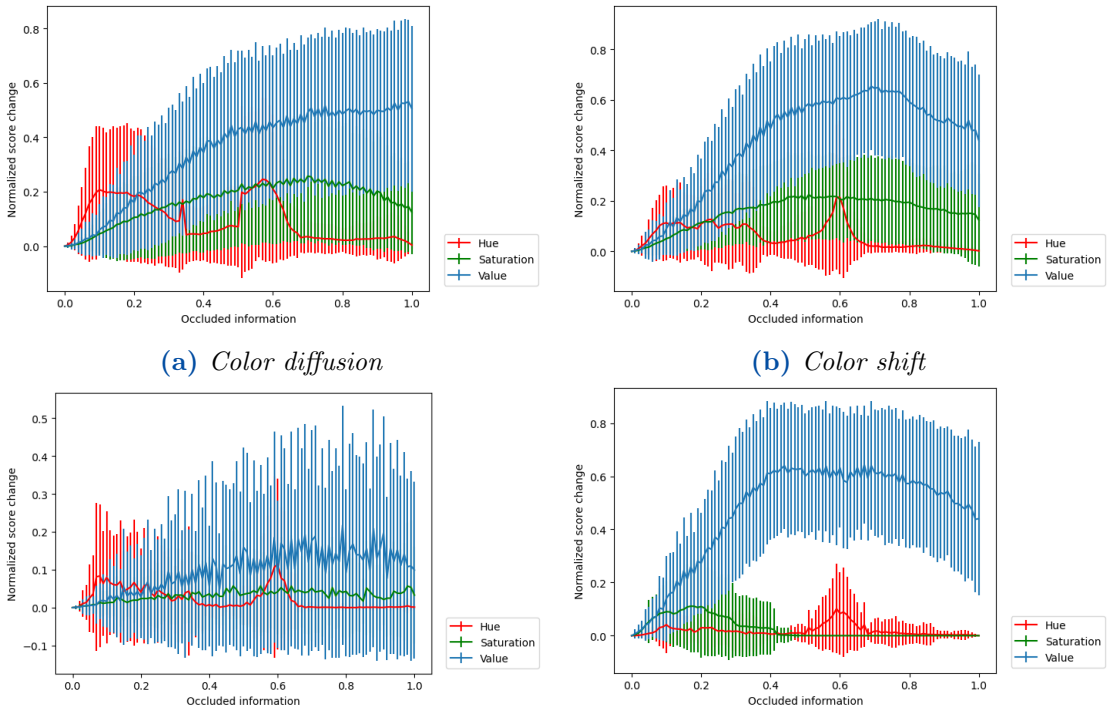
Figure 4.18: *The trend of estimated quality score change from the DBCNN model when perturbing hue, saturation and value channel. The error bars indicated the standard deviation.*

First of all, we notice that there is no common particular hue that significantly contributes to the quality score of images of all IQA models. If we *delete* any color

from the input image and replace it with another color, its corresponding objective quality score differs, but score differences are similar between all the perturbations. Meanwhile, changing the value channel of images leads to a significant change in objective quality scores. Figure 4.18 represents the tendency of the change of estimated quality scores from the DBCNN model on the set of the undistorted images in the TID2013 dataset Ponomarenko et al. (2013). In the graphs, the red line denotes the score change when the hue channel is perturbed and the two other channels are preserved. The green and blue lines represent the perturbation in only saturation and value channel, respectively. The correspondence graphs of the other three IQA models can be found in the Appendix (Figure A.19 -A.20). There is a clear agreement between the graphs in those figures: the blue lines are the highest, followed by the green line, and the red lines are in the lowest position. They suggest that all NR-IQA models that we are considering are more sensitive to the change in value channel than in the hue information on the set of images from the TID2013.

As the color space is the knowledge domain that we want to explore, the same experiments were implemented on the images which are degraded by color-based distortions in the KADID-10k database (Lin et al., 2019). Four sets of images correspond to four distortion types: color diffusion, color shift, color quantization, and color saturation are investigated separately.

The description of each distortion type from Lin et al. (2019) provides information about how they were processed from the pristine image. For example, color diffusion images were collected by applying Gaussian blur to the color channels (a and b) in the Lab color space; color shift: randomly translating the green channel, and blending it into the original image; color quantization: quantization and dithering the original image to 8 -64 colors; and color saturation: multiplying the saturation channel in the HSV color space by a factor then convert back to RGB space. The subfigures in Figure 4.19 show the tendency of the change in estimated quality scores by the DBCNN model of these distorted images when each channel: hue, saturation and value is perturbed while the two other channels are kept remaining. We can see that the three graphs 4.19a, 4.19b and 4.19d share the same pattern in which the scores change corresponds to the perturbation in hue channels is the largest, followed by the scores changes of the perturbation of saturation and hue channels. However, the red and green lines in these graphs are not completely separated but intersect at some points. Especially, the three lines which represent the score change of perturbation in three channels in the graph 4.19c are located in the same position. It indicates that the model reacts to the change in the three channels similarly to each other. Combining the observation in Figure 4.18 and this observation, we could not claim that the model is more sensitive to the change in achromatic than the change in chromatic information.



(a) *Color diffusion*

(b) *Color shift*

(c) *Color quantization*

(d) *Color saturation*

Figure 4.19: *The trend of estimated quality score change from the DBCNN model when perturbing hue, saturation, and value channel on subsets of color-distorted images. The error bars indicated the standard deviation.*

In this section, we discussed the results of using the perturbation method on color space to check if a NR-IQA model follows the same property of HSV with regard to the change of information in achromatic and chromatic images. Our experiment results suggest that we can not claim whether the model work with this principle or not. Although the explanation method which was used is understandable and easy to extend to other information domains, it has several drawbacks. Its first disadvantage is data-dependent: the explanation results rely on the input image and they can be contradictory with different images. Thus, it is difficult to form a reliable explanation of the model. Additionally, the perturbation process which hides a part of color information may cause unaware distortion on the perturbed version of the image, which in the end, affects the explanation results. Therefore, future work which focuses on designing a better way of removing information without seeding more distortions on the original image is important to XAI for image quality assessment.

4.6 Other XAI methods

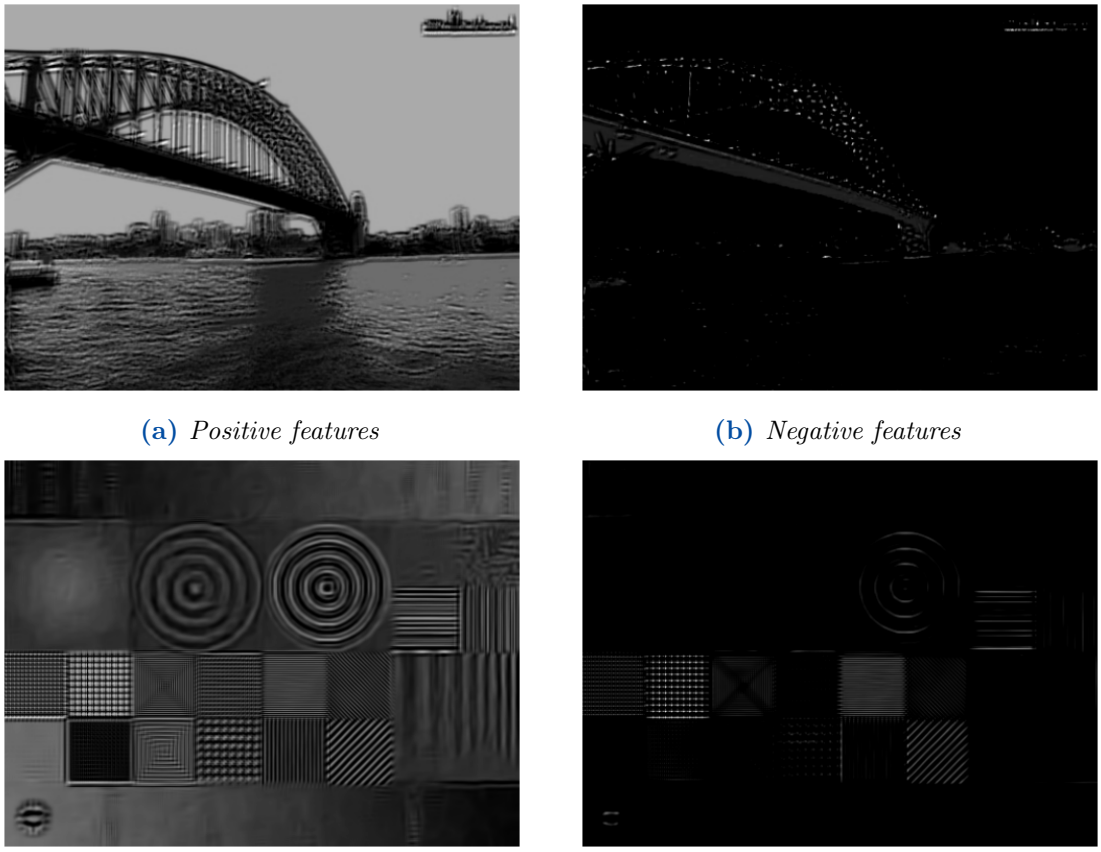
In the previous sections, we treated NR-IQA models as black boxes and implemented experiments without the knowledge of the model's architecture. In this section, we will use XAI methods (Selvaraju et al., 2017), (Springenberg et al., 2014) for CNN models that require access to each layer of the models. These XAI methods were originally proposed for classification (Schöttl, 2022) and are modified for regression models in our work. In the following, we will describe the selection of XAI methods for each NR-IQA model.

CNNIQA

Among the five NR-IQA that are considered in this work, CNNIQA has the simplest architecture with only one convolutional layer, follows by three fully connected layers. Because of its simplicity, using Grad-CAM already provides good explanation results. Figure 4.20 shows the feature in the images that increases the estimated quality score (in the left) and those will lower the quality score (in the right) of the two images in the left of Figure 4.12. More visualization results of other images can be seen in the Appendix in Figure A.21. The brighter pixels in the positive feature maps indicate a larger contribution toward the increase of model output, and those in the negative feature maps indicate a large contribution toward the decrease of model output. On the other hand, the dark pixels that appear in both types of feature maps have less influence on the quality of the image. These feature maps address the second condition of expected explanation which is mentioned in section 3.1.

We can see that in the photo of the bridge, because of the sky, the model predicts the image has high subjective quality; meanwhile, the details of the bridge lower the overall quality score. However, when judging this image, human observers tend to not focus on the white sky but notice the faded color. That can be a reason why the model fails to estimate the subjective quality score of this image. In the image of texture (in the second row), we can see the highlight in the smooth area in the maps of positive features, and some texture appears on the negative map. It is important to point out that many smooth areas in this image are caused by distortion. Thus, they should lower the estimated score of the model, which is in contrast to what is presented in the positive feature map. Therefore, the model is not able to estimate the image's perceived quality.

From the above feature maps, it seems like this model rates an image of high quality if they have a large homogenous area. However, looking at the corresponding feature maps (in Figure A.21) of other images, we have a contrast observation: the positive features in those images highlight the edges of objects (for example, the tables, the chair, or the border of the flashlight); and the negative features indicates the smooth area. The inconsistency in the influence of different image



(a) *Positive features* (b) *Negative features*

(c) *Positive features* (d) *Negative features*

Figure 4.20: *The positive and negative features visualized by Grad-CAM with the CNNIQA model. The brighter pixels in the positive feature maps indicate a larger contribution toward the increase of model output, and those in the negative feature maps indicate a large contribution toward the decrease of model output.*

features (edges, smooth regions) on the predicted quality of images brings distrust in this model.

DBCNN

The DBCNN has an architecture of two feature extractors: one was tailored from the pretrained classification model VGG16, and the other was adopted from a pretrained distortion classification model S-CNN. Because the model consists of two sub-networks, and using Grad-CAM requires the selection of a convolutional network to visualize the learned feature; we can not decide which of the last convolutional layers in two branches is more important to the prediction score. Therefore, Guided Backpropagation was used to interpret the extracted features

from the input image.

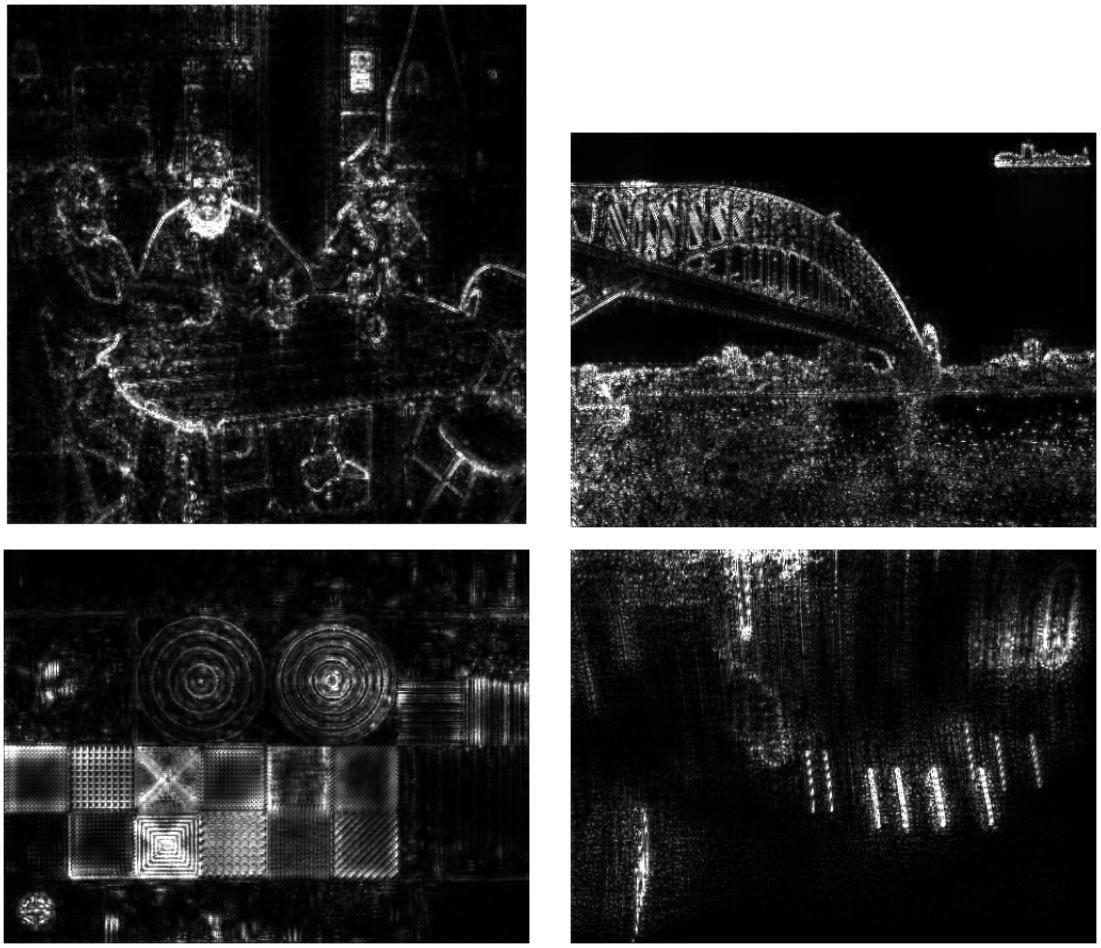


Figure 4.21: *The relevant features visualized by Guided Backpropagation with the DBCNN model. The brighter pixels indicate more attribution to the output of the model.*

Figure 4.21 shows the features in the outlier images that are relevant to the objective quality score estimated by the DBCNN model. They represent the answer to the second condition of expected explanation which is mentioned in section 3.1. From the four relevant feature maps, we can see that the models mostly detect the edges of the image to make their prediction about image quality In the top-left image, the faces of three persons are highlighted, and interestingly, the neck of the man and a small poster on the windows seems to be the most important features that affect the quality score of the image. In the image of the bridge, all the edges of the objects (bridges, buildings, and even the waves) are detected and contribute significantly to the output of the model. Similar observations can be seen from the

bottom-row images, in which the edges of strong gradients are illustrated in the relevant maps.

From these feature maps, users can decide whether to trust the model or not, based on their subjective evaluation of the relevant features. For example, if a human observer pays attention to the necks of a person in an image to rate its perceptual quality as the model does.

SPAQ

The SPAQ model adopted the pretrained ResNet-50 classification model and retrained it for image quality assessment. We also use Guided Backpropagation to visualize the pixels in the input images that affect the output of the model the most.

Figure 4.22 shows the features in the outlier images that are relevant to the objective quality score estimated by the SPAQ model. We can see that these feature maps highlight not only the edges of objects in images but also many pixels in the horizontal and vertical directions. Even in the regions of the sky, there are many bright dots shown in the relevant feature maps. In the bottom right images, the feature maps focus on all the regions over the images. This suggests that the model either fails to extract features from this image or looks at all the regions in the image with the same attribute to predict the quality score. Moreover, the appearance of the bright dots in both horizontal and vertical lines can suggest that this model is sensitive to the distortions that cause degradation in those directions. For example, compression artifact that leads to block distortions (Unterweger, 2013). However, a large number of images needed to be considered to confirm this observation.

KONIQ

The KonIQ model adopted the deep network InceptionResNetv2 classification model and retrained it for image quality assessment. We also use Guided Backpropagation to visualize the pixels in the input images that affect the output of the model the most.

Figure 4.23 shows the features in the outlier images that are relevant to the objective quality score estimated by the KonIQ model. address the second condition of expected explanation which is mentioned in section 3.1. These feature maps are somewhat similar to those of the DBCNN model (Figure 4.21). However, they are different in some extent. For example, in the top-left image, the KonIQ model focuses less on the facial features of the people but pays more attention to the shoe of the woman on the right. And for the top-right image, this model only treats a part of the bridge as well as some buildings as important features to predict the image's quality.

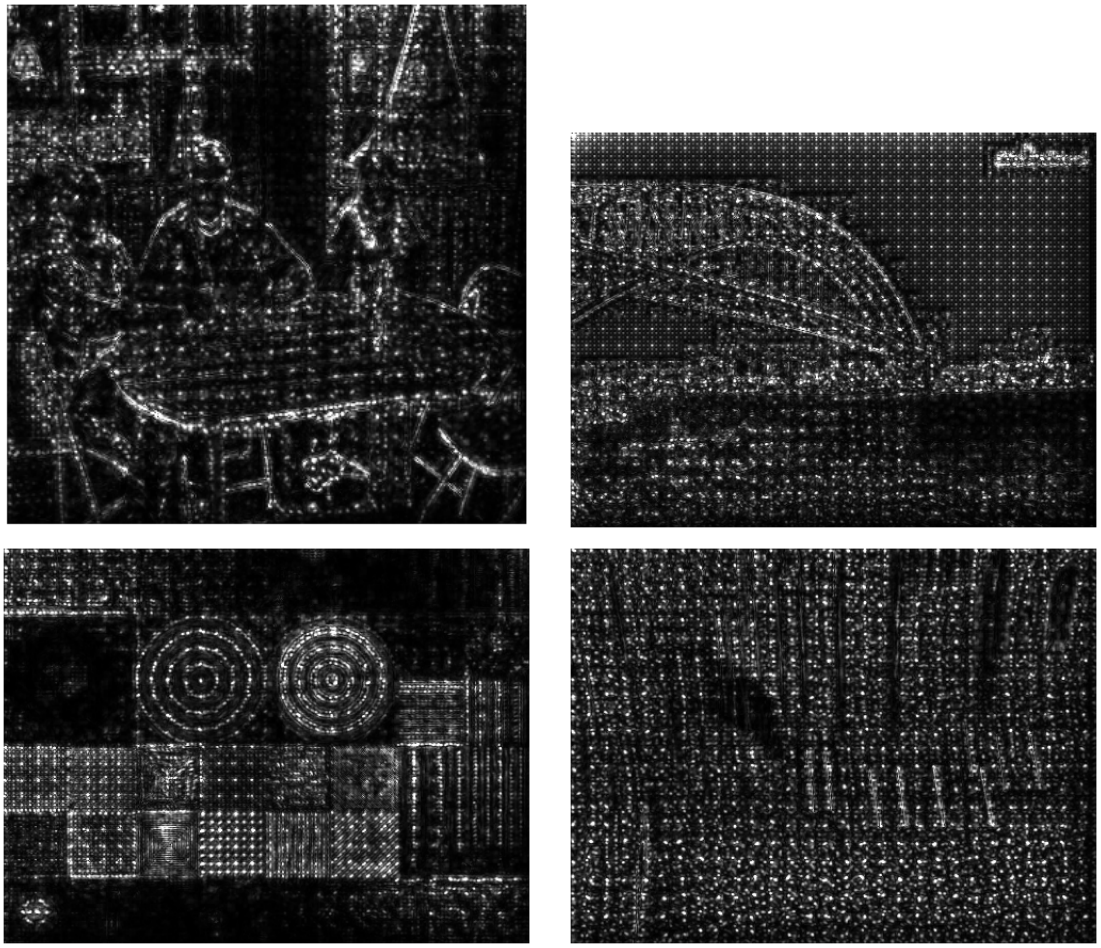


Figure 4.22: *The relevant features visualized by Guided Backpropagation with the SPAQ model. The brighter pixels indicate more attribution to the output of the model.*

In this section, we showed the explanation provided by using the popular XAI methods on NR-IQA problems. Our analysis suggests that these techniques are able to explain the prediction of the model by showing which pixels are relevant to the model's output. This type of explanation satisfies the second condition of good explanations, which was mentioned in section 3.1. However, they are only applicable to CNN-bsed models, and can not explain other types of models such as transformers. Based on the attribution maps of many images, a user can decide whether to trust the model or not based on their subjective evaluation because there is currently no objective framework for assessing the explanation of NR-IQA models. Although it is difficult, this could be an important direction for future work in XAI for NR-IQA.

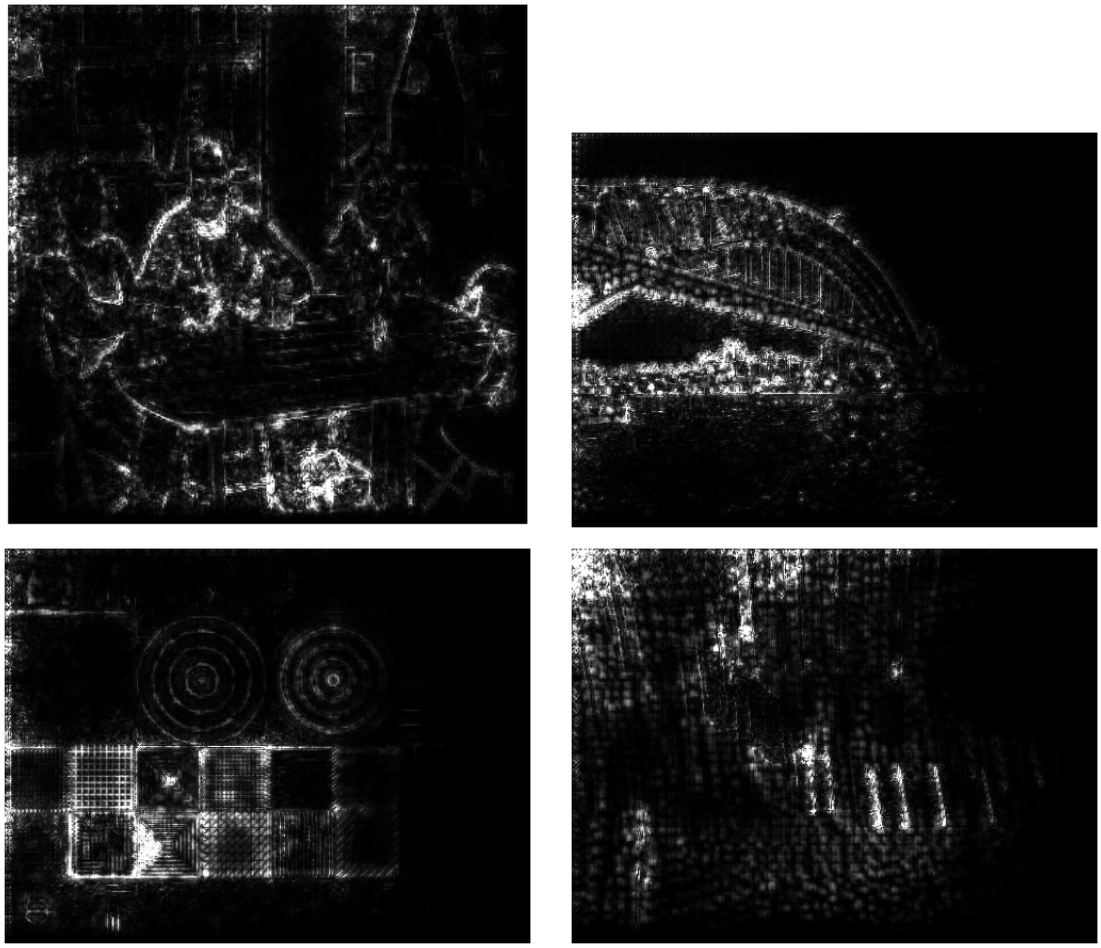


Figure 4.23: *The relevant features visualized by Guided Backpropagation with the KonIQ model. The brighter pixels indicate more attribution to the output of the model.*

5 Conclusions and Future work

5.1 Conclusion

The objective of this work was to provide the explanation for NR-IQA models and answer four questions which were presented in section 1.3. We conducted interviews to find the expected properties of goods explanation for IQA models. A workflow was proposed to provide knowledge about the model. First of all, we provide the limitations of the model by finding the set of images that the model fails to predict their subjective quality. Different outlier detection methods were introduced and compared. As each of them has its own drawbacks and advantages, we figured out that combining the three methods will provide robust results. Our experiments show that the existing NR-IQA models face difficulty in predicting the perceptual quality for images with blurring appearance.

We used perturbation methods, which were applied to any black-box model, on three information domains: spatial, frequency, and color to explain the NR-IQA model. Our experiment shows that the attribution maps from this method can be interpreted for the network with a small number of hidden layers. However, if the networks are deeper, it is difficult to explain the attribution of image features in the overall image quality. Our experiment results in the frequency domain and color space show that it is not feasible to claim whether the model mimics HSV in predicting image quality or not by using the explanation from the perturbation method. Although this XAI method does not require access to the architecture of the model and can be used to explain any *black box* model, its explanation results need to be validated.

We also used the popular XAI methods (Grad-CAM and Guided Backpropagation) which were originally proposed for classification problems to visualize the relevant features in the image that affect the objective quality of images. They resulted in the attribution maps which show what in the images matter to the predicted score of a CNN-based model. These attribution maps shed some light on the understanding of the NR-IQA model, however, users still need to have expertise in the field to build trust in the model.

5.2 Future works

Although our work studied several ways to explain an NR-IQA model, there are a few limitations that can be resolved in the future to bring deeper contribution to the field of IQA:

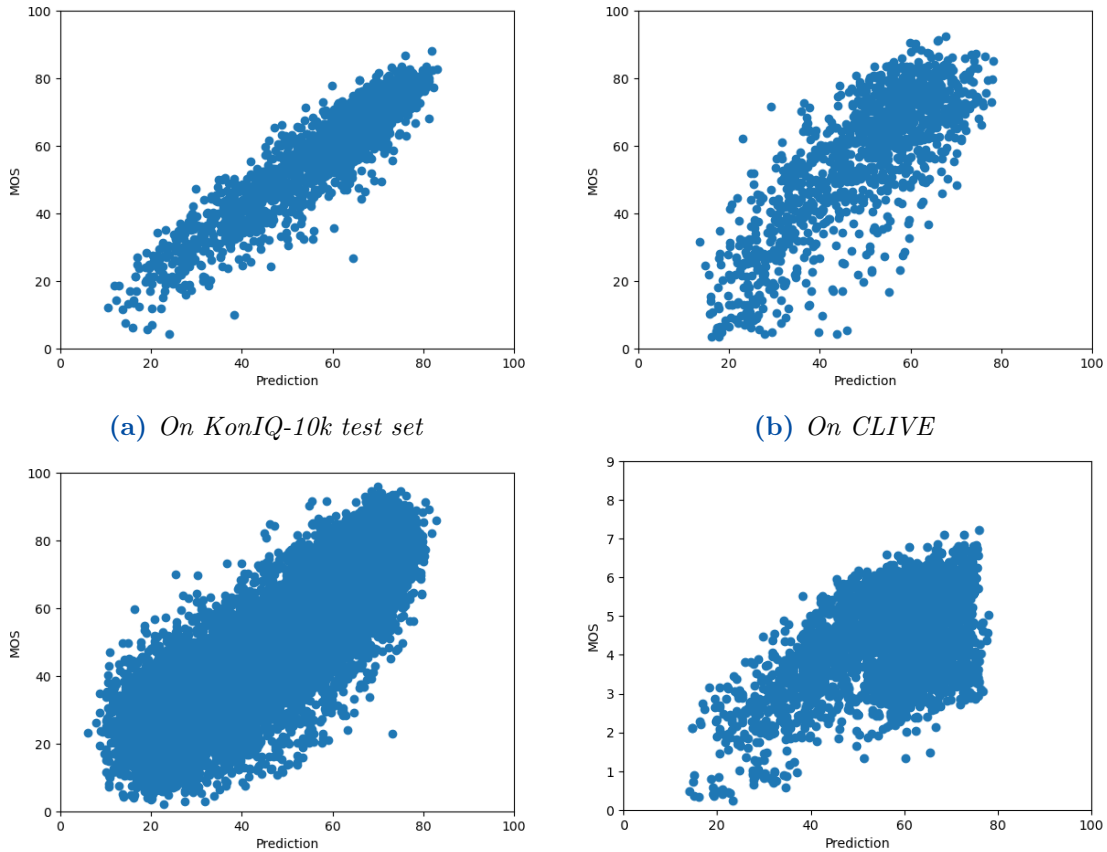
- There are no objective criteria to assess the effectiveness of each perturbation type in the spatial domain.
- The color perturbation results may be affected by the content of the image, if an image has a dominant hue, replacing that with another hue value can significantly change the objective quality score. When we consider a group of images, their dominant color can be different. In the end, if we take the average score changes, they may compensate for each other.

From the preliminary of this work, we can extend the studies in XAI for NR-IQA with several directions, which are the followings:

- Finding a more effective method to hide the information in each patch of the image, in order to not produce more artifacts on the image. In this case, the attribution maps from the perturbation methods will become robust.
- Creating a framework to objectively evaluate the explanations produced by the proposed methods.
- Finding other explanations that use the information from the embedded space in the deep learning model.

A Appendix

The following images are the graphs from the experiments of outliers detections.



(a) *On KonIQ-10k test set*

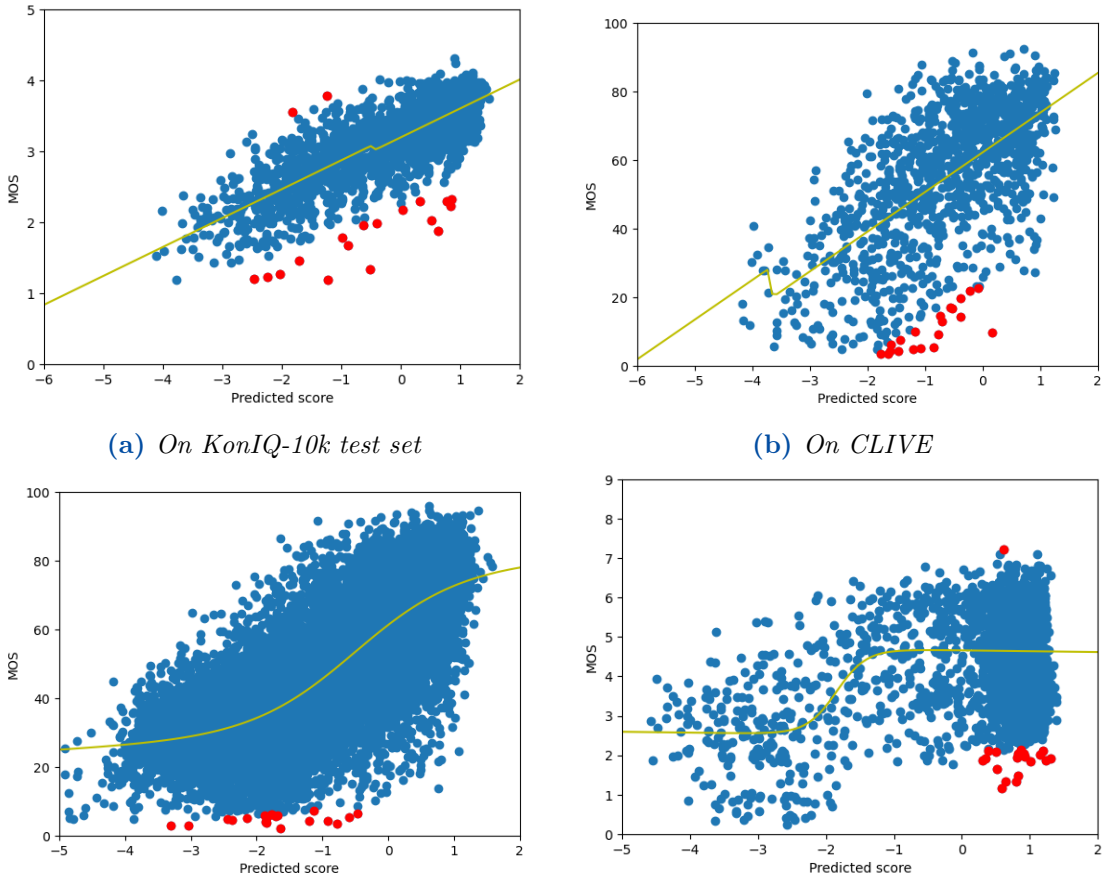
(b) *On CLIVE*

(c) *On SPAQ*

(d) *On TID2013*

Figure A.1: *MOS vs quality score predicted by DBCNN model on four databases.*

The following figures are from the experiments to select which perturbation type is suitable in the spatial domain.



(a) *On KonIQ-10k test set*

(b) *On CLIVE*

(c) *On SPAQ*

(d) *On TID2013*

Figure A.2: *Examples of outlier finding based on correlation coefficient using the CNNIQA model on different databases. In the plots, the red dots represent the outlier images, while the green dots indicate the best prediction.*

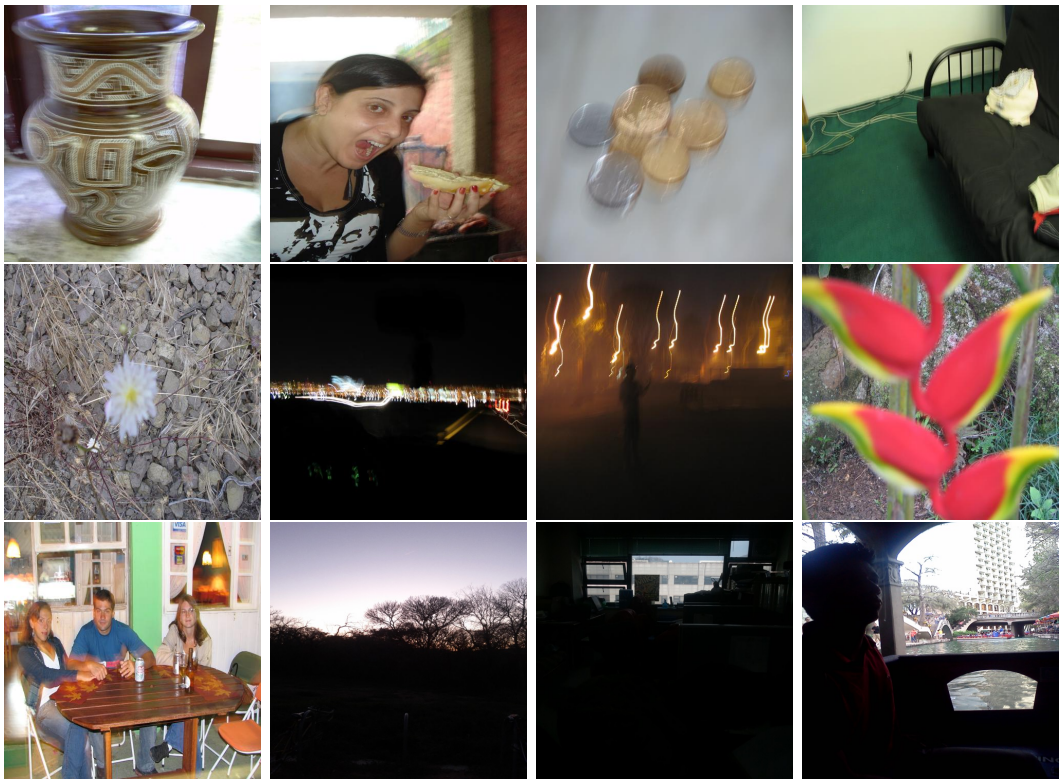


Figure A.3: *The twelve most outliers from CLIVE datasets with the CNNIQA model.*

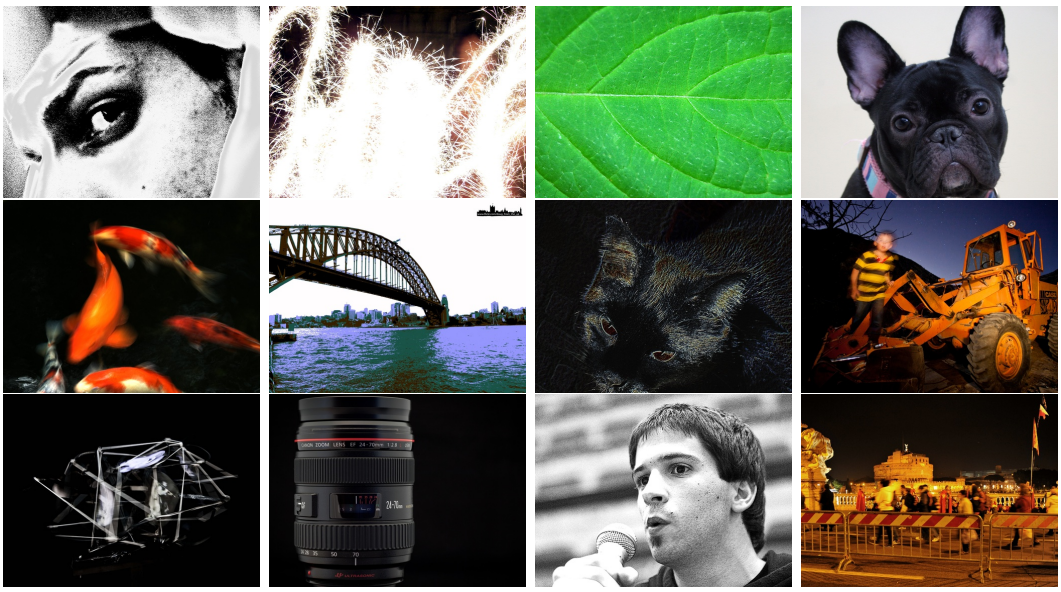


Figure A.4: *The twelve most outliers from KonIQ-10k datasets with the CNNIQA model.*

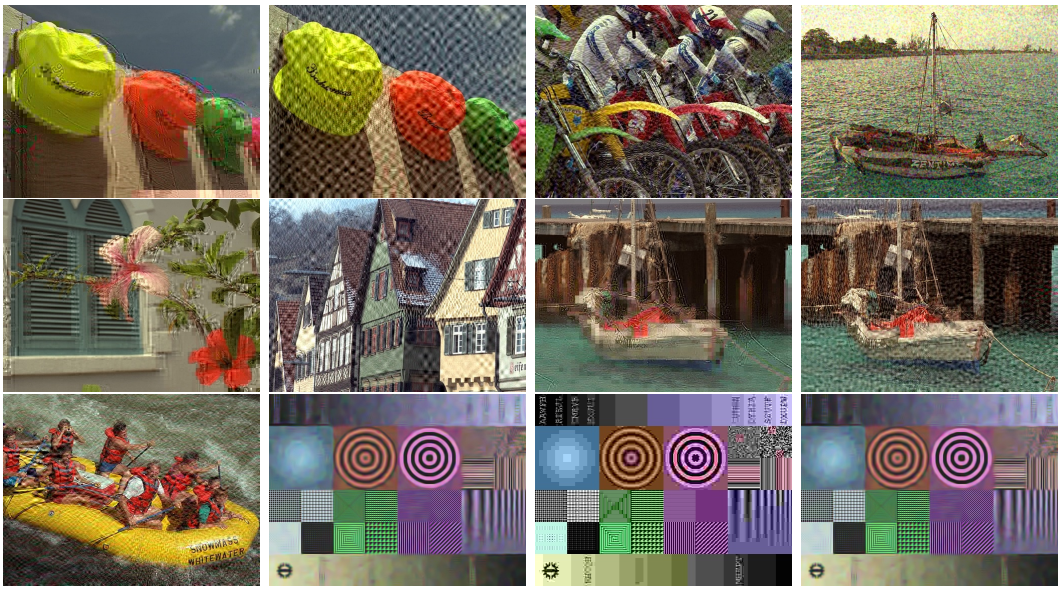
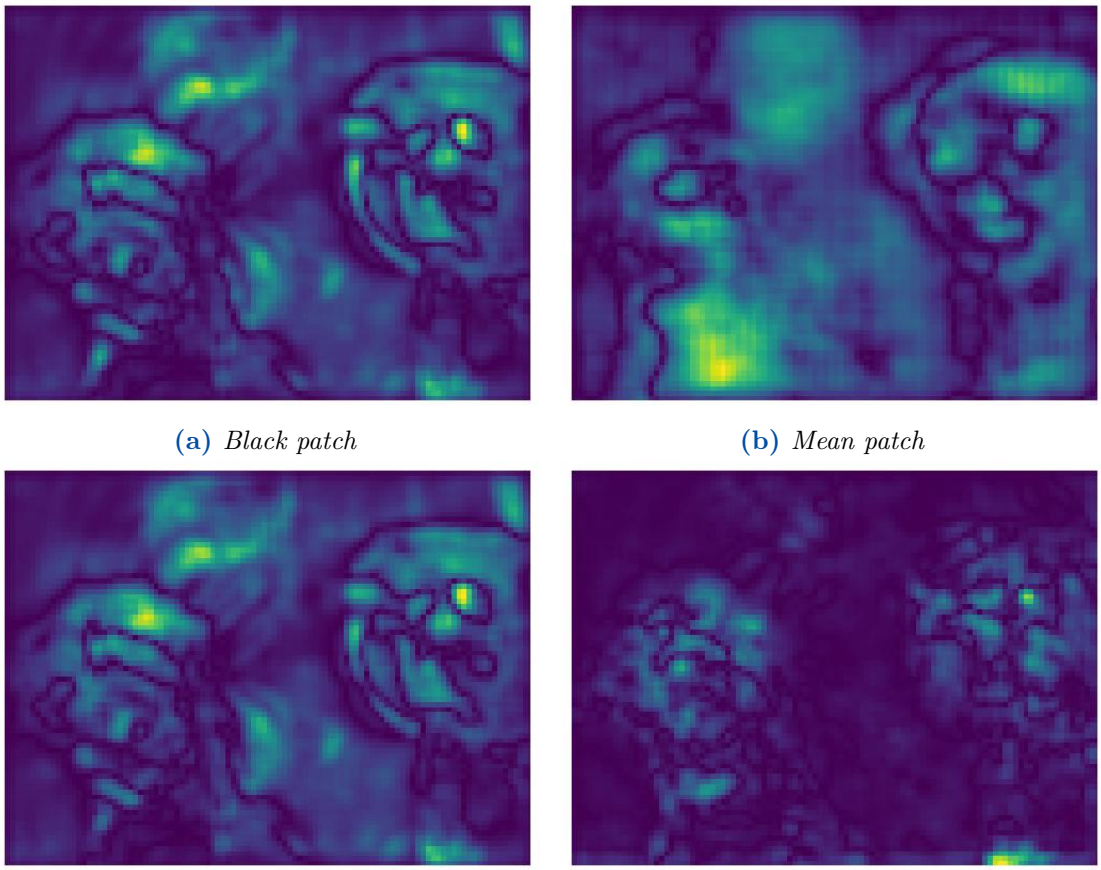


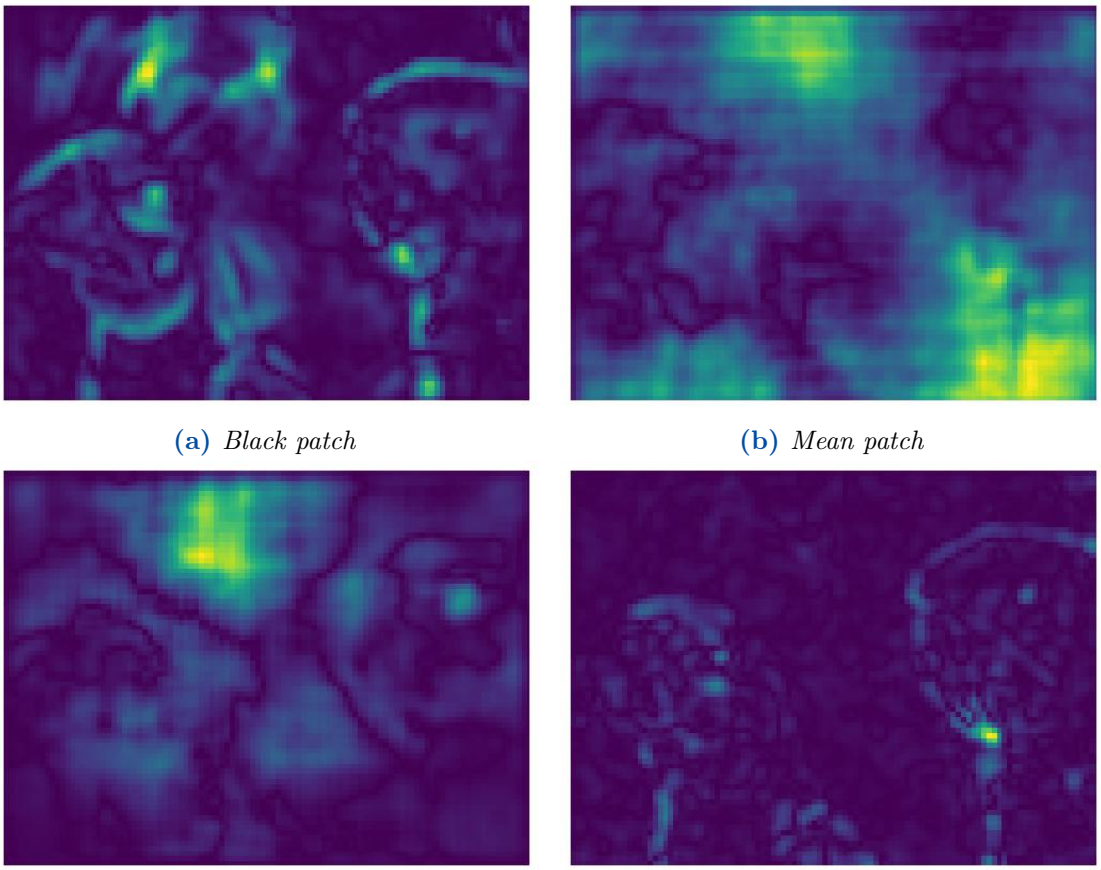
Figure A.5: *The twelve most outliers from TID2013 datasets with the CNNIQA model.*



(a) *Black patch* **(b)** *Mean patch*

(c) *Median patch* **(d)** *Smooth blur patch*

Figure A.6: *The attribution maps produced by using four types of patch perturbation with the input image in Figure 3.6a, and the DBCNN model. The brighter color represents the more importance of the pixels to the predicted score.*



(a) *Black patch* (b) *Mean patch*

(c) *Median patch* (d) *Smooth blur patch*

Figure A.7: *The attribution maps produced by using four types of patch perturbation with the input image in Figure 3.6a, and the SPAQ model. The brighter color represents the more importance of the pixels to the predicted score.*

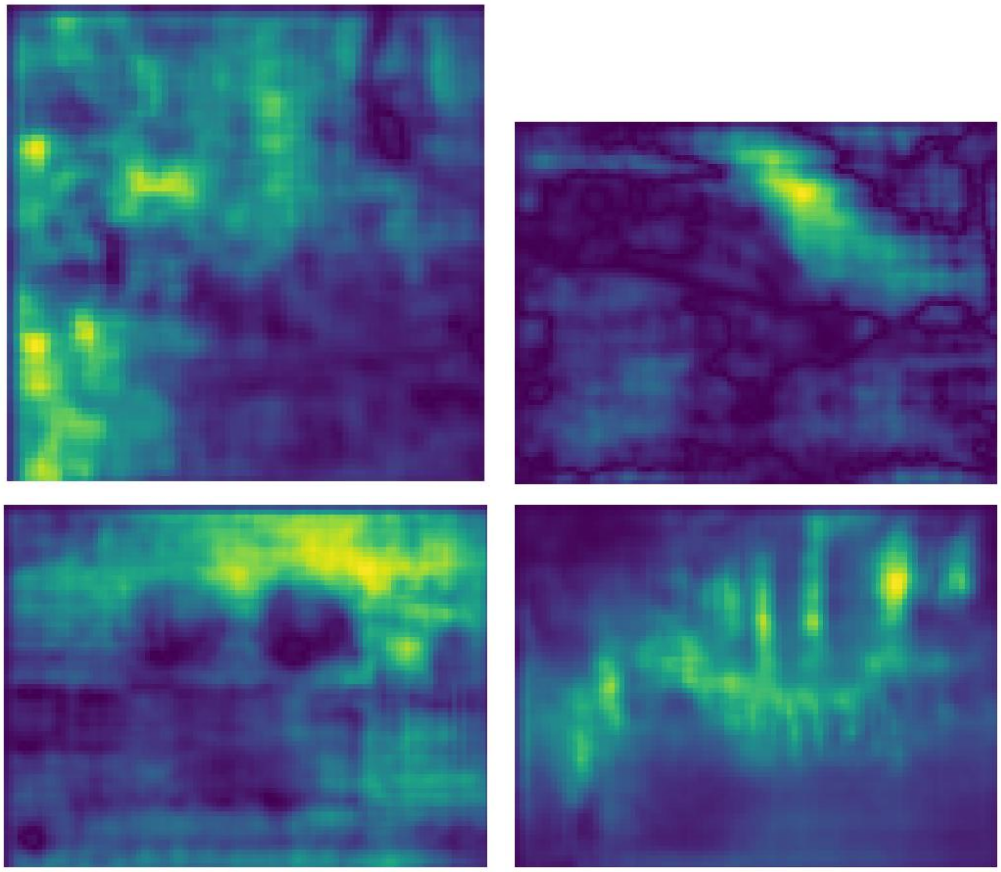


Figure A.8: *Attribution maps corresponding to four images in Figure 4.12 to the quality prediction by the SPAQ model.*

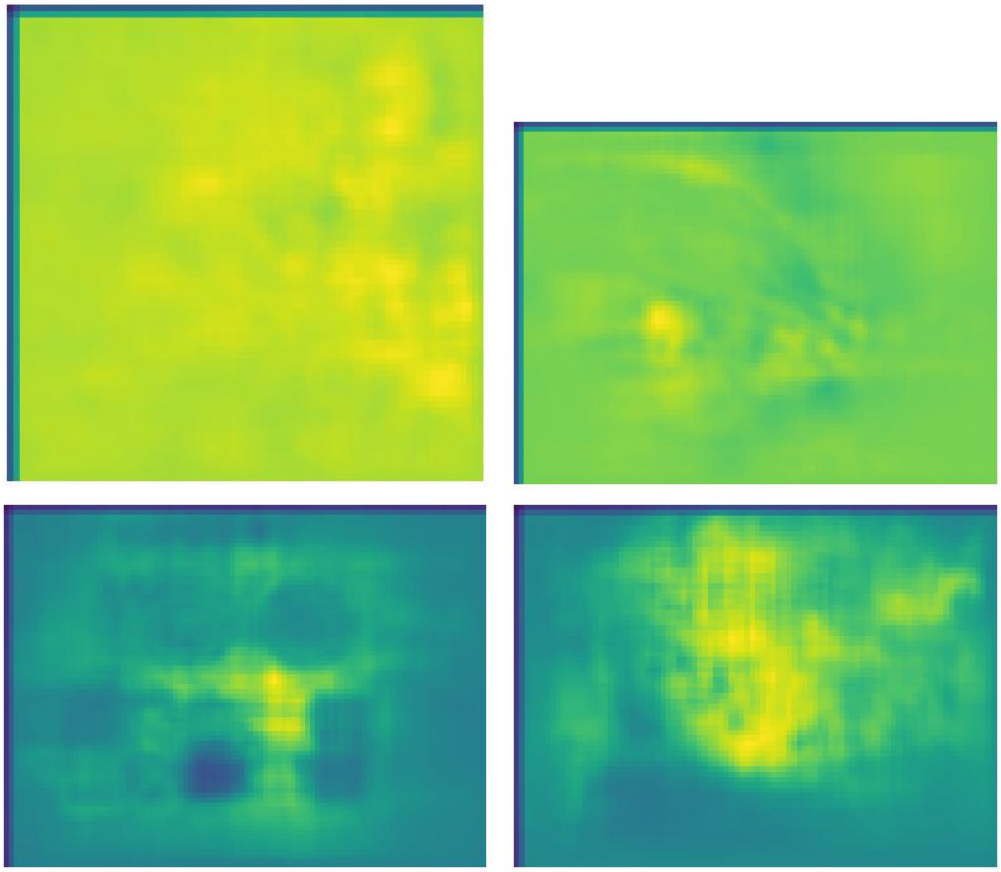


Figure A.9: *Attribution maps corresponding to four images in Figure 4.12 to the quality prediction by the KonIQ model.*

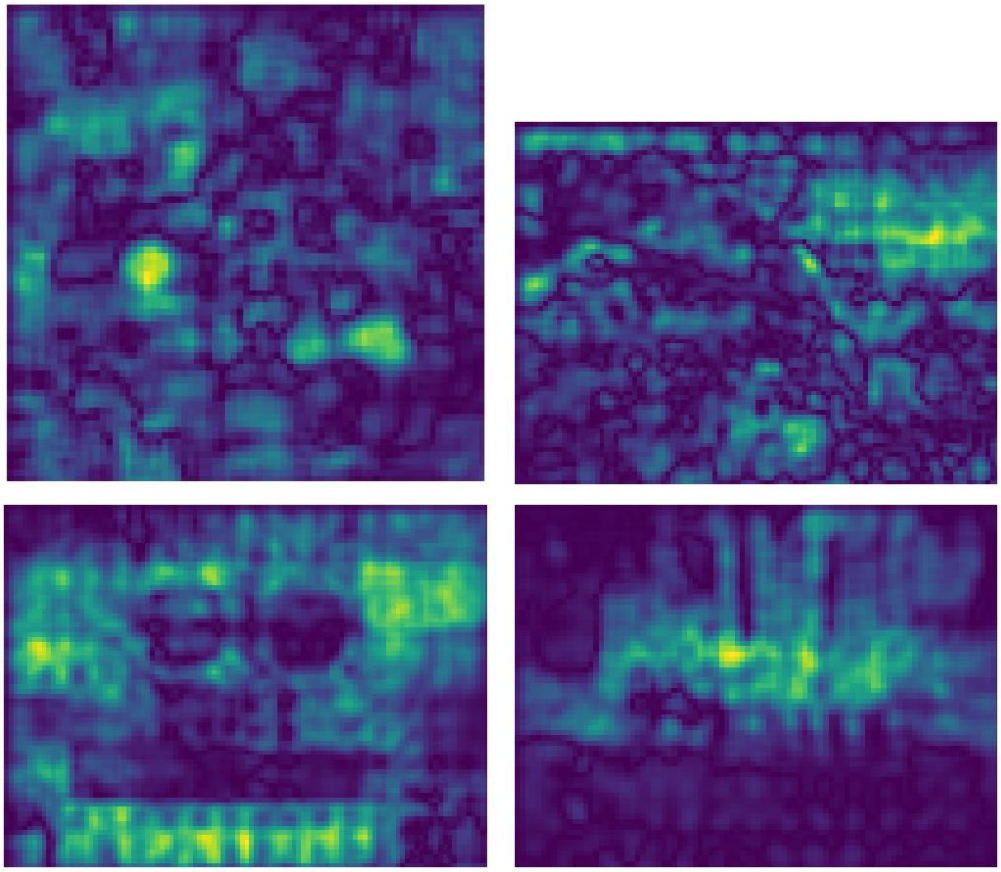


Figure A.10: *Attribution maps corresponding to four images in Figure 4.12 to the quality prediction by the MUSIQ model.*

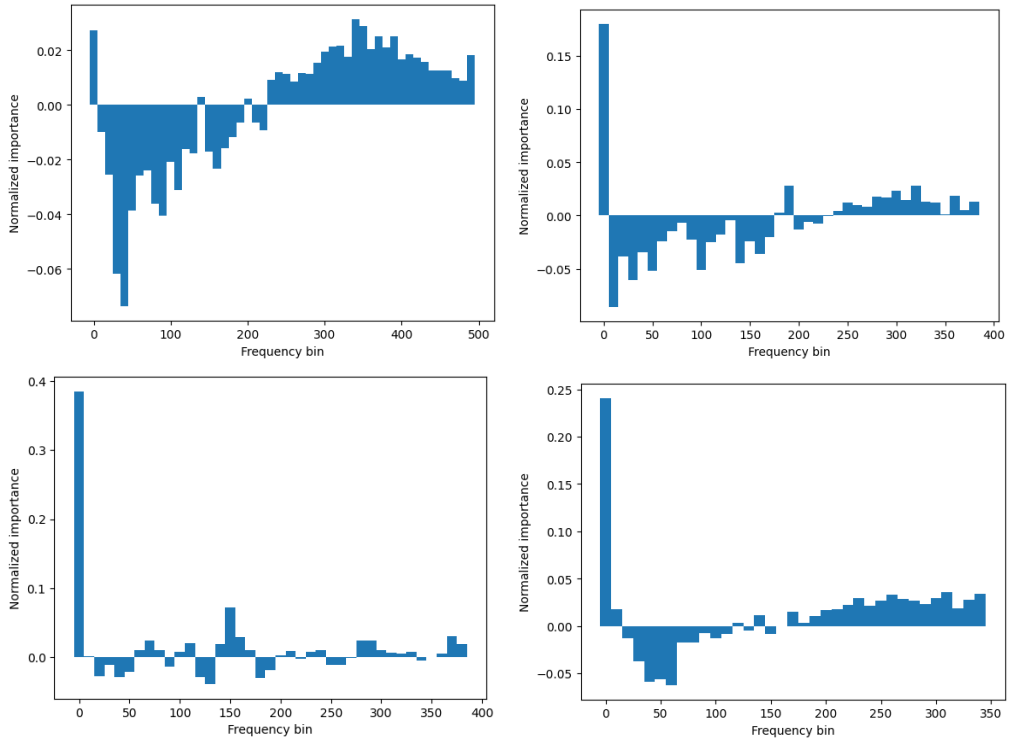
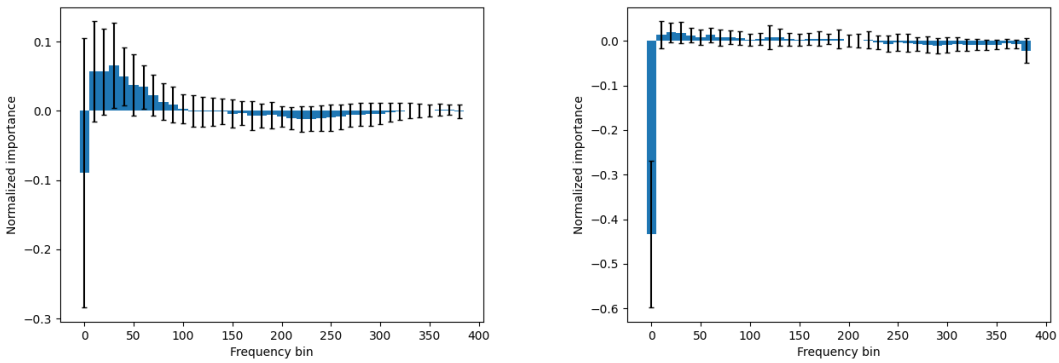


Figure A.11: *Contribution of data in each frequency bands of the Gaussian blurred images to the estimated quality prediction by the CNNIQA model.*



(a) *Gaussian blur* **(b)** *High frequency noise*

Figure A.12: *Importance of data in each frequency level of two types of distorted images to the estimated quality prediction by the CNNIQA. The error bar indicates the variation of importance values.*

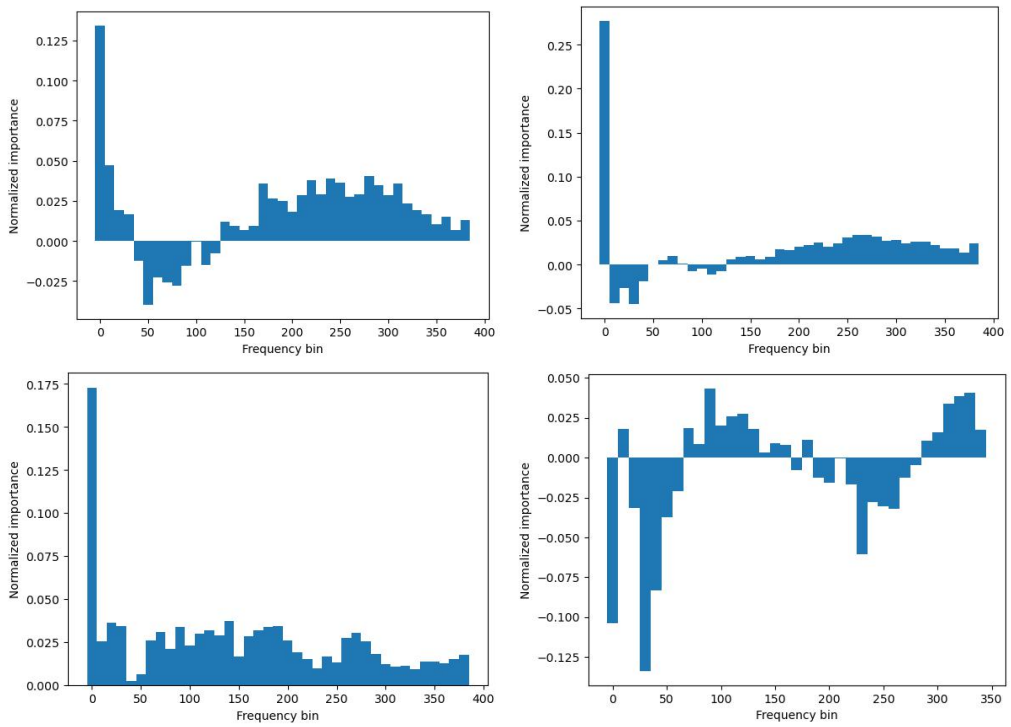
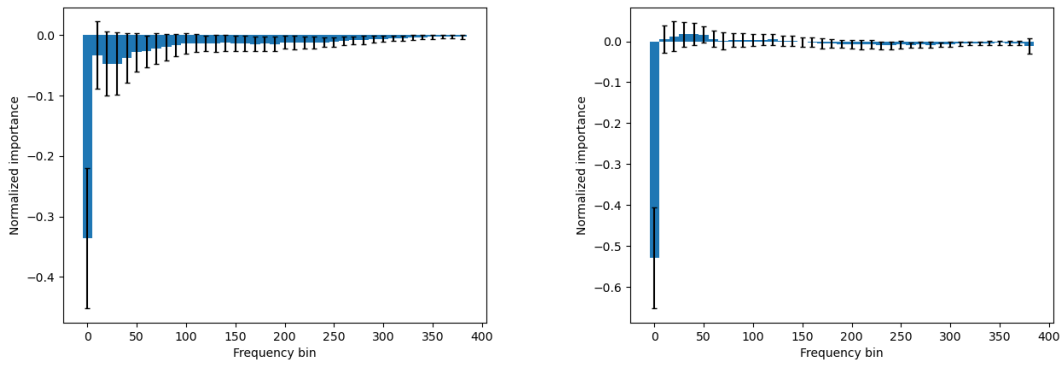


Figure A.13: *Importance of data in each frequency level of the Gaussian blurred images to the estimated quality prediction by the SPAQ model.*



(a) *Gaussian blur* **(b)** *High frequency noise*

Figure A.14: *Contribution of data in each frequency level of two types of distorted images to the estimated quality prediction by the SPAQ. The error bar indicates the standard deviation.*

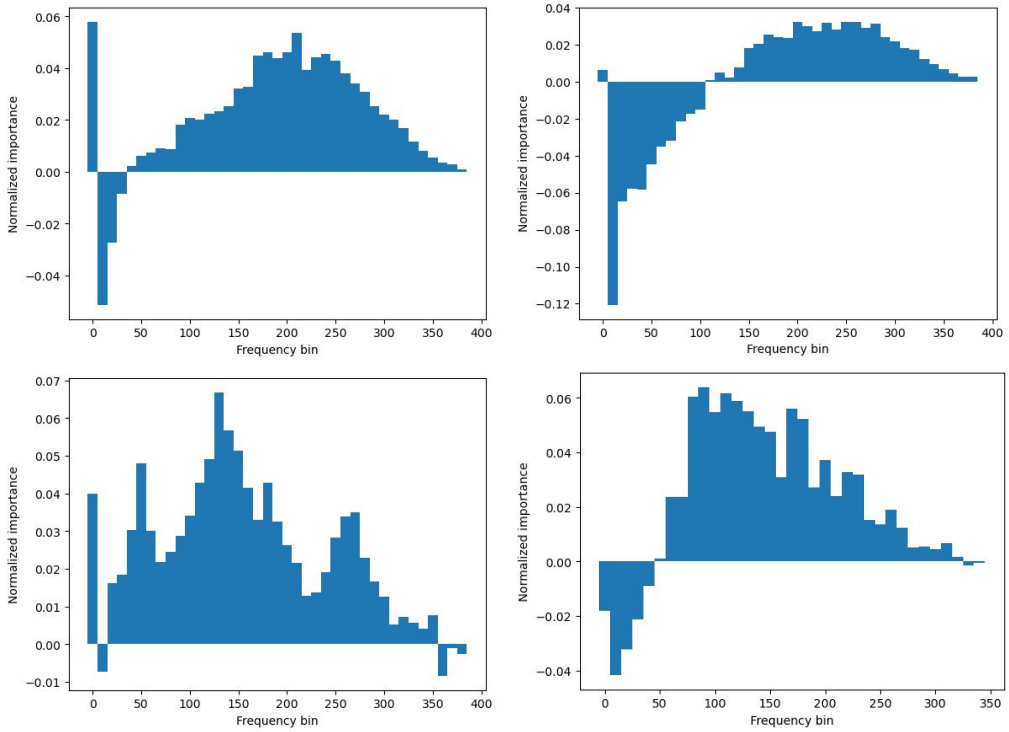
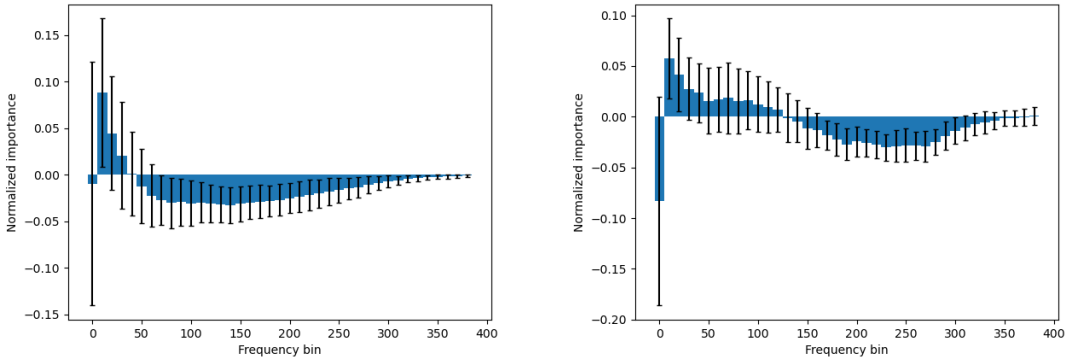


Figure A.15: *Importance of data in each frequency level of the Gaussian blurred images to the estimated quality prediction by the SPAQ model.*



(a) *Gaussian blur* **(b)** *High frequency noise*

Figure A.16: *Contribution of data in each frequency level of two types of distorted images to the estimated quality prediction by the KONIQ. The error bar indicates the standard deviation.*

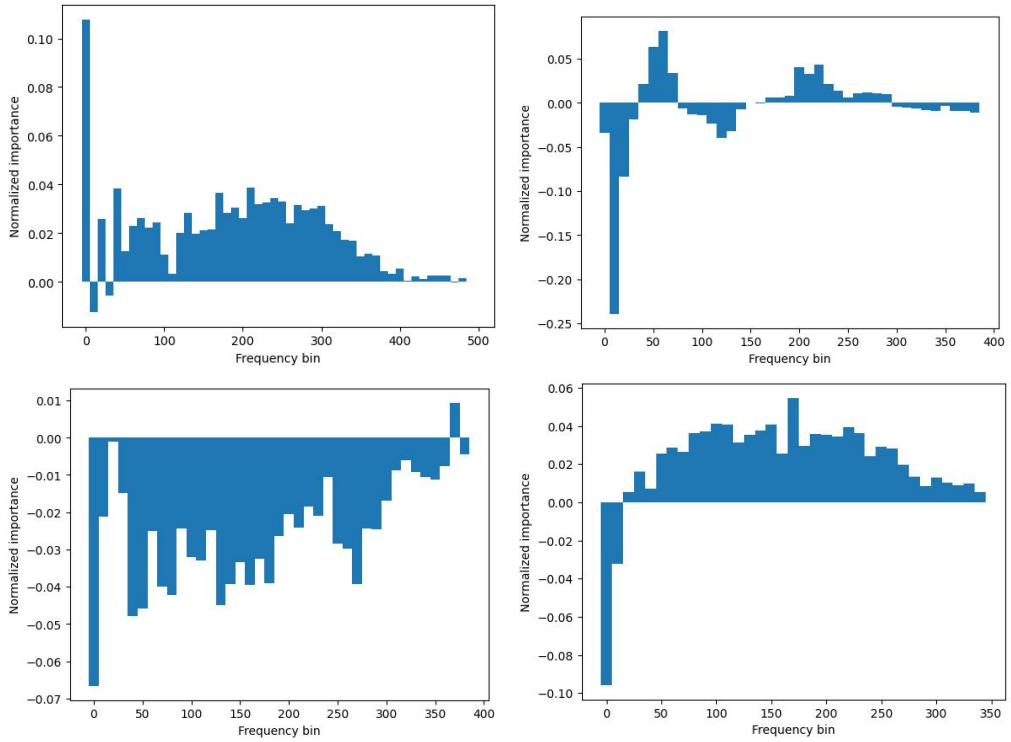
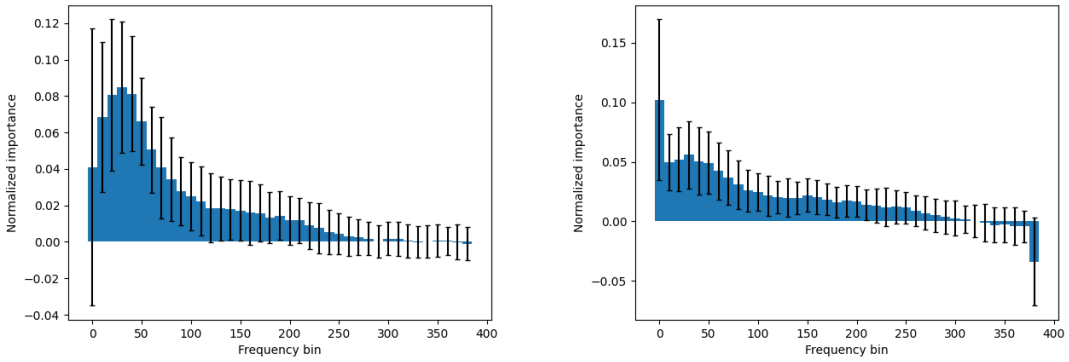
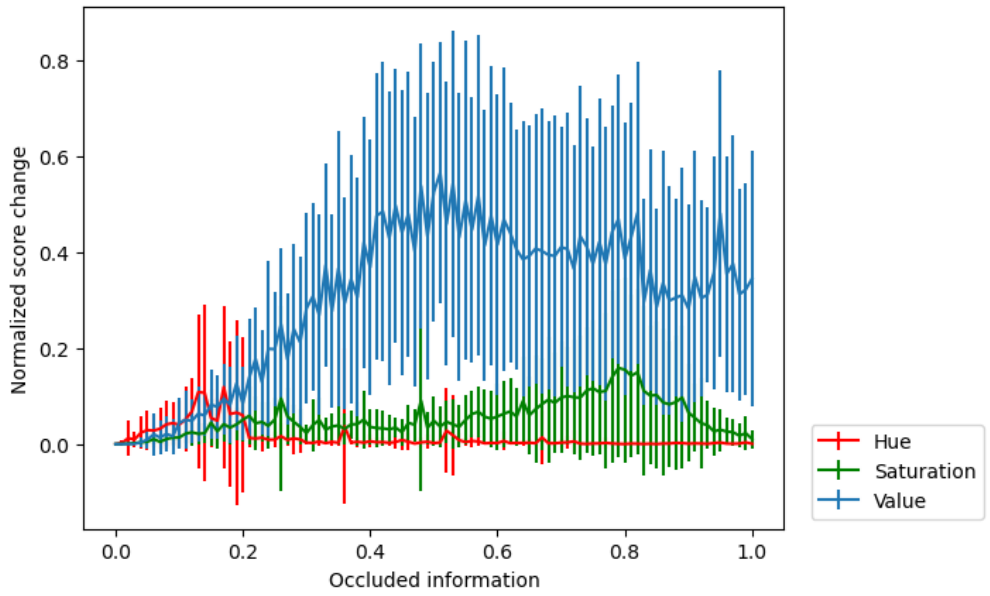


Figure A.17: *Contribution of data in each frequency level of the Gaussian blurred images to the estimated quality prediction by the MUSIQ model.*

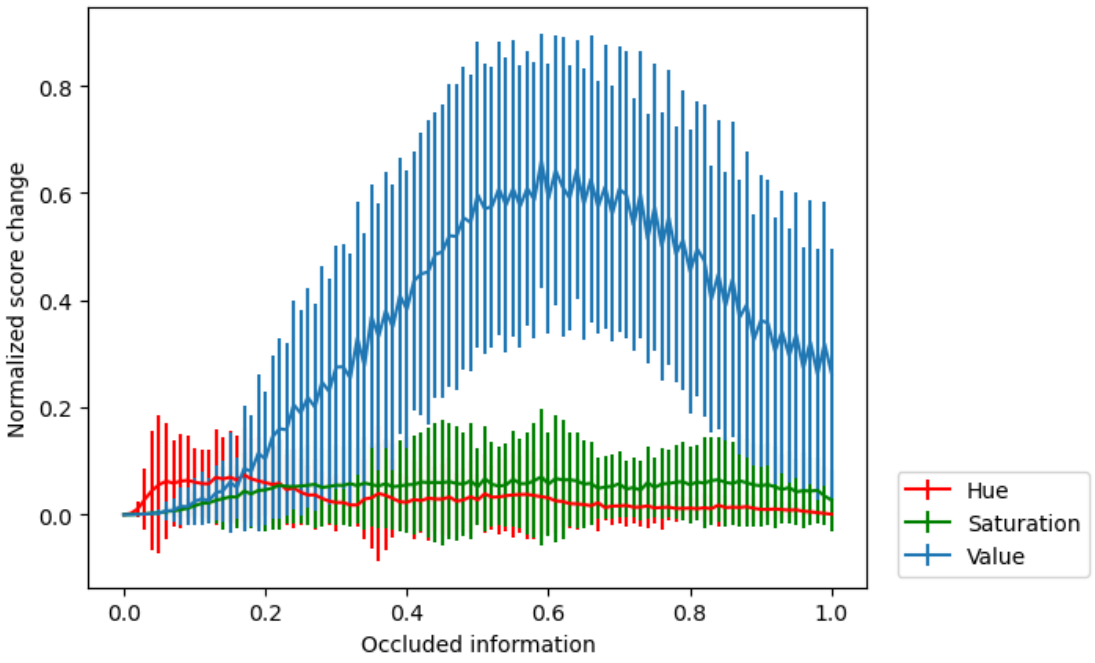


(a) *Gaussian blur* **(b)** *High frequency noise*

Figure A.18: *Contribution of data in each frequency level of two types of distorted images to the estimated quality prediction by the MUSIQ. The error bar indicates the standard deviation.*

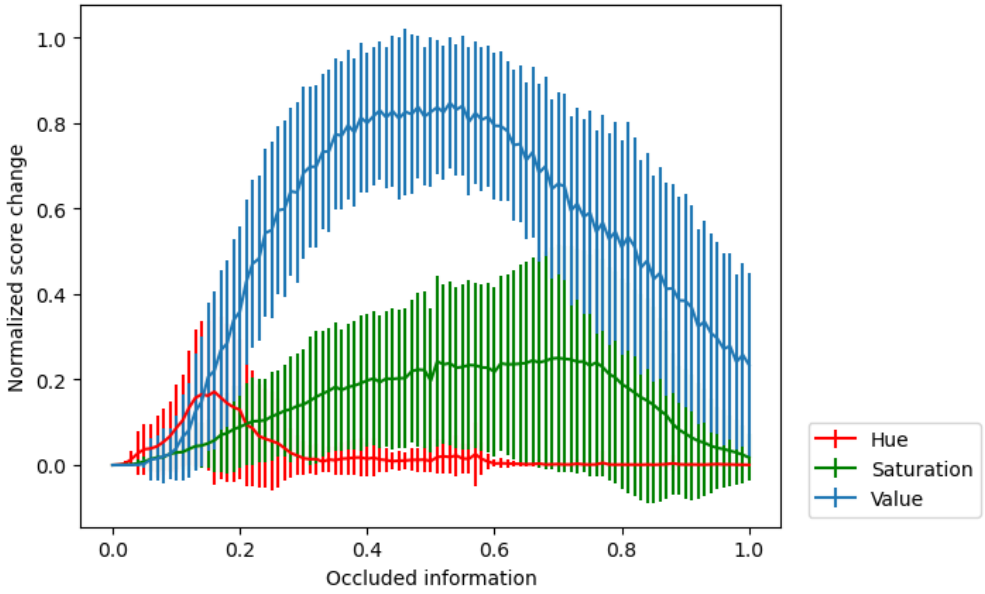


(a) *CNNIQA model*

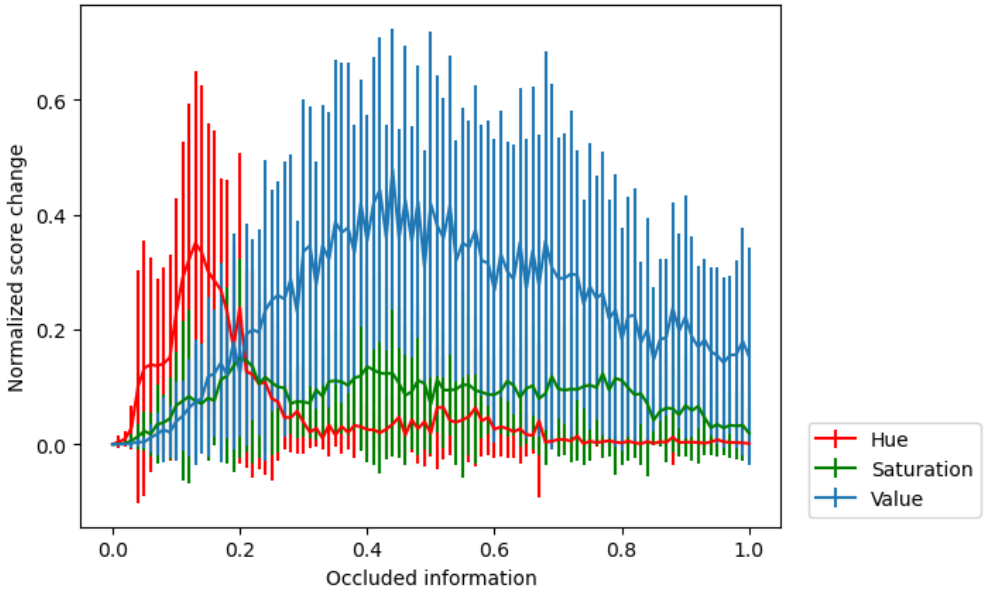


(b) *KonIQ model*

Figure A.19: *The trend of estimated quality score change when perturbing hue, saturation and values.*

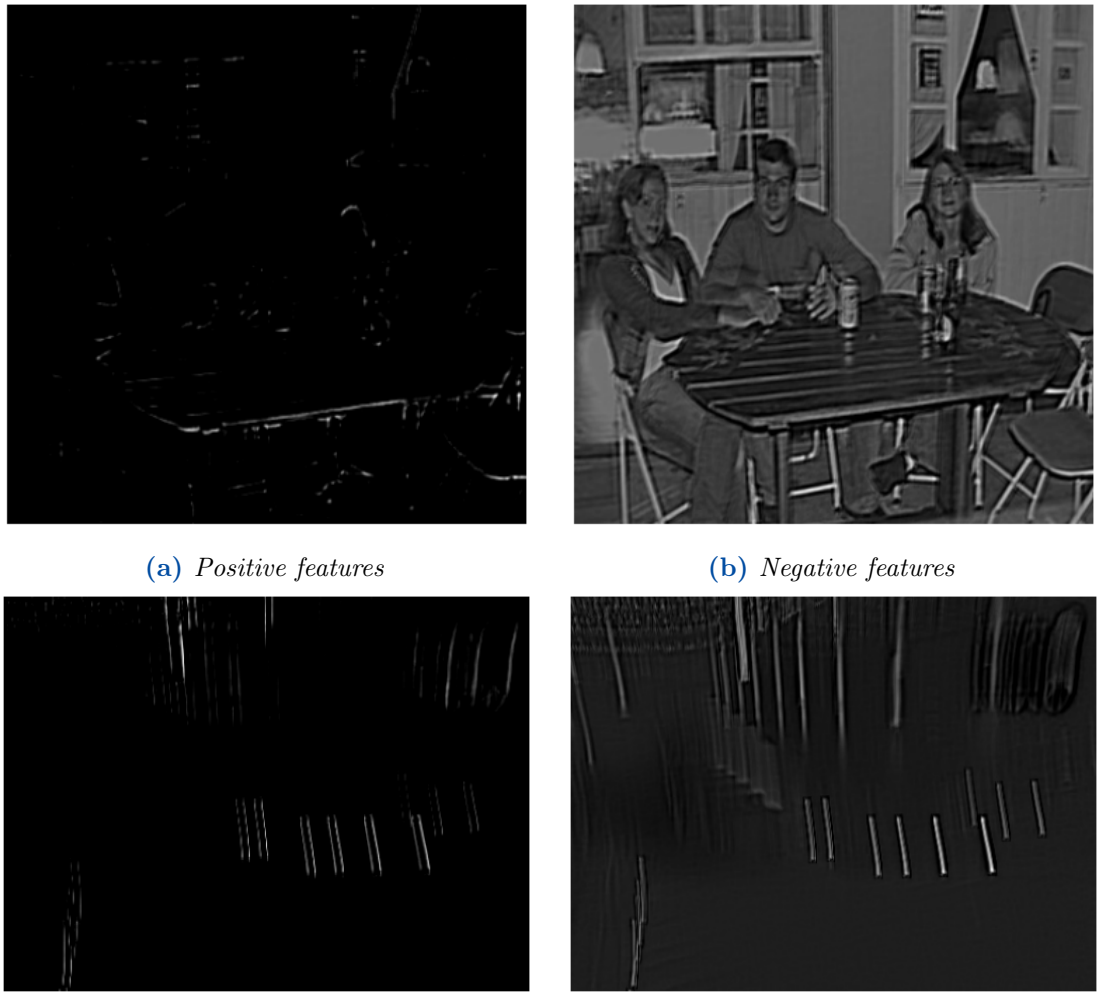


(a) *SPAQ model*



(b) *MUSIQ model*

Figure A.20: *The trend of estimated quality score change when perturbing hue, saturation and values.*



(a) *Positive features* (b) *Negative features*

(c) *Positive features* (d) *Negative features*

Figure A.21: *The positive and negative features visualized by Grad-CAM on CNNIQA model.*

Bibliography

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160. (cited on pages 18 and 20)

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424. (cited on page 18)

Athar, S. and Wang, Z. (2019). A comprehensive performance evaluation of image quality assessment algorithms. *Ieee Access*, 7:140030–140070. (cited on page 3)

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140. (cited on page 23)

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549. (cited on page 19)

Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, page 39. (cited on page 18)

Blog, L. (2020). Deep learning: Guided backpropagation. (cited on pages 43, 44, and 102)

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2017). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219. (cited on pages 9, 10, and 101)

Bull, D. R. and Zhang, F. (2021). Measuring and managing picture quality. *Intelligent Image and Video Compression*, pages 335–384. (cited on page 33)

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE. (cited on page 22)

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee. (cited on page 10)

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. (cited on page 12)

Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. (2020). Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686. (cited on pages 11 and 45)

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395. (cited on page 32)

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437. (cited on page 25)

Ghadiyaram, D. and Bovik, A. C. (2015). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387. (cited on pages 3 and 15)

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. (2022). No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230. (cited on pages 11 and 12)

Hautière, N., Tarel, J.-P., and Brémond, R. (2007). Perceptual hysteresis thresholding: Towards driver visibility descriptors. In *2007 IEEE International Conference on Intelligent Computer Communication and Processing*, pages 89–96. IEEE. (cited on pages 6 and 101)

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. (cited on page 11)

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. (cited on page 11)

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. (2020). Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056. (cited on pages 3, 11, 12, 15, 45, 46, and 101)

Joshi, K. (2021). Arya-xai - a distinctive approach to explainable ai. (cited on pages 19 and 101)

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740. (cited on pages 8, 9, 45, 46, and 101)

Kang, L., Ye, P., Li, Y., and Doermann, D. (2015). Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE international conference on image processing (ICIP)*, pages 2791–2795. IEEE. (cited on page 8)

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. (2021). Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157. (cited on pages 12, 13, 45, and 101)

Kim, J. and Lee, S. (2016). Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220. (cited on page 8)

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., and Dähne, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*. (cited on page 23)

Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., and Lapuschkin, S. (2020). Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. (cited on page 23)

Krasula, L. and Le Callet, P. (2018). Emerging science of qoe in multimedia applications: Concepts, experimental guidelines, and validation of models. *Academic Press Library in Signal Processing, Volume 6*, pages 163–209. (cited on page 33)

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., and Klambauer, G. (2019). Neuralhydrology–interpreting lstms in hydrology. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 347–362. (cited on page 26)

Larson, E. C. and Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006. (cited on pages 9, 14, 47, and 102)

Legge, G. E. (1981). A power law for contrast discrimination. *Vision research*, 21(4):457–467. (cited on page 5)

Letzgus, S., Wagner, P., Lederer, J., Samek, W., Müller, K.-R., and Montavon, G. (2022). Toward explainable artificial intelligence for regression models: A methodological perspective. *IEEE Signal Processing Magazine*, 39(4):40–58. (cited on page 25)

Li, Q., Lin, W., and Fang, Y. (2016a). No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Processing Letters*, 23(4):541–545. (cited on pages 8 and 12)

Li, Q., Lin, W., Xu, J., and Fang, Y. (2016b). Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia*, 18(12):2457–2469. (cited on page 8)

Lin, H., Hosu, V., and Saupe, D. (2019). Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE. (cited on pages 3, 14, and 64)

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18. (cited on page 18)

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. (cited on page 24)

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. (2017). End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213. (cited on page 8)

Madhuri, G. and Bindu, C. H. (2015). Performance evaluation of multi-focus image fusion techniques. In *2015 International Conference on Computing and Network Communications (CoCoNet)*, pages 248–254. IEEE. (cited on pages 40 and 102)

Marziliano, P., Dufaux, F., Winkler, S., and Ebrahimi, T. (2002). A no-reference perceptual blur metric. In *Proceedings. International conference on image processing*, volume 3, pages III–III. IEEE. (cited on pages 7 and 12)

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012a). No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708. (cited on page 7)

Mittal, A., Soundararajan, R., and Bovik, A. C. (2012b). Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212. (cited on page 7)

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15. (cited on page 23)

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080. (cited on page 18)

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2022). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*. (cited on page 18)

Papadopoulos, S. and Kontokosta, C. E. (2019). Grading buildings on energy performance using city benchmarking data. *Applied Energy*, 233:244–253. (cited on page 26)

Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. (2013). Color image database tid2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*, pages 106–111. IEEE. (cited on pages 9, 14, and 64)

Poynton, C. (1997). Frequently asked questions about color. *Retrieved June*, 19(449):2004. (cited on page 6)

Prabhushankar, M., Kwon, G., Temel, D., and AlRegib, G. (2020). Contrastive explanations in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3289–3293. IEEE. (cited on pages 26, 27, and 101)

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. (cited on page 24)

Saad, M. A., Bovik, A. C., and Charrier, C. (2012). Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352. (cited on pages 6 and 7)

Saleem, R., Yuan, B., Kurugollu, F., Anjum, A., and Liu, L. (2022). Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*. (cited on page 20)

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278. (cited on pages 18 and 22)

Schöttl, A. (2022). Improving the interpretability of gradcams in deep classification networks. *Procedia Computer Science*, 200:620–628. (cited on page 66)

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. (cited on pages 22, 23, 42, 66, and 101)

Sharma, A., Bhosle, A., and Chaudhary, B. (2012). Consumer perception and attitude towards the visual elements in social campaign advertisement. *IOSR Journal of Business and Management (IOSRJBM)*, 3(1):6–17. (cited on page 1)

Sheikh, H. R., Bovik, A. C., and Cormack, L. (2005). No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on image processing*, 14(11):1918–1927. (cited on page 7)

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451. (cited on pages 9, 10, 14, and 16)

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR. (cited on page 22)

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. (cited on pages 20, 21, and 101)

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. (cited on pages 9 and 10)

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*. (cited on page 22)

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (xai) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250. (cited on page 18)

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*. (cited on pages 21, 43, and 66)

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. (2020). Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676. (cited on page 10)

Sun, Q., Huang, F.-C., Wei, L.-Y., Luebke, D., Kaufman, A., and Kim, J. (2020). Eccentricity effects on blur and depth perception. *Optics express*, 28(5):6734–6739. (cited on page 55)

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR. (cited on page 21)

Tamaddon-Jahromi, H. R., Chakshu, N. K., Sazonov, I., Evans, L. M., Thomas, H., and Nithiarasu, P. (2020). Data-driven inverse modelling through neural network (deep learning) and computational heat transfer. *Computer Methods in Applied Mechanics and Engineering*, 369:113217. (cited on page 26)

Toet, A. and Lucassen, M. P. (2003). A new universal colour image fidelity metric. *Displays*, 24(4-5):197–207. (cited on page 5)

Unterweger, A. (2013). Compression artifacts in modern video coding and state-of-the-art means of compensation. In *Multimedia Networking and Coding*, pages 28–49. IGI Global. (cited on page 69)

Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470. (cited on page 13)

Wang, Z. and Bovik, A. C. (2002). A universal image quality index. *IEEE signal processing letters*, 9(3):81–84. (cited on page 5)

Wang, Z. and Bovik, A. C. (2011). Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40. (cited on page 7)

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612. (cited on page 5)

Wang, Z., Sheikh, H. R., and Bovik, A. C. (2002). No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings. International conference on image processing*, volume 1, pages I–I. IEEE. (cited on page 7)

Wu, Q., Wang, Z., and Li, H. (2015). A highly efficient method for blind image quality assessment. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 339–343. IEEE. (cited on page 7)

Xue, W., Zhang, L., and Mou, X. (2013). Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 995–1002. (cited on page 7)

Yang, S., Jiang, Q., Lin, W., and Wang, Y. (2019). Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1383–1391. (cited on page 9)

Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., and Yang, Y. (2022). Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200. (cited on pages 12, 13, and 101)

Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. (2020). From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585. (cited on page 11)

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer. (cited on pages 21, 25, and 35)

Zhang, L., Zhang, L., and Bovik, A. C. (2015). A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591. (cited on page 7)

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386. (cited on page 8)

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. (2018). Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47. (cited on pages 10, 45, 46, and 101)

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929. (cited on page 22)

Zhou, L., Deng, W., and Wu, X. (2019). Robust image segmentation quality assessment. *arXiv preprint arXiv:1903.08773*. (cited on page 3)

Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. (2020). Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152. (cited on page 1)

Zhu, M., Hou, G., Chen, X., Xie, J., Lu, H., and Che, J. (2021). Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1953–1962. (cited on page 12)

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*. (cited on page 25)

BIBLIOGRAPHY

List of Figures

List of Tables