Abou Bakr Belkaid Tlemcen University
Internship at XLIM UMR CNRS 7252,
Poitiers, France

Master Thesis
Field: Artificial Intelligence / Computer Science

A Benchmark Study on Image Quality Assessment and Interpretability in Super-Resolution

---

Submitted by
Mohammed Seyf Elislem LASFER

Supervisor: Houcine MATALLAH
Co-supervisor: Chaker LARABI
Co-supervisor: Abderrezzaq SENDJASNI

President: Sidi Mohammed CHOUITI
Examiner: Sid Ahmed BERRABAH

Defense in Tlemcen, October 2025

Dedication

I dedicate this thesis to my parents, whose love, support, and sacrifices have been my source of inspiration and perseverance.

To my family, who have always encouraged me with kindness.

To my teachers, for their guidance and valuable advice throughout my academic journey.

To all those who, directly or indirectly, contributed to the completion of this work, I express my deep gratitude.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to ALLAH for granting me the strength, patience, and courage to accomplish this modest work.

I would like to extend my sincere thanks and profound appreciation to Professor Chaker Larabi and Dr. Abderrezzaq Sendjasni for supervising my internship at the XLIM laboratory (UMR CNRS 7252). I am truly grateful for their scientific guidance, valuable advice, availability, and encouragement throughout this project. I also wish to thank Mr. Matallah Houcine, my academic supervisor at the university, for his follow-up, support, and constructive feedback during the preparation of this thesis.

My special thanks go to the XLIM laboratory (UMR CNRS 7252) for hosting me and providing an excellent research environment during my final year project.

I would also like to express my heartfelt appreciation to the jury members, the honorable President Professor Sidi Mohammed CHOUITI, and the honorable Examiner Professor Sid Ahmed BERRABAH, for the interest they have shown in my work by agreeing to evaluate this thesis and enrich it with their insightful remarks.

Finally, I am equally thankful to all the teachers who supported me throughout my studies.

Abstract

Super-resolution (SR) is a vital technique in fields such as medical imaging, surveillance, and satellite observation. However, evaluating the quality of super- resolved images remains challenging, especially without reference images. This thesis focuses on No-Reference Image Quality Assessment (NR-IQA) for SR outputs, providing a detailed review of existing methods, their principles, strengths, limitations, and real-world applications. Additionally, the integration of Explainable Artificial Intelligence (XAI) offers insights into the decision-making processes of SR models, enhancing their transparency and reliability. This work contributes to the development of more effective and interpretable SR systems.

Keywords: Super-resolution, No-reference image quality assessment (NR-IQA), Image enhancement, Explainable Artificial Intelligence (XAI), Image quality evaluation, Deep learning, Interpretable models.

Résumé

La super-résolution (SR) d'image est une technique essentielle dans divers domaines comme l'imagerie médicale, la surveillance et l'observation satellite. Cependant, évaluer la qualité des images générées reste un défi, surtout en l'absence d'images de référence. Cette mémoire s'intéresse à l'évaluation sans référence (No- Reference Image Quality Assessment, NR-IQA) des images super-résolues. Elle propose une étude approfondie des méthodes existantes, en analysant leurs principes, avantages, limites et applications. Par ailleurs, l'intégration de l'Intelligence Artificielle Explicable (XAI) permet de mieux comprendre les décisions des modèles SR, offrant une meilleure transparence et fiabilité. Ce travail apporte des pistes pour le développement de systèmes SR plus performants et interprétables.

Mots-clés : Super-résolution, Évaluation de la qualité d'image sans référence (NR-IQA), Amélioration d'image, Intelligence artificielle explicable (XAI), Évaluation de la qualité d'image, Apprentissage profond, Modèles interprétables.

ملخص

تعد تقنية تحسين دقة الصور من الأدوات الأساسية في مجالات مثل التصوير الطبي، والمراقبة، والاستشعار عن بُعد. ومع ذلك، لا تزال عملية تقييم جودة الصور المحسّنة تمثل تحدياًا في غياب الصور المرجعية. تركز هذه الأطروحة على تقييم جودة الصور دون مرجع ، من خلال دراسة شاملة للأساليب الحالية، وتحليل مبادئها، ومزاياها، وقيودها، وتطبيقاتها العملية. كما تسلط الضوء على أهمية الذكاء الاصطناعي القابل للتفسير لفهم آليات عمل نماذج التحسين، مما يعزز الشفافية والمصداقية. وتسهم هذه الدراسة في تطوير أنظمة ضض أكثر فعالية وقابلية للتفسير.

الكلمات المفتاحية: تحسين دقة الصور، تقييم جودة الصور دون مرجع ، تحسين الصور، الذكاء الاصطناعي القابل للتفسير ، تقييم جودة الصورة، التعلم العميق، النماذج القابلة للتفسير.

Contents

List of Figures

List of Tables

General Introduction

Context and Motivation

With the rapid evolution of digital technologies, large amounts of data are generated quickly and are used everywhere. High-quality visual information is now essential in many fields such as medical diagnosis, satellite monitoring, smart transportation, and entertainment. These areas increasingly rely on clear and accurate images to support precise analysis and effective decision-making.

However, due to physical constraints of image acquisition systems (e.g., sensor limitations), as well as practical constraints such as bandwidth, storage, and computational efficiency, low-resolution (LR) images are frequently encountered. These LR images often lack fine details and textures, which can hinder downstream processing and human interpretation.

Image Super-Resolution (SR) seeks to overcome this limitation by reconstructing high-resolution (HR) images from one or more LR inputs through computational means. Recent advances in deep learning have led to powerful SR models, including convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformer-based architectures, that are capable of generating photorealistic, perceptually enhanced images.

Yet, with the growing sophistication of SR methods comes a major challenge: how to reliably and meaningfully evaluate the quality of the super-resolved images. The most commonly used evaluation metrics—such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [8]—are part of the Full-Reference IQA (FR-IQA) category, which means they require access to the original high-quality ground truth image. This requirement is unrealistic in many real-world situations, particularly in medical imaging, remote sensing, or historical photo enhancement, where ground truth is

unavailable or subjective.

This shortcoming has led to increased interest in No-Reference IQA (NR-IQA) models, which aim to assess the visual quality of an image without requiring a reference. NR-IQA models are particularly important in the context of SR, as they allow for autonomous quality evaluation in practical deployment scenarios. However, the reliability, generalization ability, and interpretability of NR-IQA models in SR applications remain open questions.

In parallel, as SR models become more complex and less transparent, the issue of explainability has gained prominence. These models are often used as black boxes, and their internal decision-making processes are difficult to interpret. For practical and ethical reasons—including user trust, compliance, and scientific insight—it is increasingly important to make the behavior of these models explainable and interpretable. This has given rise to Explainable AI (XAI) techniques, which aim to provide interpretable explanations for the outputs of complex models.

Combining SR, IQA, and XAI therefore constitutes a critical and timely research area, as it not only advances the development of high-quality image reconstruction techniques but also ensures objective evaluation of their performance and provides interpretability for trustworthy deployment in real-world applications.

Research Problem

Despite the rapid development of both super-resolution and no-reference quality assessment methods, significant gaps remain in how these approaches are evaluated and understood. This thesis addresses the following core research questions:

Evaluation reliability: Which NR-IQA models are most effective at assessing the perceptual quality of super-resolved images? How do they correlate with human subjective perception?

Dataset generalization: How do these models perform across different types of SR models and datasets? Are they robust to variations in texture, content, and artifacts?

Artifact awareness: What types of visual artifacts are introduced by different SR methods (e.g., checkerboard effects, over-smoothing, hallucinations), and how do NR-IQA models respond to them?

Interpretability: Can we use XAI techniques to understand how SR models reconstruct details? What regions of the input images are the models focusing on during the reconstruction process?

Trust and transparency: To what extent can the integration of interpretability tools improve our understanding and trust in both SR and IQA models?

Answering these questions is crucial to advancing the field of image quality assessment and enhancing the deployment of SR systems in critical and real-world settings.

Thesis Objectives

This thesis aims to address the above research questions through a structured and multi-faceted approach. The main objectives are:

- To establish a comprehensive benchmark of multiple NR-IQA models on super-resolved images generated by diverse SR approaches.
- To evaluate the performance of NR-IQA models by measuring their correlation with human judgments (e.g., MOS) through standard metrics such as SROCC and PLCC.
- To explore the application of state-of-the-art XAI methods for visualizing and interpreting the behavior of different SR models.
- To critically assess the strengths and limitations of interpretability tools in the context of super-resolution.

Expected Contributions

The anticipated contributions of this research include:

- A comprehensive benchmark of NR-IQA models applied specifically to super-resolution, providing guidance on their relative strengths and limitations.
- A curated taxonomy of SR artifacts, along with an analysis of how they affect perceived image quality and NR-IQA scores.
- Application and evaluation of XAI methods to explain and visualize the behavior of SR models, yielding novel insights into how these models reconstruct visual detail.
- A critical reflection on the role of interpretability in image enhancement tasks, including recommendations for future research in integrating XAI into the SR development pipeline.

Document Structure

The document is organized into three main chapters:

- **Chapter 1: State of the Art and Theoretical Foundations**

  This chapter introduces the background and literature on image super-resolution, image quality assessment (FR and NR), and explainable artificial intelligence.

- **Chapter 2: Methodology**

  This chapter describes the experimental protocol, including the selection of SR and IQA models, dataset preparation, artifact classification, and explainability tools.

- **Chapter 3: Experiments and Results**

  This chapter is dedicated to the evaluation and critical analysis of super-resolution models. It presents the benchmarking results, correlation analyses, and visualizations of attention maps and attribution regions, providing quantitative and qualitative insights into model performance. Furthermore, it interprets these findings, discusses the limitations and practical implications of NR-IQA metrics and model explainability, summarizes the key contributions, and outlines future research directions in super-resolution evaluation and interpretable AI.

Chapter 1

State of the Art and Theoretical Foundations

Contents

1.1. Image Super-Resolution (SR)

1.1.1. Introduction to Super-Resolution Concepts

Super-resolution (SR) refers to the process of generating high-resolution (HR) images from one or more low-resolution (LR) inputs. Over the years, it has found applications across a broad spectrum of domains, including satellite and aerial imaging [9], medical imaging [10], ultrasound imaging, line fitting, automated mosaicking, infrared imaging, facial and text image enhancement [11], compressed image and video restoration, sign and license plate recognition [12], iris and fingerprint recognition [13], digital holography, and high-dynamic-range (HDR) imaging. Fundamentally, SR algorithms aim to recover fine image details beyond the sampling limitations of the original acquisition device by effectively increasing the number of pixels per unit area [14].

Figure 1.1: *Example of an image enhanced using a super-resolution algorithm.*

1.1.2. Classification of SR Methods

Super-resolution techniques are generally categorized into three major categories, each defined by its core principle, benefits, and limitations. In the following sections, we detail each category.

1.1.2.1. Interpolation-Based Methods

Interpolation techniques, such as those introduced by Lertrattanapanich et al. [15], Su et al. [16], and Zhao et al. [17], enhance resolution by estimating new pixel values based

on neighboring pixels in the image grid. Methods like nearest-neighbor, bilinear, bicubic, and Lanczos interpolation fall under this category. Their simplicity and low computational cost make them accessible and widely used in real-time applications. However, due to their simplistic assumptions, these methods often produce blurred outputs and lack the ability to reconstruct fine textures or sharp edges—particularly problematic when high magnification is needed [18]. For instance, in Figure 1.2, the bicubic interpolation result clearly illustrates this limitation.

1.1.2.2. Reconstruction-Based Methods

In contrast to interpolation-based methods, reconstruction ones such as those proposed by Irani et al. [19], Tekalp et al. [20], Patti et al. [21], and Liu et al. [22], frame super-resolution as an optimization problem. These approaches typically use transformation matrices or motion estimation to iteratively reconstruct a high-resolution image from multiple low-resolution (LR) inputs. Common techniques in this category include Iterative Back-Projection (IBP) [23], Projection Onto Convex Sets (POCS) [24], and regularized reconstruction methods such as Total Variation (TV) minimization [25]. Other notable approaches include MAP-based SR (Maximum a Posteriori estimation), as illustrated in Figure 1.2, and Bayesian methods that incorporate prior knowledge to stabilize the reconstruction. Additionally, motion-compensated SR methods align multiple LR frames to exploit sub-pixel shifts, improving detail recovery. These methods generally provide better preservation of structural and fine-grained information compared to interpolation-based techniques. However, they are computationally expensive, sensitive to motion estimation errors, and highly dependent on parameter tuning. Small variations in transformation assumptions can lead to significant changes in output quality, which limits their effectiveness in dynamic scenes or resource-constrained environments [18].

1.1.2.3. Deep Learning-Based Methods

The remarkable progress in super-resolution can be largely attributed to advances in artificial intelligence, and deep learning in particular. Specialized architectures such as Convolutional Neural Networks (CNNs) [26], as illustrated in Figure 1.2, have demonstrated outstanding performance in learning complex mappings from low-resolution to high-resolution images, Generative Adversarial Networks (GANs) [27], and, more recently, vision Transformers (ViT) [28] have significantly pushed the boundaries of performance. These models are designed to learn complex mappings from low-resolution to high-

resolution images by extracting hierarchical features and progressively refining them to reconstruct fine details.

One of the earliest deep learning models for SR, the SRCNN model proposed by Dong et al. [29], introduced a shallow CNN to directly learn the end-to-end mapping between LR and HR images. This was later improved by deeper architectures like EDSR (Enhanced Deep Super-Resolution) [30], which removed unnecessary layers such as batch normalization and increased model capacity to boost accuracy.

GAN-based models, such as ESRGAN (Enhanced Super-Resolution GAN) [31], introduced adversarial learning to enhance perceptual quality by encouraging the network to produce more photo-realistic and sharper textures, closer to human visual perception.

More recently, ViT architectures have emerged, leveraging self-attention mechanisms to capture long-range dependencies and contextual information. SwinIR (Swin Transformer for Image Restoration) [32] is a prominent example, combining the strengths of Swin Transformers with convolutional modules to achieve state-of-the-art results in both distortion- and perception-oriented SR tasks.

These deep models have demonstrated significant improvements over traditional and reconstruction-based approaches in terms of both quantitative accuracy (e.g., PSNR, SSIM) and perceptual quality. Moreover, their flexibility allows them to generalize well across diverse image domains. However, their impressive performance comes at the cost of high computational complexity, and they are often sensitive to architectural choices and hyperparameter tuning, which can substantially influence the quality and stability of the output [18].

Figure 1.2: *Illustration of three categories of image super-resolution: bicubic interpolation (blurry), MAP-based reconstruction (over-smoothed), and learning-based SRCNN (sharper and closer to HR).*

1.1.3. Artifacts, Fidelity–Perception Trade-off, and Generalization in Super-Resolution

Reconstructing a high-resolution image I_{HR} from a low-resolution input I_{LR} can be formulated as an inverse problem:

$$I_{LR} = D(I_{HR}) + \epsilon,$$

where $D(\cdot)$ represents a degradation operator (e.g., downsampling or blurring) and ϵ is noise. This problem is inherently ill-posed because multiple high-resolution images can correspond to the same low-resolution input. A common approach is to minimize the Mean Squared Error (MSE):

$$\hat{I}_{HR} = \arg\min_{I} \|D(I) - I_{LR}\|_2^2.$$

However, MSE encourages averaging over all possible solutions, which often suppresses high-frequency details, leading to overly smooth textures and blurry reconstructions. To better preserve fine details, one can add a high-frequency loss term:

$$\hat{I}_{HR} = \arg\min_{I} \|D(I) - I_{LR}\|_2^2 + \lambda \|F_{HF}(I) - F_{HF}(I_{HR})\|_2^2,$$

where F_{HF} extracts high-frequency components and λ balances overall fidelity and detail preservation.

On the flip side, methods based on perceptual loss or adversarial training—particularly those involving GANs—can introduce artificial details. While these details may enhance the visual appeal of images, they do not necessarily preserve the original content. This is clearly illustrated in Figure 1.3, where SRGAN generates textures that appear realistic but do not correspond to the actual structures in the original image. Such artifacts pose a significant issue in fields where accuracy is critical, such as medical imaging or surveillance. The generated textures may look plausible to the human visual system (HVS) but often have little to no correlation with the true image content [33, 31].

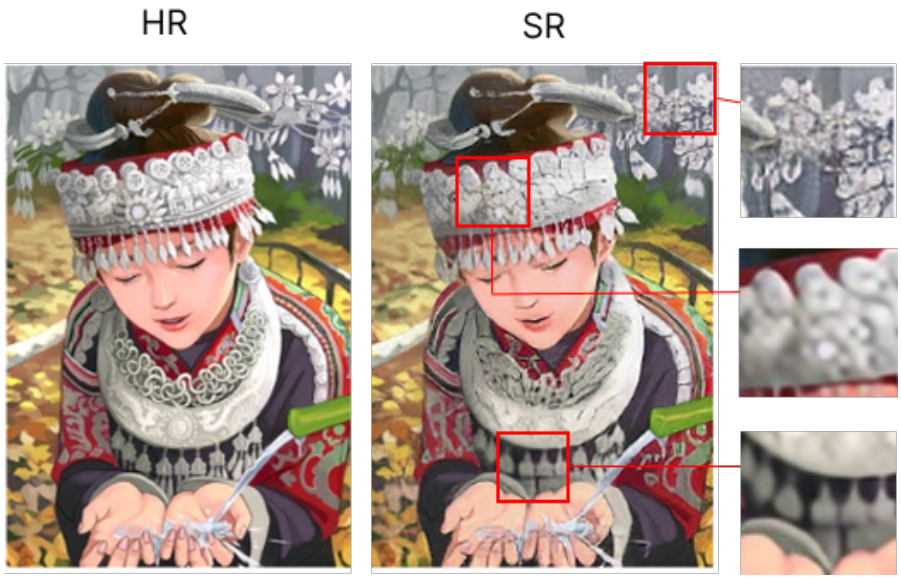


Figure 1.3: *Illustration of visual artifacts produced by the SRGAN method.*

Besides these, there are other common artifacts in super-resolution outputs. Oversharpening, caused by aggressive edge enhancement, can make contours appear unnaturally sharp or glowing. Ringing artifacts—wave-like distortions near strong edges—usually arise from overcorrected deblurring or compression. Aliasing effects, like jagged edges or *"moiré"* patterns, typically stem from improper sampling or interpolation, especially in areas rich in fine textures. Poor handling of lighting transitions can lead to inconsistencies in brightness and unnaturally sharp edges. It has been observed that using feature maps after activation for perceptual loss can cause luminance distortions, whereas using them before activation results in better brightness reconstruction [31]. Other issues, such as halo effects, blocking, banding, color bleeding, ghosting, and checkerboard patterns, are also common. These are often the result of poorly configured upsampling, compression artifacts, or deconvolution layers.

The network's architecture itself can also contribute to instability during training. For example, Batch Normalization layers have been found to introduce unpredictable artifacts in deep networks. Removing these layers has been shown to improve both stability and generalization [31]. On the other hand, shallow networks often struggle to preserve structural details, while very deep networks may produce incoherent or noisy textures if not carefully optimized—especially in uniform or repetitive regions [33, 31].

One of the core challenges in super-resolution is finding the right balance between fidelity (how accurate the reconstruction is) and perceptual realism (how good it looks to a human observer). Traditional loss functions like MSE tend to produce high PSNR

values but result in blurry or dull images. Perceptual and adversarial losses, by contrast, generate more visually pleasing images but can deviate from the ground truth. Striking the right balance is particularly important in applications that demand both accuracy and realism. Strategies such as using pre-activation feature maps in the perceptual loss, employing relativistic discriminators, and interpolating between different network variants during training have been effective in achieving this balance [31].

Another major concern is generalization. Many super-resolution models perform well on clean, synthetic datasets but fail when applied to real-world images with unknown or complex degradation patterns. This is especially problematic when the model is deployed outside its training conditions. Certain architectural components—like Batch Normalization—can hurt generalization by making the model too dependent on training-specific statistics. Removing them has shown to improve performance in more diverse settings [31]. Broader and more realistic training data, domain adaptation techniques, and self-supervised learning are promising directions for improving how well super-resolution models generalize in the wild. The figure 1.4 illustrates an example of the different artifacts that can occur due to super-resolution.

Figure 1.4: *Visual examples of typical artifacts introduced by image super-resolution methods, including blur, aliasing, ringing, oversharpening, banding, and other perceptual distortions.*

1.2. Image Quality Assessment (IQA)

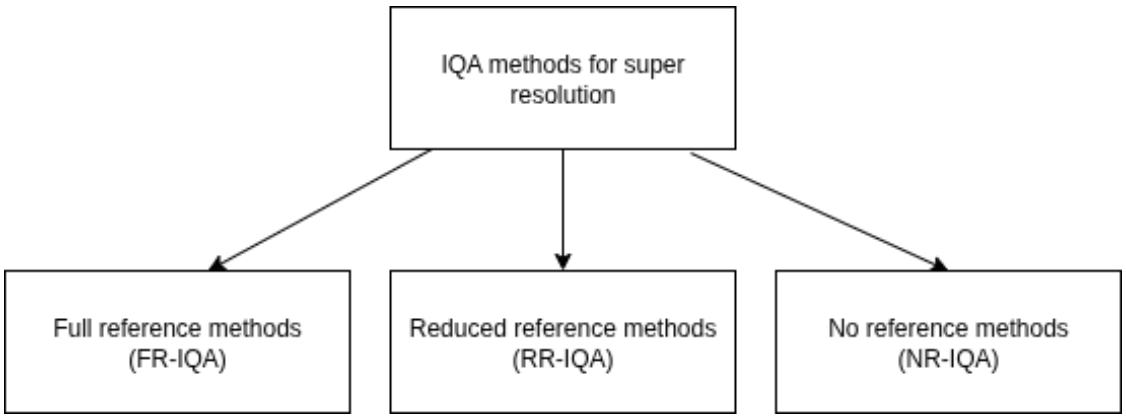


Figure 1.5: *Classification of IQA methods for super-resolution.*

Image quality assessment (IQA) is the process of evaluating how visually pleasing or accurate an image is, and it plays a vital role in various fields, with some examples shown in Figure 1.6. The main objective is to determine how closely an image resembles its ideal version or how well it maintains critical visual details. IQA techniques are typically divided into subjective and objective assessments [34]. Subjective assessment involves human observers who rate the image quality based on visual perception. While this approach provides reliable results, it is time-consuming, costly, and not scalable. In contrast, objective assessment methods use mathematical models and algorithms to estimate image quality automatically and consistently.

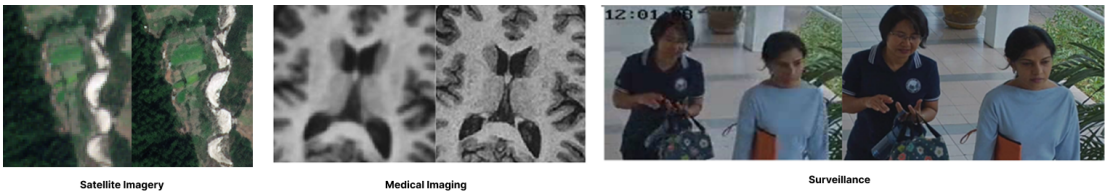


Figure 1.6: *Examples of imaging applications: satellite imagery, medical diagnostics, and surveillance systems [1, 2, 3].*

Objective IQA methods analyze various image features including structure, texture, luminance, contrast, and sharpness. Traditional techniques rely on signal processing tools such as the Discrete Cosine Transform (DCT) for analyzing frequency components, edge detection algorithms for identifying contours and boundaries, and statistical measures to assess image clarity or degradation. More recently, deep learning-based approaches

have emerged, where neural networks are trained to predict image quality in a way that approximates human visual perception. These models have shown promising results but require large annotated datasets and significant computational resources.

In figure 1.5, we observe that the IQA methods can also be classified based on the availability of a reference image. In full-reference methods, the original high-quality image is available for direct comparison. Reduced-reference methods rely on partial information from the reference image. In contrast, no-reference methods attempt to evaluate image quality without any reference, making them particularly challenging and relevant in real-world scenarios [5].

Image quality is often affected by various types of distortions, including noise, blur, compression artifacts, color inaccuracies, and artifacts introduced during upscaling processes like super-resolution. Identifying and quantifying these distortions is essential for evaluating image processing algorithms and for ensuring that visual content meets acceptable standards for interpretation, usability, and aesthetic appeal [35].

1.2.1. Full-Reference Image Quality Assessment (FR-IQA)

In general, Full-Reference Image Quality Assessment (FR-IQA) delivers the best performance since it relies on a distortion-free reference image that provides complete information for comparison. However, a high-quality reference image is not always available. Among the FR-IQA methods, we have Peak Signal to Noise Ratio (PSNR) [8], which evaluates the loss between an original image and a reconstructed one, and is defined by the formula:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{1.1}$$

where MAX is the maximum possible pixel value (for example, 255 for an 8-bit image), and MSE (Mean Squared Error) is given by:

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 . \tag{1.2}$$

In this context, $I(i,j)$ denotes the pixel value at position (i,j) in the original image, and $K(i,j)$ denotes the corresponding pixel value in the reconstructed image. The variables m and n represent the dimensions (height and width) of the image.

However, PSNR does not always correlate well with human visual perception, especially in cases where structural distortions are more noticeable than pixel-wise errors. To overcome these limitations, the Structural Similarity Index Measure (SSIM) was proposed

by Wang et al. [34]. Unlike PSNR, SSIM evaluates image quality by comparing local patterns of pixel intensities that have been normalized for luminance and contrast. It considers three components: luminance similarity, contrast similarity, and structural similarity, is defined by the formula:

$$\text{SSIM}(x, y) = l(x, y) \cdot c(x, y) \cdot s(x, y), \tag{1.3}$$

Where the components are defined as:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}. \tag{1.4}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}. \tag{1.5}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{1.6}$$

Typically, the constants are set such that $C_3 = C_2/2$. The values μ_x and μ_y are the local means of x and y, σ_x, σ_y are their standard deviations, and σ_{xy} is their covariance.

In addition, the *Learned Perceptual Image Patch Similarity* (LPIPS) metric [36] represents a more recent approach that leverages deep learning to better capture perceptual similarity. LPIPS extracts deep features from both the reference and distorted images using a pretrained CNN, such as AlexNet [37] or VGG [38].

For each layer l of the CNN, the distance between the corresponding feature maps of the two images x and y is computed as:

$$d_l(x, y) = \left\| w_l \cdot (F_l^x - F_l^y) \right\|_2^2 \tag{1.7}$$

where:

- F_l^x and F_l^y are the feature maps of image x and image y at layer l,
- w_l is a learned weighting vector that adjusts the contribution of each channel.

The final LPIPS distance is computed by summing over all layers:

$$\text{LPIPS}(x, y) = \sum_l d_l(x, y) \tag{1.8}$$

However, in the specific context of super-resolution, we have several evaluation methods for super-resolution (SR-IQA), among which is Super Resolution Image Fidelity (SRIF) [4].

The SRIF method as illustrated in 1.7 evaluates the visual quality of super-resolved (SR) images by dividing both the SR and high-resolution (HR) images into small regions called patches, and analyzing them using two complementary fidelity dimensions: deterministic fidelity (DF) and statistical fidelity (SF). Regarding deterministic fidelity, a 3-level Gaussian pyramid is applied to both the SR and HR images in order to retain only the global structural components such as edges and contours. Then, a local structural comparison is performed using the SSIM-based formula:

$$D_{\text{local}}(x, y) = \frac{\sigma_{xy} + C_1}{\sigma_x \sigma_y + C_1} \tag{1.9}$$

where x and y are local patches from the HR and SR images, respectively, σ_x and σ_y are their standard deviations, σ_{xy} is the cross-correlation between the patches, and C_1 is a small stabilizing constant. Each patch is then weighted based on its information content to produce a scale-specific DF score:

$$D_{l,j} = \frac{\sum_i w_{l,j,i} D_{\text{local}}(x_{l,j,i}, y_{l,j,i})}{\sum_i w_{l,j,i}} \tag{1.10}$$

The scores across scales are combined using a multiplicative weighted average:

$$D_l = \prod_{j=1}^{K} (D_{l,j})^{\alpha_j} \tag{1.11}$$

where K is the total number of scales (typically 5), and α_j is the weight assigned to scale j. The overall deterministic fidelity is then obtained by summing the results across pyramid levels with specific weights:

$$D = \sum_l w_l D_l \tag{1.12}$$

This hierarchical and multi-scale fusion process ensures accurate assessment of structural integrity across different levels of detail.

Concerning statistical fidelity, a 3-level Laplacian pyramid is used to extract textures and fine details. This is done by first constructing a Gaussian pyramid, then upsampling (interpolating) the image at level $l + 1$ to the size of level l, and computing the difference:

$$L_l = G_l - \text{expand}(G_{l+1}) \tag{1.13}$$

Each Laplacian map is locally normalized:

$$L_l(i,j) = \frac{L_l(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{1.14}$$

where $\mu(i,j)$ and $\sigma(i,j)$ denote the local mean and standard deviation, and C is a small constant. The similarity between the statistical distributions of the Laplacian maps from the HR and SR images is then computed using the Kullback-Leibler divergence:

$$S_l = \int p_l(x) \log\left(\frac{p_l(x)}{q_l(x)}\right) dx \tag{1.15}$$

where $p_l(x)$ and $q_l(x)$ are the normalized probability distributions from the HR and SR Laplacian maps, respectively. Finally, the overall statistical fidelity is calculated by a weighted summation:

$$S = \sum_l w_l S_l \tag{1.16}$$

To combine the deterministic fidelity (DF) and statistical fidelity (SF) scores into a final super-resolution quality score, the SRIF method employs an **uncertainty weighting** strategy. Instead of assigning fixed weights to DF and SF, this approach computes adaptive weights based on the confidence (or uncertainty) associated with each score. First, two content-aware metrics are calculated: the *sharpness ratio sr* and the *texture richness ratio tr*. The sharpness ratio compares the local phase coherence (LPC-SI) of the super-resolved (SR) and high-resolution (HR) images:

$$sr = \frac{\text{LPC-SI}_{\text{SR}}}{\text{LPC-SI}_{\text{HR}}} \tag{1.17}$$

The texture richness ratio measures entropy differences between Laplacian and Gaussian pyramid levels:

$$tr = \frac{ed\text{-}L_1}{eo\text{-}G_2} \tag{1.18}$$

These two indicators are combined to produce an *assorted factor* f that reflects the image's structural and textural complexity:

$$f = sr^{\alpha} + tr^{\alpha} \tag{1.19}$$

The training data is grouped into bins based on the assorted factor f, and within each bin, the variance of the prediction errors for DF and SF is computed, denoted as v_d and v_s respectively. These variances represent the uncertainty of each metric: higher

variance implies lower reliability. The weights assigned to DF and SF are then inversely proportional to their respective uncertainties:

$$w_d = \frac{v_s}{v_d + v_s} \quad , \quad w_s = \frac{v_d}{v_d + v_s} \tag{1.20}$$

Finally, the overall super-resolution image fidelity score is computed as the weighted sum:

$$Q_{ds} = w_d \cdot D + w_s \cdot S \tag{1.21}$$

This uncertainty-aware fusion ensures that the more reliable metric (whether structural or statistical) contributes more to the final quality score, resulting in a robust and content-adaptive quality assessment.

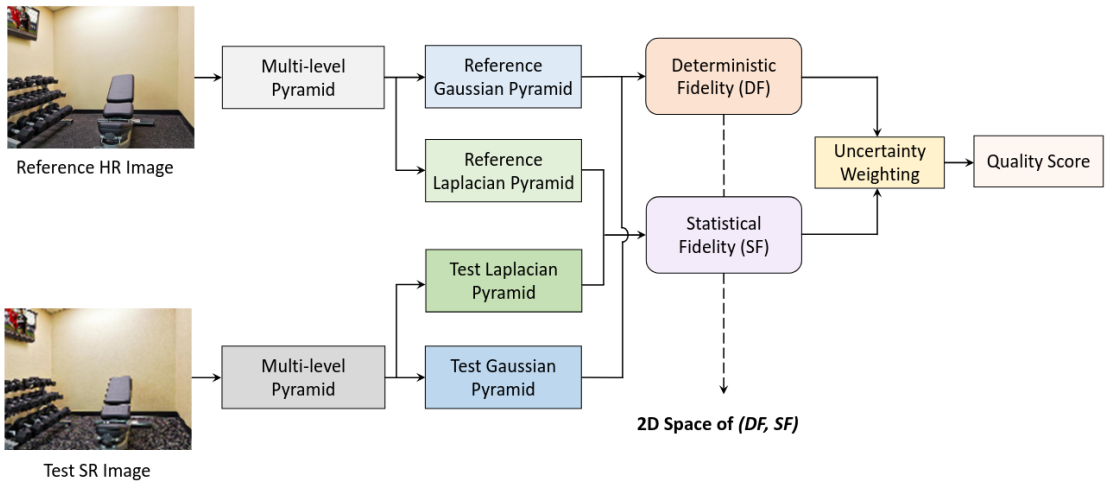


Figure 1.7: *SRIF: A Super-Resolution Quality Assessment Method Based on Deterministic and Statistical Fidelity [4].*

Among the most prominent current methods for full-reference super-resolution image quality assessment (FR-SR-IQA) is BiAtten-Net (Li24) [39], a dual-stream neural network architecture. It consists of two parallel branches: one processes the high-resolution (HR) reference image, and the other processes the super-resolved (SR) image. This bi-directional structure allows the network to compare the two images in both directions: assessing how well the SR image resembles the HR reference, and conversely, how the HR image could be transformed into the SR version.

The process begins by dividing both the HR and SR images into small patches. Each patch is passed through an encoder—a neural network that extracts feature maps.

These features are then fed into a Bi-directional Attention Block (BAB), which applies a transformer-inspired attention mechanism. Within each BAB, the feature maps are projected via 1×1 convolutions into three matrices: Q (Query), K (Key), and V (Value). The Query and Key matrices are used to compute attention scores, highlighting the relative importance of each patch, while the Value matrix helps reconstruct refined feature representations. Additionally, the variance of the dot products between Q and K from both HR and SR streams is calculated to normalize the attention.

To further refine the extracted features, a second BAB is applied. The outputs from both attention blocks undergo pooling to reduce dimensionality, followed by concatenation. Finally, the concatenated features are passed through a Fully Connected Layer (FCL), which outputs the final quality score.

1.2.2. Reduced-Reference Image Quality Assessment (RR-IQA)

In the specific context of super-resolution, image quality assessment can be addressed using Reduced-Reference methods (RR-IQA), which rely on partial information about the reference image. Instead of requiring the original high-resolution (HR) image, these methods utilize the low-resolution (LR) image that serves as input to the super-resolution process. Although this LR image does not fully represent the final output, it retains valuable information—particularly regarding edge structures—that can assist in assessing image quality. As such, it functions as a partial reference, which justifies the term reduced reference [4]. However, RR-IQA methods also come with inherent limitations. Because they rely only on limited information from the LR image, they may fail to capture subtle distortions or fine-grained textures introduced during the super-resolution process. This partial reference may not be sufficient to accurately assess artifacts such as over-smoothing or hallucinated details that do not appear in the LR input. Consequently, the assessment may lack precision compared to full-reference approaches, especially in challenging scenarios where high perceptual fidelity is critical.

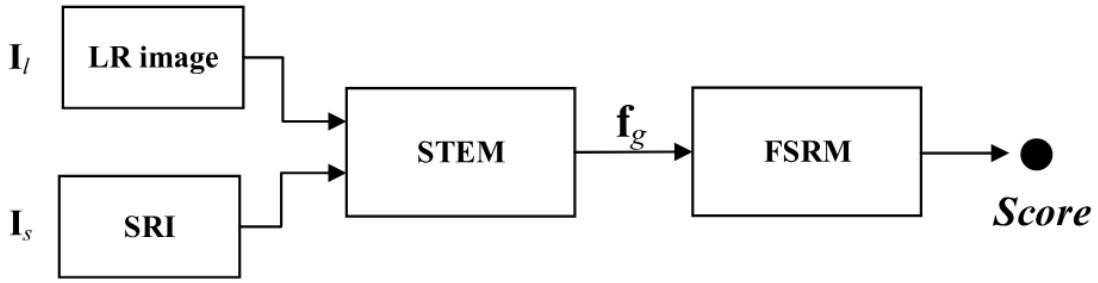


Figure 1.8: *RR-SRIQA framework by Zhou et al. (2022), using LR and SR images to predict quality via STEM and FSRM modules [5].*

Among the various Reduced-Reference Image Quality Assessment (RR-IQA) methods, one notable example is the deep learning-based approach proposed by Fei Zhou and collaborators [5]. The model consists of two main modules as shown in the figure 1.8: the STEM (Structure–Texture feature Extraction Module) and the FSRM (Feature-to-Score Regression Module). The STEM module as can be seen in Figure1.9 extracts key visual features related to image quality. It starts with preprocessing, where the low-resolution (LR) image is bilinearly interpolated to match the size of the super-resolved image (SRI), followed by a structure-texture decomposition applied to both images to isolate structural parts (orientation, edges) and textural parts (fine patterns). This results in four components, but the textural component of the LR image is ignored due to unreliability. Next, local feature extraction is performed through three parallel branches: a manual method comparing edge orientations, a deep CNN capturing high-frequency content differences, and another deep CNN learning texture similarities between the SRI and high-resolution (HR) images during training. Each branch generates local feature maps. Finally, these maps are globally aggregated into a 192-dimensional feature vector, combining similarity histograms and spatial pyramid statistics. The FSRM module takes this vector as input and uses a shallow three-layer fully-connected MLP network to predict an image quality score.

1.2.3. No-Reference Image Quality Assessment (NR-IQA)

No-Reference Image Quality Assessment (NR-IQA) involves evaluating the quality of an image without relying on a reference image. This assessment is based solely on the image's intrinsic characteristics, such as local contrast, sharpness, and texture. In

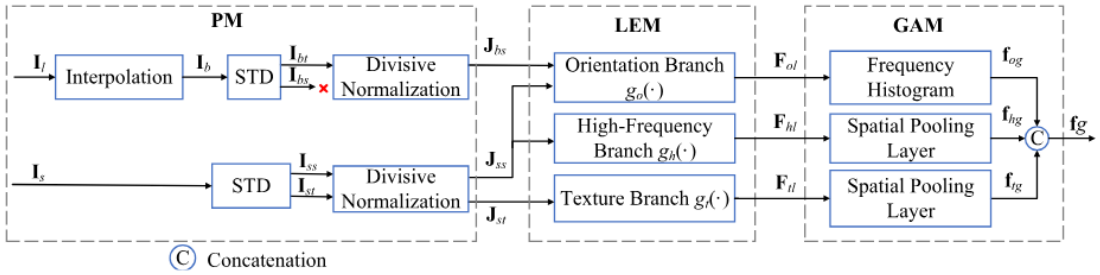


Figure 1.9: *Structure of the STEM module, which extracts orientation, high-frequency, and texture features from both LR and SR images through normalization, feature decomposition (LEM), and pooling operations (GAM) [5].*

practical scenarios, obtaining corresponding reference images is often unrealistic, which highlights the importance of NR-IQA methods [18]. Image Quality Assessment (IQA) methods are generally categorized into traditional approaches that mainly rely on statistical analysis of images. These methods extract global or local features—such as intensity histograms, gradient distributions, or texture properties—to estimate perceived image quality [18].Among traditional no-reference image quality assessment (NR-IQA) methods, BRISQUE *(Blind/Referenceless Image Spatial Quality Evaluator)* [40] stands out as a technique that evaluates image quality by analyzing deviations from natural scene statistics in the spatial domain. The core of BRISQUE involves computing the Mean Subtracted Contrast Normalized (MSCN) coefficients of an image, mathematically defined as:

$$I_{norm}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1}, \tag{1.22}$$

where $I(i,j)$ denotes the pixel intensity at location (i,j), $\mu(i,j)$ is the local mean, and $\sigma(i,j)$ is the local standard deviation computed over a neighborhood. These MSCN coefficients are modeled using a Generalized Gaussian Distribution (GGD), and the extracted statistical parameters serve as features that characterize the perceived image quality. By quantifying the deviation of these parameters from those observed in pristine natural images, BRISQUE effectively predicts perceptual quality without requiring a reference image.

Similar to BRISQUE, the *Natural Image Quality Evaluator* (NIQE) [41] begins by extracting features from the Mean Subtracted Contrast Normalized (MSCN) coefficients of the image. These features are used to fit a multivariate Gaussian model based on a corpus of high-quality natural images. For a given test image, the same features are computed and the Mahalanobis distance between the test feature vector and the reference

Gaussian model is calculated as the NIQE score:

$$\text{NIQE}(I) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}, \tag{1.23}$$

where x is the feature vector of the test image, and μ and Σ represent the mean and covariance of the model derived from natural images.

Another traditional method is *Distortion Identification-based Image Verity and INtegrity Evaluation* (DIIVINE) [42]. The image is first decomposed into subbands in both frequency and orientation using a steerable pyramid (a type of multi-orientation wavelet transform). Then, a rich set of statistical features is extracted, including marginal statistics (such as shape and standard deviation of coefficients, modeled using Generalized Gaussian Distributions), spatial correlations, inter-scale correlations (e.g., high-pass/ band-pass relationships), and inter-orientation dependencies. A divisive normalization inspired by the human visual system is applied to reduce local redundancy. The quality assessment is carried out in two stages: first, a Support Vector Machine (SVM) classifier probabilistically identifies the type of distortion among known categories. Then, for each identified distortion type, a Support Vector Regressor (SVR) estimates a quality score based on the extracted features. These individual scores are combined according to the distortion probabilities to obtain the final overall quality score.

In line with other no-reference image quality assessment approaches that rely on natural scene statistics, the *Blind Image Integrity Notator using DCT Statistics* (BLIINDS-II) [43] method is based on a statistical model of Discrete Cosine Transform (DCT) coefficients. It leverages the natural scene statistics (NSS) of undistorted images in the DCT domain, under the assumption that visual distortions significantly alter these statistical properties. The process begins by dividing the input image into local blocks, on which a 2D DCT is applied. The extracted DCT coefficients are modeled using a generalized Gaussian distribution, and the model parameters—particularly the shape parameter γ—serve as indicators of visual quality. Several types of features are derived, including frequency variation, energy ratios across frequency subbands, and orientation-based variations. These features are extracted at multiple spatial scales to capture the effects of distortions at different resolutions. A simple Bayesian probabilistic model, trained on subjectively labeled images (e.g., from the LIVE IQA database), is then used to predict a perceptual quality score. BLIINDS-II is characterized by its computational efficiency, independence from reference images, and strong correlation with human subjective assessments of image quality.

The HOSA (High Order Statistics Aggregation) [44] method also is an advanced

blind image quality assessment (BIQA) approach based on the aggregation of high-order statistics. It consists of three main stages. First, local image patches of size $B \times B$ are extracted on a regular grid. Each patch $I(i, j)$ is normalized using:

$$x(i,j) = \frac{I(i,j) - \mu}{\sigma + 10}, \tag{1.24}$$

where μ and σ are the local mean and standard deviation of the patch, and 10 is a stabilizing constant. The resulting feature vectors are then whitened using Zero-phase Component Analysis (ZCA) to remove linear correlations among dimensions. Second, a codebook of $K = 100$ codewords is generated via K-means clustering. Each cluster (or codeword) is represented not only by its mean μ_k, but also by its diagonal variance σ_k^2 and dimension-wise skewness γ_k, thereby capturing first-, second-, and third-order statistics. Third, each patch x_i is softly assigned to its r nearest codewords using Gaussian weights:

$$\omega_{ik} = \frac{e^{-\beta \|x_i - \mu_k\|^2}}{\sum_{j:k \in rNN(x_j)} e^{-\beta \|x_j - \mu_k\|^2}}. \tag{1.25}$$

For each codeword k, the weighted differences of mean, variance, and skewness between the assigned patches and the codeword are computed as follows:

$$m_k^d = \sum_{i:k \in rNN(x_i)} \omega_{ik} \cdot x_i^d - \mu_k^d. \tag{1.26}$$

$$v_k^d = \sum_{i:k \in rNN(x_i)} \omega_{ik} \cdot (x_i^d - \hat{\mu}_k^d)^2 - \sigma_k^{2,d}. \tag{1.27}$$

$$s_k^d = \sum_{i:k \in rNN(x_i)} \omega_{ik} \cdot \frac{(x_i^d - \hat{\mu}_k^d)^3}{(\hat{\sigma}_k^{2,d})^{3/2}} - \gamma_k^d, \tag{1.28}$$

where d denotes the dimension, and $\hat{\mu}_k^d$, $\hat{\sigma}_k^{2,d}$ are the weighted local mean and variance.

The feature vectors m_k, v_k, and s_k are concatenated into a global representation V. This representation is then subjected to signed power normalization:

$$f(v) = \text{sign}(v) \cdot |v|^\alpha, \tag{1.29}$$

with $\alpha = 0.2$, followed by L_2 normalization.

Finally, a linear Support Vector Regression (SVR) model is trained to map the aggregated feature vector V to a perceptual quality score. This combination of high-order statistical modeling, soft feature assignment, and robust normalization allows HOSA to assess image quality accurately and efficiently, with strong generalization across various

distortion types and image contents.

In recent years, deep learning-based methods have emerged as powerful alternatives to traditional image quality assessment techniques. Unlike conventional approaches that rely on handcrafted features, deep learning models, particularly convolutional neural networks (CNNs), are capable of automatically learning hierarchical and discriminative features directly from raw images. These models can capture complex visual distortions and generally exhibit a stronger correlation with human perception.

Among the most notable models is HyperIQA [45], which employs a self-adaptive hypernetwork architecture to perform blind image quality assessment (BIQA) on authentically distorted images. The assessment process involves three stages. First, a ResNet-50 backbone extracts high-level semantic features from the input image, enabling the model to understand its content. Second, a hypernetwork leverages these semantic features to dynamically generate the weights of a separate prediction network, allowing the system to adapt its quality assessment strategy to the specific content of each image. Finally, a target network integrates both local distortion features and global semantic features, using the generated parameters to produce a quality score. This separation between content understanding and quality prediction closely mirrors human visual perception and improves the model's generalization across diverse real-world images.

The Deep Bilinear Convolutional Neural Network (DBCNN) [46] also relies on two distinct convolutional networks. The first network, S-CNN, is trained to recognize types and levels of artificial distortions, while the second network VGG-16, pre-trained on ImageNet [47], extracts robust features from naturally distorted images. The outputs of these networks are combined through bilinear pooling, which captures the interactions between the two types of distortions. This bilinear representation is then normalized and used to predict an overall quality score.

Another approach, CONTRIQUE [48], evaluates image quality automatically using unsupervised contrastive learning. During training, the model learns to recognize different types of degradations, such as blur, noise, and compression, by comparing images with and without defects. A convolutional neural network extracts the relevant visual features, which are then fed into a simple linear regression model to predict a quality score. CONTRIQUE performs well on both artificially and realistically degraded images, without requiring large annotated datasets. Other notable methods include ReIQA [49], ARNIQA [50], NIMA [51], Paq-2-PiQ [52], and Clip-IQA [53].

Within the domain of no-reference image quality assessment (NR-IQA) for super-resolution, both traditional and deep learning approaches have been proposed. Traditional methods include KLTSRQ [54], which exploits local texture structures to capture SR-

induced distortions, the ODU framework [55] that extracts distortion-sensitive features from DCT and paired-products under both opinion-distortion-aware and unaware settings, and the regression-based models of Zhang et al. [56, 57], which progressively refine quality prediction through cascade regression and adaptive feature selection. On the deep learning side, Bare et al. [58] introduced a CNN with a label distribution strategy for patch-wise quality estimation, while Zhang (2022) [59] proposed a distortion-aware deep model tailored for SR scenarios. More advanced architectures, such as CN-BSRIQA [60], adopt a two-stage cascaded framework combining CNN feature extraction with DBN refinement, whereas DBSRNet [61] employs a dual-branch design integrating residual/RFBNet modules and Vision Transformer blocks for local-global feature fusion. Recently, SFD-IQA [62] introduced a CLIP-based semantic feature discrimination strategy with feature and text-guided discriminators, enabling opinion-unaware quality prediction without IQA-specific training. Collectively, these approaches highlight the evolution from handcrafted distortion-sensitive descriptors toward hybrid deep learning models leveraging semantic alignment and attention mechanisms, significantly enhancing robustness and generalization in SR-IQA.

1.3. Explainability and Interpretability of Artificial Intelligence Models (XAI)

1.3.1. Background and Motivation for Explainable AI (XAI)

Explainability is now regarded as an essential component in the development and deployment of artificial intelligence models, particularly those based on deep learning. Indeed, the complex and opaque nature of these models prevents users from understanding the mechanisms behind their decisions, which undermines trust in AI systems. Making models explainable, therefore enables professionals—even those from non-technical backgrounds—to understand better and accept the results produced by these tools and promote their adoption. Moreover, explainability plays a key role in debugging models by making it easier to identify errors, biases, or inconsistencies related to model design or input data. Ultimately, it serves as a fundamental lever for compliance, particularly in contexts where legal and ethical requirements—such as transparency, accountability, and the right to explanation—are becoming increasingly unavoidable. Explainability is therefore at the heart of a responsible AI approach that is reliable, understandable, and aligned with current standards [63].

1.3.2. Distinction Between Explainability and Interpretability

In the field of Explainable Artificial Intelligence (XAI), a fundamental conceptual distinction is made between the notions of interpretability and explainability. This distinction, often overlooked or used interchangeably in the literature, was clearly established by Barredo Arrieta et al. (2020) [63] in a seminal review. According to the authors, interpretability refers to an intrinsic property of a model—its ability to be directly and intuitively understood by a human without the need for external analytical tools. It is closely related to model transparency and is expressed through three levels:

- **Simulatability:** the model can be entirely understood mentally.
- **Decomposability:** each part of the model is understandable individually.
- **Algorithmic transparency:** the overall mathematical functioning of the model can be analyzed.

These characteristics are typical of models such as linear regressions, simple decision trees, or Bayesian models [63]. In contrast, explainability is an extrinsic and active property. It refers to the mechanisms, tools, or post-hoc methods designed to provide explanations for the behavior of complex models, often referred to as "black boxes." These explanation techniques are commonly applied to deep neural networks, random forests, or support vector machines, where direct interpretation is not feasible due to architectural complexity. Among such methods are saliency map visualizations, input sensitivity-based approaches, and simplified surrogate models like LIME [63].

This distinction is not merely terminological—it has profound implications for model design objectives, usage constraints, and user trust in AI systems. While interpretability enables immediate and reliable understanding of the model, which is essential in fields such as healthcare, law, or finance, explainability is often a retrospective workaround to mitigate the opacity of more performant yet non-transparent models. Therefore, within the scope of this thesis, the emphasis will be placed on interpretability as a central focus, with the aim of promoting the development of transparent, fair, and auditable models aligned with the principles of Responsible AI [63].

1.3.3. Typologies of Interpretability

Interpretability in artificial intelligence can be categorized along two fundamental dimensions: intrinsic vs. post-hoc interpretability, and local vs. global interpretability

[63]. On one hand, intrinsic interpretability refers to the ability of a model to be directly understood due to its simple and transparent structure. Models such as linear regression or shallow decision trees are considered intrinsically interpretable, as their functioning can be analyzed without external tools. In contrast, post-hoc interpretability involves techniques applied after model training to explain the behavior of complex and opaque models (so-called black-box models), such as deep neural networks. Post-hoc methods—including LIME, saliency maps, or local surrogate models—aim to make such models understandable to humans after the fact [63].

On the other hand, local interpretability focuses on explaining a specific prediction made by the model for an individual instance, whereas global interpretability seeks to provide an overall understanding of the model's behavior across the entire dataset. Local interpretability is crucial in contexts where each individual decision must be justified (e.g., medical diagnosis), while global interpretability is essential for assessing the consistency, internal logic, or potential biases of a model [63]. Although these typologies differ in scope and application, they are complementary and should be selected according to the usage context, stakeholder needs, and the requirements for transparency and accountability [63].

1.3.4. Relevant Explainability Methods for Imaging

Among the different methods of visual explainability, there are local methods based on gradients such as Grad-CAM, LIME, and SHAP. Grad-CAM (*Gradient-weighted Class Activation Mapping*) [64] is a technique used to visualize the regions of an image that most strongly influence a convolutional neural network's (CNN) decision for a given class, for example, Table 1.1 illustrates which regions of the image were used to predict whether the output corresponds to a dog or a cat. It relies on the gradients of the class score with respect to the feature maps of a convolutional layer. First, a forward pass is performed through the network to compute the raw score y^c for the target class c. Then, the gradient of this score with respect to the activations A^k of a selected convolutional layer is computed:

$$\frac{\partial y^c}{\partial A^k_{ij}} \tag{1.30}$$

These gradients are spatially averaged (over the spatial dimensions i and j) to obtain importance weights α^c_k for each feature map k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1.31}$$

where Z is the total number of pixels in each feature map A^k. Then, a weighted combination of the activation maps is performed, followed by a ReLU function to retain only the positive contributions:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{1.32}$$

The resulting map $L_{\text{Grad-CAM}}^c$ is a coarse heatmap that highlights the important regions for predicting class c. It is then upsampled to the original image resolution using bilinear interpolation and overlaid on the input image to visualize the most influential areas. Grad-CAM is valuable because it requires no modification to the original model architecture and can be applied to various tasks such as image classification, captioning, and Visual Question Answering (VQA).

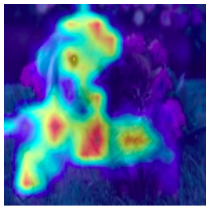
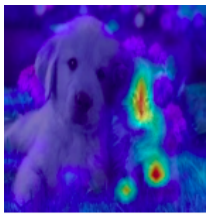
Category	**Image**	**GradCAM**
Dog		
Cat		

Table 1.1: *Grad-CAM visualizations of a ResNet model's classification of the same input image as 'Dog' and 'Cat'. The heatmaps show the image regions that most influenced each prediction [7].*

Although Grad-CAM is very effective for convolutional networks, an alternative approach such as LIME (Local Interpretable Model-agnostic Explanations) [65] begins by dividing the original image into several regions called superpixels, which are groups of visually similar pixels. This representation makes the image easier for humans to interpret. Each superpixel is then encoded in binary form: 1 indicates that the region is present, and

0 indicates that it is absent. LIME then generates many modified versions of the image by randomly masking certain superpixels (usually by graying them out) while keeping the others visible. These modified images are passed to the original classification model to obtain the corresponding predictions. This process produces a dataset of perturbed images and their associated predictions. Each version is weighted according to its similarity to the original image, giving greater importance to those closer to the original. Finally, LIME trains a simple interpretable model (such as a linear regression with a small number of variables) on these weighted data to approximate the behavior of the complex model in the vicinity of the original image. The outcome is a visual explanation that highlights the most important superpixels contributing to the prediction, making the model's decision more interpretable.

SHAP (SHapley Additive exPlanations) also [66] applies Shapley values—a concept from game theory—to the interpretation of computer vision models. For a given image, it assigns each pixel (or superpixel) a value representing its influence on the model's prediction. SHAP systematically assesses the impact of each region by masking it in various combinations and computing its average contribution. For instance, in classifying a handwritten digit like "8", the pixels forming the loops receive positive SHAP values, while the background has a neutral or negative influence. To efficiently handle images, techniques such as Kernel SHAP (for general models) or Deep SHAP (optimized for CNNs) approximate these complex computations. The results are visualized through heatmaps (red indicating positive impact, blue negative), highlighting which parts of the image truly guided the model's decision.

1.3.5. Application of XAI to SR Models

Despite the growing importance of image super-resolution and the emergence of numerous enhancement methods, relying confidently on a single approach remains a major challenge—especially in critical fields such as security and medicine. One of the key concerns is that we often do not understand how the model reconstructs degraded regions—such as blurred areas, compressed artifacts, or illegible text in low-resolution images. It is unclear how the model transforms unreadable characters, hidden object details, or an unrecognizable face into a super-resolved image that appears sharp and detailed. For example, Figure 1.10 shows the low-resolution (LR) input and the corresponding super-resolved (SR) output. While the SR image restores some details in the license plate, the reconstructed text differs from the original high-resolution (HR) image. What information does the model rely on to reconstruct these distorted areas? Can it introduce errors in the

process—for instance, mistaking a '2' for a '0' on a car license plate? These uncertainties raise critical questions about the reliability of super-resolution models. This is precisely why interpretability and explainability are essential: to evaluate the trustworthiness of the generated content and to ensure safe and informed use of such models in high-stakes applications.

The interpretability of super-resolution models is becoming increasingly important, enabling designers and quality inspectors to conduct in-depth image analyses and make more informed decisions. However, existing interpretability methods struggle to cope with the complexity of image degradation and the diversity of image models, making it difficult to provide reliable and accurate explanations [67].

Figure 1.10: *Example of a super-resolution (SR) reconstruction error in license plate recognition using SRGAN model, highlighting the need for explainability.*

1.4. Conclusion

This chapter provides an overview of the fundamentals of Image Super-Resolution (SR), outlining the evolution of methods from classical approaches to advanced deep learning architectures such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Transformers. Key challenges in SR were also discussed, with

particular attention to the trade-off between fidelity and perceptual quality, as well as the generalization ability of existing models.

In parallel, the main categories of image quality assessment techniques are reviewed, namely Full-Reference, Reduced-Reference, and No-Reference methods, emphasizing their respective roles in evaluating SR performance.

Finally, the concepts of explainability and interpretability in artificial intelligence (XAI) are introduced, underlining their growing relevance in computer vision and, more specifically, in SR. The principal interpretation methods applied to image generation models were highlighted, setting the stage for a deeper investigation into their usefulness in understanding and improving SR models in the following chapters.

Chapter 2

Methodology

Contents

2.1. Benchmarking NR-IQA Models for Super-Resolution

2.1.1. Selection of No-Reference Image Quality Assessment (NR-IQA) Methods

In this work, we introduce a comprehensive benchmark of No-Reference Image Quality Assessment (NR-IQA) models tailored for Super-Resolution (SR) tasks. This choice is driven by the increasing demand for quality evaluation methods that do not require reference images—particularly relevant in real-world applications where high-resolution (ground truth) images are usually unavailable.

In practice, the low-resolution (LR) image cannot serve as a reliable reference for evaluating the quality of its super-resolved (SR) counterpart. This is because the LR image lacks essential visual details, such as textures and fine structures, which are reconstructed or enhanced in the SR image. Consequently, comparisons between SR and LR images are inherently flawed, underscoring the necessity of NR-IQA approaches in such contexts.

The considered NR-IQA models in our benchmark were selected according to three key criteria:

Relevance in recent literature: Preference was given to models that have shown competitive performance in recent studies.

Popularity and adoption: Widely used models were prioritized to facilitate reproducibility and enable fair comparisons.

Methodological diversity: A broad range of approaches was considered, encompassing deep learning–based methods, handcrafted feature models, and transformer-based techniques, to ensure a comprehensive evaluation.

This diversity enables a thorough assessment of the advantages and limitations of each method in the specific context of SR image quality evaluation.

2.1.2. Datasets Used

In our study, we selected four datasets based on several essential criteria. The first criterion concerns the number of scale factors used (for example $\times 2$, $\times 3$, $\times 4$, etc.), which allows the evaluation of super-resolution algorithms' performance at different enlarge- ment levels. The second criterion relates to the number of super-resolution methods applied within each dataset, ensuring sufficient diversity for a thorough comparison. Finally, an important criterion involves the method used for the subjective evaluation of image quality, through Mean Opinion Score (MOS) ratings. These scores are ob- tained from human judgments, using various approaches such as pairwise comparisons, which measure the

perceived quality of the super-resolved images. In table 2.1, we present a comparative overview of the selected dataset, and details on each one are provided in the following.

2.1.2.1. Computer Vision and Image Understanding dataset (CVIU-2017)

To construct the CVIU-2017 [68] dataset, 30 high-resolution (HR) images were selected from the Berkeley Segmentation Dataset (BSD200) [69]. These images were then degraded by downsampling using a Gaussian filter, parameterized by a scaling factor s and a kernel width sigma, in order to generate low-resolution (LR) versions. From these LR images, 1,620 super-resolved (SR) images were generated using 9 different super-resolution methods, including Bicubic, SRCNN, Dong11, Timofte13,among others. To assess the perceptual quality of the SR images, a subjective test was conducted with 50 participants, who assigned each image a visual score ranging from 0 to 10. The final score for each SR image corresponds to the average of the 40 central scores, with the 10 most extreme scores discarded to reduce bias.

2.1.2.2. Quality Assessment Database for SRIs (QADS)

In the QADS [70] database, 20 high-resolution (HR) images were selected as reference images, mainly from the MDID dataset [71], based on the diversity of their visual content. To generate the super-resolved images (SRIs), each reference image was first downsampled by a scaling factor of k=2,3k=2,3, or 44 using bicubic interpolation. Then, 21 super-resolution methods were applied: 4 based on interpolation, 11 based on dictionary learning, and 6 based on deep neural networks (DNNs). This process resulted in a total of 980 super-resolved images.

The MOS (Mean Opinion Score) ratings were obtained through a subjective evaluation involving 100 participants under controlled experimental conditions (low light- ing, calibrated screen, stable environment). A specific graphical interface allowed the subjects to compare two super-resolved images (SRIs) derived from the same high- resolution reference image, in order to judge which one presented better visual quality, or if they were equivalent. The results of these comparisons were then processed using the Pair Comparison Sorting (PCS) algorithm, which ranked the 49 SRIs corresponding to each reference image. The obtained scores were normalized, inconsistent or aber- rant responses were eliminated, and a final MOS score was assigned to each image by calculating the average of the valid evaluations.

2.1.2.3. Real-world SISR Quality dataset (RealSRQ)

The RealSRQ [54] dataset consists of 60 high-resolution (HR) images captured with a DSLR camera at different focal lengths, along with their corresponding low-resolution (LR) versions generated at scaling factors $\times 2$, $\times 3$, and $\times 4$. For each LR image, ten representative SISR algorithms (both classical and deep learning-based) are used to generate super-resolved images, resulting in a total of 1,620 images. The visual quality of these results is assessed through a subjective study based on pairwise comparisons between images, conducted with 60 participants. The binary preference data collected is then converted into continuous scores, known as B-T scores, using the Bradley-Terry statistical model.

2.1.2.4. SR Image quality database with Semi-Automatic Ratings (SISAR)

The SISAR [72] database is one of the largest dedicated to image super-resolution, containing 12,700 SR images generated from 100 reference images across six categories: animals, buildings, humans, sports, plants, and landscapes. These images were created using six super-resolution methods (BICUBIC, RLLR [73], SRCNN [74], VDSR [75], SRCNN+BICUBIC, VDSR+BICUBIC) and various scaling factors (1.5, 2, 2.7, 3, 3.6, and 4). For image quality annotation, SISAR employs an innovative semi-automatic method for generating MOS scores. The core idea is to iteratively apply a downsampling followed by a super-resolution operation, leading to a gradual decrease in the visual quality of the resulting high-resolution image. This degradation follows an exponential law:

$$Q(t) = e^{-bt} \tag{2.1}$$

where $Q(t)$ represents the perceived quality after t iterations, and b is a parameter depending on the image content and the SR algorithm used. To estimate this parameter, the authors conducted subjective tests on a sample of 300 images (30 scenes) evaluated by 21 participants. Once the b values were determined for each combination (image, method, factor), the remaining images were automatically annotated using this formula, eliminating the need for human assessment in every case while maintaining high consistency and reliability of the scores.

Criteria	QADS	CVIU-2017	RealSRQ	SISAR
Year of publication	2019	2017	2022	2020
HR image Number	20 images	30 images	60 images	100 images
SR Image Number	980 images	1620 images	1620 images	12700 images
Methodology for Obtaining Mean Opinion Scores (MOS)	Pair Comparison Sorting	Scores between 0 and 10	Pair Comparison Sorting	A semi-automatic method based on a decreasing exponential law
SR method number	21 methods	9 methods	10 methods	6 methods
SR methods used	Bilinear interpolation Bicubic interpolation Orientational interpolation Fast up-sampling Example-based SR LLE Sparse KRR SCSR ASDS SCDL SF SPM A+ Self-exemplars CCS VDSR VGGNet DCSCN DRRN LapSRN SRGAN	Bicubic interpolation Back Projection Shan08 Glasner09 Yang10 Dong11 Yang13 Timofte13 SRCNN	BCI ASDS SPM Aplus AIS SRCNN CSCN VDSR SRGAN USRnet	BICUBIC RLLR SRCNN VDSR SRCNN + BICUBIC VDSR + BICUBIC RCAN SAN RCAN + BICUBIC RSAN + BICUBIC
Scale factors	2, 3, 4	2, 3, 4, 5, 6, 8	2, 3, 4	1.5, 2, 2.7, 3, 3.6, 4
Number of Participants in Subjective Tests	100 people	50 people	60 people	23 people

Table 2.1: *Comprehensive Comparison of Super-Resolution Benchmark Datasets*

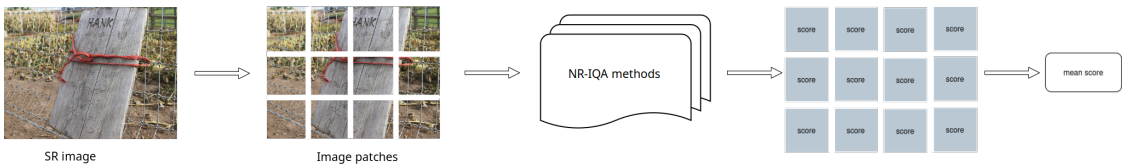


Figure 2.1: *Architecture of the proposed benchmark for evaluating Super-Resolution (SR) models using No-Reference Image Quality Assessment (NR-IQA) on patches*

2.1.3. Experimental Protocol

In this work, we consider four categories of image quality assessment methods: (i) traditional No-Reference IQA (NR-IQA) techniques, (ii) deep learning–based IQA methods, (iii) traditional Super-Resolution IQA (SR-IQA) methods, and (iv) deep learning–based SR-IQA approaches.

For the training and evaluation, two complementary strategies were adopted: local assessment, performed on cropped regions of the images (as illustrated in 2.1 and 2.2), and global assessment, carried out by resizing the images (as illustrated in 2.2).

Figure 2.2: *Image preprocessing strategies: super-resolution (SR), fixed-size crops, and uniform resizing to 224×224.*

Traditional methods were implemented in MATLAB and executed on an Intel i7 processor running at 2.50 GHz. Deep learning–based approaches, on the other hand, were trained on a GPU compute server equipped with Tesla V100S GPUs. Training followed a 5-fold cross-validation protocol with splits defined at the reference image level: 80% of

the data were used for training and 20% for testing, with all predicted scores normalized to the [0, 1] range.

For each fold, we computed correlation metrics to assess the monotonicity of the algorithms, namely the Spearman Rank-Order Correlation Coefficient (SROCC) and the Kendall Rank-Order Correlation Coefficient (KROCC):

- The **SROCC** evaluates how well the ranking of images by the SR-IQA estimated quality corresponds to the ranking based on human subjective evaluations. It is defined as:

$$\text{SROCC} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2.2}$$

  where d_i is the difference between the ranks of the i-th image in the predicted and subjective rankings, and n is the number of images.

- The **KROCC** measures the similarity of orderings (ranks) between the automatic SR-IQA method and human perception based on the number of concordant and discordant pairs of images. Like SROCC, it assesses consistency in relative order rather than linear relationship. It is computed as:

$$\text{KROCC} = \frac{C - D}{\frac{1}{2} n(n-1)} \tag{2.3}$$

  where C is the number of concordant pairs, D the number of discordant pairs, and n the number of images.

Additionally, we evaluated the accuracy of the algorithms using the Pearson Linear Correlation Coefficient (PLCC) and the Root Mean Squared Error (RMSE).

- The **PLCC** measures the strength of the linear relationship between the scores predicted by the SR-IQA method and human subjective scores:

$$\text{PLCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.4}$$

  where x_i are the predicted scores, y_i are the subjective scores, and $\bar{x}$, $\bar{y}$ are their respective means.

- The **RMSE** indicates the average magnitude of the error between predicted and

true values, with lower values indicating better accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.5}$$

Finally, to obtain a robust evaluation, we took the median values of the correlation coefficients (SROCC, KROCC, PLCC) and RMSE over the 5 folds.

2.2. Proposed Method

We propose an exploratory No-Reference Image Quality Assessment method for super-resolved images, termed SR-NR-IQA. Let the input super-resolved image be $I \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width, and 3 represents the RGB channels. The proposed method, illustrated in Figure 2.3, consists of four main steps: patch extraction, multi-scale decomposition, deep feature extraction, and feature fusion with quality prediction.

2.2.1. Patch Extraction

The image I is divided into N non-overlapping patches:

$$\{P_i\}_{i=1}^{N}, \quad P_i \in \mathbb{R}^{h\times w\times 3}, \quad h \leq H, w \leq W$$

Each patch captures localized quality variations. The patches are converted to grayscale:

$$\tilde{P}_i = \text{Gray}(P_i) \in \mathbb{R}^{h\times w}$$

2.2.2. Multi-scale Pyramid Decomposition

Each grayscale patch $\tilde{P}_i$ is decomposed into two pyramid representations:

Gaussian Pyramid: The Gaussian pyramid emphasizes global structural information:

$$G_i^0 = \tilde{P}_i, \quad G_i^l = \text{Downsample}(G_i^{l-1}), \quad l = 1, \ldots, L_G$$

Laplacian Pyramid: The Laplacian pyramid captures fine textural details:

$$L_i^l = G_i^l - \text{Upsample}(G_i^{l+1}), \quad l = 0, \ldots, L_L - 1$$

At each pyramid level, features are extracted and transformed into a compact representation via a fully connected layer:

$$f_i^G = \mathrm{FC}(\{G_i^l\}_{l=0}^{L_G}), \quad f_i^L = \mathrm{FC}(\{L_i^l\}_{l=0}^{L_L})$$

2.2.3. Deep Feature Extraction with ResNet50

Simultaneously, the original RGB patch P_i is input to a ResNet50 network to extract high-level semantic and visual features:

$$f_i^R = \mathrm{ResNet50}(P_i) \in \mathbb{R}^{d_R}$$

ResNet50 employs residual blocks to alleviate the vanishing gradient problem:

$$\mathbf{x}_{l+1} = F(\mathbf{x}_l, \{W_l\}) + \mathbf{x}_l$$

where F denotes a stack of convolutional layers with weights W_l, and the residual connection $\mathbf{x}_l$ helps preserve low-level information while enabling very deep feature extraction.

2.2.4. Feature Fusion and Quality Prediction

The features from Gaussian, Laplacian, and ResNet50 representations are concatenated into a unified vector:

$$f_i = [f_i^G; f_i^L; f_i^R] \in \mathbb{R}^{d_G+d_L+d_R}$$

A fully connected layer predicts the perceptual quality score q_i for each patch:

$$q_i = \mathrm{FC}_{\mathrm{score}}(f_i)$$

The overall image quality score Q is obtained by averaging all patch scores:

$$Q = \frac{1}{N}\sum_{i=1}^{N} q_i$$

Summary: Mathematically, the proposed SR-NR-IQA method can be summarized as:

$$Q = \frac{1}{N}\sum_{i=1}^{N} \mathrm{FC}_{\mathrm{score}}\Big([\mathrm{FC}(\{G_i^l\}), \mathrm{FC}(\{L_i^l\}), \mathrm{ResNet50}(P_i)]\Big)$$

Overall, this formulation effectively captures global structural information, local texture details, and high-level semantic cues, resulting in a robust no-reference quality assessment for super-resolved images, as shown in Figure 2.3.

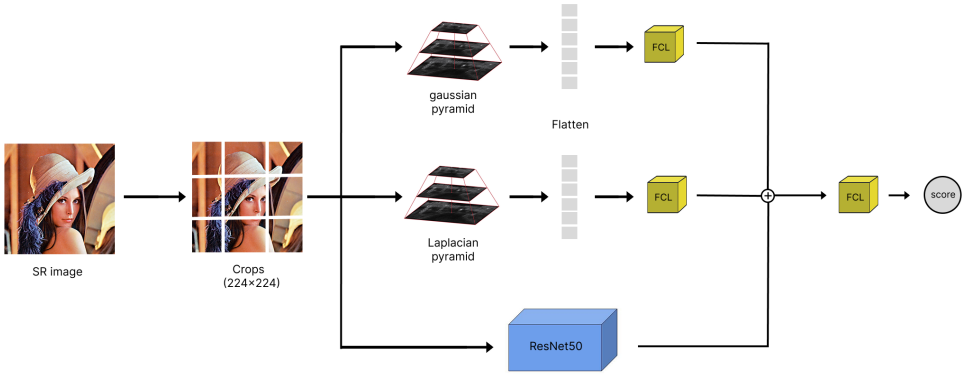


Figure 2.3: *Proposed SR-NR-IQA architecture combining multi-scale pyramids and ResNet50 features.*

2.3. Advancing Interpretability in Super-Resolution Models

2.3.1. Selecting Super-Resolution Models for Interpretation

In order to explore the challenges related to the interpretability of super-resolution models, we selected a diverse set of architectures, each characterized by specific internal representations, information flow mechanisms, and varying levels of decision-making complexity. This diversity allows us to analyze how the architectural design itself influences the potential to interpret the image reconstruction process. The selection includes models based on traditional convolutional neural networks, such as EDSR (*Enhanced Deep Super-Resolution*) [30], CARN (*Cascading Residual Network*) [76], SAN (*second-order attention network*) [77] and RCAN (*Residual Channel Attention Networks* [78].EDSR employs a deep stack of 32 optimized residual blocks, each composed of two 3×3 convolutional layers and a residual (skip) connection, notably omitting batch normalization to preserve the range and distribution of features critical for accurate reconstruction. CARN, on the other hand,operates using a cascade of residual blocks, where the output of each block is shared with the others, enabling the network to combine multi-level features effectively for reconstructing the final high-resolution image. The SAN model is composed of four

main modules: a shallow feature extraction module based on convolution, a deep feature extraction module using attention-enhanced residual groups (NLRG), an upsampling module, and a reconstruction layer. The NLRG incorporates non-local operations and LSRAG blocks, which consist of simplified residual blocks followed by a Second-Order Channel Attention (SOCA) module. The SOCA mechanism enhances inter-channel dependencies by leveraging second-order feature statistics, thereby improving the quality and discriminative power of the learned representations. the RCAN architecture adopts a hierarchical structure called Residual in Residual (RIR), which incorporates both long and short skip connections to facilitate training by allowing the direct passage of low-frequency information. Each Residual Group (RG) contains multiple Residual Channel Attention Blocks (RCAB), which include a channel attention mechanism. This mechanism uses global spatial pooling followed by a recalibration module based on fully-connected layers, activated by ReLU and Sigmoid, to dynamically weight each channel according to its importance. This enables the network to focus on the most informative channels for high-resolution image reconstruction. Finally, the output is produced by an upscaling module and a final reconstruction step. Finally, we used the RRDNet model, which is integrated into the architecture of a GAN model known as ESRGAN [31]. This model relies on RRDB (Residual-in-Residual Dense Blocks), which cleverly combine residual and dense connections to enhance information flow between layers. Instead of using batch normalization, ESRGAN employs techniques such as residual scaling to facilitate learning.

In addition, we selected a GAN-based model known as SRGAN [33], which consists of a generator that converts low-resolution images into high-resolution ones, and a discriminator that attempts to differentiate between generated and real high-resolution images. SRGAN leverages a perceptual loss function that combines content loss—based on features extracted by a pre-trained VGG network—and adversarial loss, which encourages the generation of more realistic images.

To comprehensively cover all architectures, we used a transformer-based model, namely SwinIR [32]. This architecture consists of three main components: an initial convolutional layer that extracts low-level features, followed by a series of Residual Swin Transformer Blocks (RSTB), which capture both local and global dependencies through a shifted window attention mechanism. Finally, a reconstruction module combines these features to produce a high-quality image.

2.3.2. Selection of Explainability Methods (XAI)

To select appropriate XAI methods for super-resolution tasks, we considered several criteria, including compatibility with the image reconstruction nature of the models. Unlike classification tasks where methods like Grad-CAM rely on final class scores, super-resolution generates high-resolution images as outputs, making such approaches unsuitable. Therefore, we focused on techniques capable of interpreting image-to-image transformations, providing meaningful explanations at both local and global levels through visual or feature-based analyses.

2.3.2.1. Local Attribution Maps (LAM)

The *Local Attribution Map* (LAM) [6] is a technique designed to interpret deep super-resolution (SR) networks by revealing which input pixels from the low-resolution (LR) image most strongly contribute to the reconstruction of a specific high-resolution (HR) output patch. As indicated in Figure 2.4 LAM focuses on local patches and seeks to understand how localized features such as textures and edges are recovered by the model. The method builds upon the integrated gradients (IG) framework but introduces two critical modifications tailored for the SR context. First, the baseline input image I' is constructed by blurring the original LR image I using a Gaussian kernel $\omega(\sigma)$, leading to the formulation $I' = \omega(\sigma) \otimes I$, where $\otimes$ denotes convolution. This baseline removes high-frequency components, allowing the attribution process to isolate the contribution of details. Second, rather than employing linear interpolation between the baseline and the original image, LAM defines a progressive blurring path function $\gamma_{pb}(\alpha) = \omega(\sigma - \alpha\sigma) \otimes I$, where $\alpha \in [0, 1]$, enabling a smooth and natural transition from the blurred to the sharp input. Given an SR network F and a simple feature detector D (such as one based on image gradients), LAM computes the contribution of each pixel i via the following integrated gradient formula:

$$\mathrm{LAM}_{F,D}(\gamma)_i = \int_0^1 \frac{\partial D(F(\gamma(\alpha)))}{\partial \gamma(\alpha)_i} \cdot \frac{\partial \gamma(\alpha)_i}{\partial \alpha} \, d\alpha. \tag{2.6}$$

This integral is numerically approximated using a Riemann sum over m discrete steps:

$$\mathrm{LAM}_i \approx \sum_{k=1}^{m} \frac{\partial D(F(\gamma(k/m)))}{\partial \gamma(k/m)_i} \cdot (\gamma(k/m)_i - \gamma((k+1)/m)_i) \cdot \frac{1}{m}. \tag{2.7}$$

To quantify how broadly the model distributes attention over input pixels, the Diffusion Index (DI) is introduced, derived from the Gini coefficient G, which is defined as

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |g_i - g_j|}{2n^2 \bar{g}}, \tag{2.8}$$

where g_i denotes the absolute attribution value at pixel i, $\bar{g}$ is the mean of all attributions, and n is the number of pixels. The Diffusion Index is then computed as DI $= (1 - G) \times 100$, where a higher DI indicates broader utilization of input information by the SR network. Empirical analysis across various SR architectures demonstrates a strong positive correlation between DI and reconstruction quality as measured by the Peak Signal-to-Noise Ratio (PSNR), with a Pearson correlation coefficient of 0.851 and a Spearman rank correlation of 0.880, both with p-values below 10^{-12}. These results confirm that networks which attend to a wider spatial area in the LR image tend to produce more accurate HR outputs. Therefore, LAM combined with the Diffusion Index offers a rigorous framework for interpreting SR networks, diagnosing architectural strengths and weaknesses, and guiding future model design for improved performance.

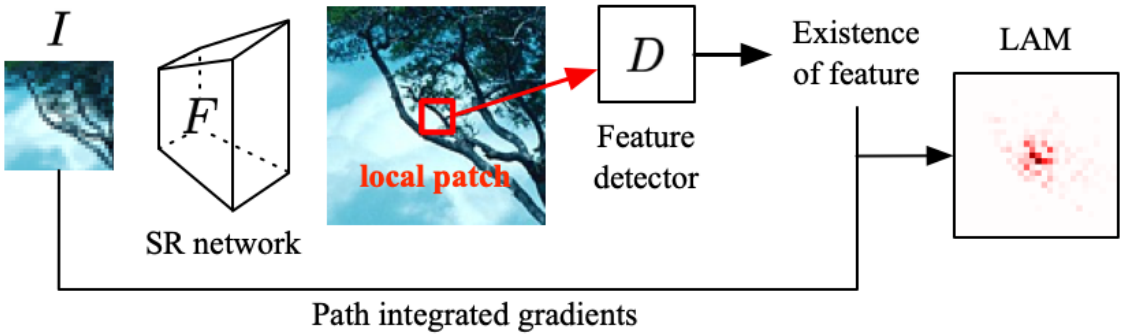


Figure 2.4: *Workflow of Local Attribution Map (LAM) generation [6].*

2.3.2.2. Glocal Attribution Maps

The Glocal Attribution Map (GL-AM) [79] is a gradient-based explainability approach designed to interpret deep super-resolution (SR) networks by combining both local and global perspectives. GL-AM provides insights into how input pixels contribute to the final reconstructed high-resolution output by computing gradients across all layers of the network. The local attribution focuses on determining which low-resolution pixels affect the reconstruction of a specific region of interest (ROI) in the super-resolved image. This is achieved by calculating the gradient of the texture in the ROI with respect to the input,

denoted as $\text{grad}_{\text{local}}^{i\to\text{in}} = g(T_{\text{ROI}})^{i\to\text{in}}$, where $g(\cdot)$ is the gradient function and i refers to the layer index. These gradients are weighted by importance scores $W_{\text{local}}^{i\to\text{in}}$ and aggregated across all layers to obtain the local attribution map using the following expression:

$$G_{\text{local}} = \sum_{i=0}^{n} \text{grad}_{\text{local\_key}}^{i} + \sum_{i=1}^{n-1} \left(\text{grad}_{\text{local\_key}}^{i} - \text{grad}_{\text{local\_key}}^{i-1} \right). \tag{2.9}$$

This formulation captures both the consensus (shared importance across layers) and variation (layer-specific sensitivity) to better understand how different features are used by the network during reconstruction. In contrast, the global attribution captures the influence of entire regions of the input image on the whole SR output. For this purpose, GL-AM calculates the gradient of the output of each layer j with respect to its previous layer i, using $\text{grad}_{\text{global}}^{j\to i} = g(T_{\text{Global}})^{j\to i}$, and multiplies these with corresponding feature maps fea_j to obtain:

$$\text{fea}_{\text{global}}^{j\to i} = \text{fea}_j \times \text{grad}_{\text{global\_key}}^{j}. \tag{2.10}$$

The final global attribution map is then given by:

$$G_{\text{global}} = \sum_{i=0}^{n} \text{fea}_{\text{global}}^{j\to i} + \sum_{j=1}^{n-1} (\text{fea}_{\text{global}}^{j} - \text{fea}_{\text{global}}^{j-1}). \tag{2.11}$$

All attribution maps are resized to match the output dimensions and normalized for visual interpretability, where darker regions in the heatmap indicate stronger influence on the reconstruction. To evaluate these heatmaps, both qualitative and quantitative approaches are employed. One commonly used metric is the Diffusion Index (DI), which measures how many input pixels significantly contribute to the reconstruction, reflecting the model's effective receptive field. GL-AM thus provides a unified and detailed understanding of SR models by explaining their behavior at both fine-grained (local) and holistic (global) levels, offering valuable insights into model design, attention mechanisms, and reconstruction dynamics.

2.3.2.3. Real Attribution Maps (RAM)

Real Attribution Maps (RAM) [67] are an advanced attribution method designed to interpret super-resolution models under real-world industrial conditions. RAM is built upon the Integrated Gradients (IG) framework, which attributes the model's output to its input by accumulating gradients along a path from a baseline image to the actual input

image. The fundamental IG formula is:

$$IG = (\mathcal{I} - \mathcal{I}') \cdot \int_0^1 \frac{\partial \mathcal{F}(\mathcal{I}' + \alpha(\mathcal{I} - \mathcal{I}'))}{\partial \mathcal{I}} \, d\alpha \tag{2.12}$$

where $\mathcal{I}$ is the input image, $\mathcal{I}'$ is the baseline image (a degraded version with minimal meaningful content), and $\mathcal{F}$ is the neural network under analysis.

RAM enhances this attribution by introducing two critical components: Multi-Path Downsampling (MPD) and Multi-Progressive Degradation (MPG). MPD generates multiple downsampled input images using various realistic degradation operators to simulate different industrial scenarios:

$$\mathcal{I}_i^{PT} = \mathcal{I}_{HR}^{(i)} \downarrow_s \tag{2.13}$$

MPG creates degraded baseline images by applying blur and noise to each downsampled input:

$$\mathcal{I}_i'^{PT} = k(\sigma_b) \otimes \mathcal{I}_i^{PT} + n(\sigma_n) \tag{2.14}$$

Then, for each path i, RAM defines a progressive and realistic interpolation function:

$$\gamma_i(\alpha) = k(\sigma_b - \alpha\sigma_b) \otimes \mathcal{I}_i^{PT} + n(\sigma_n - \alpha\sigma_n), \quad \alpha \in [0, 1] \tag{2.15}$$

where $\otimes$ denotes convolution, k is a blur kernel parameterized by σ_b, and n is noise with intensity σ_n.

The attribution for each path is computed by discretizing the integrated gradients:

$$\phi_i = \sum_{j=0}^{k} \left(\gamma_i\left(\frac{j}{k}\right) - \gamma_i\left(\frac{j+1}{k}\right) \right) \cdot \frac{\partial \mathcal{F}(\gamma_i(j/k))}{\partial \gamma_i(j/k)} \tag{2.16}$$

The final Real Attribution Map is obtained by averaging the attributions from all m paths:

$$\phi_{RAM} = \frac{1}{m} \sum_{i=0}^{m} \phi_i \tag{2.17}$$

RAM interprets the results by generating heatmaps that highlight the input regions most influential to the output. These heatmaps are evaluated both qualitatively (via visual inspection) and quantitatively using metrics such as PSNR and SSIM on insertion experiments, and AUC-DEL for deletion-based evaluation. High attribution scores in meaningful image regions (e.g., edges, defects, critical textures) indicate that the model

relies on relevant visual cues, while low attributions elsewhere confirm robustness. RAM therefore provides a reliable and interpretable explanation of model behavior in degraded, real-world conditions.

2.4. Conclusion

This chapter outlined the approach we followed to evaluate image super-resolution models, both in terms of perceived image quality and understanding how the models work internally. On one hand, we built a benchmark to compare several NR-IQA models, selecting those most commonly used and cited in the literature. We also explained how the low- and high-resolution images were generated, as well as the evaluation metrics and experimental protocol applied.

On the other hand, we introduced a set of explainability techniques (XAI) designed to be applied to SR models in later stages of the study. Specifically, we selected LAM, GL-AM, and RAM, which produce attribution maps at both local and global levels. These methods will allow us to visualize the regions of the image that SR models focus on during reconstruction. By setting up this interpretability framework, we aim to better understand model behavior in future analyses.

Chapter 3

Experiments and Results

Contents

3.1. NR-IQA Benchmark Results

3.1.1. Correlation analysis

We present four comparative figures [3.1, 3.2, 3.3, 3.4], each corresponding to a specific dataset: CVIU-2017, QADS, RealSRQ, and SISAR, respectively. Each figure includes the results of the considered image quality assessment metrics. These metrics comprise both traditional no-reference IQA (TR IQA), deep learning-based IQA methods (DL IQA), traditional (TR SRIQA), and deep learning-based super-resolution IQA (DL SRIQA) metrics. For each one, we computed the following correlation coefficients: SROCC, PLCC, KROCC, along with the RMSE.

In the CVIU-2017 dataset, the analysis shows that both traditional image quality assessment methods and super-resolution-specific approaches encounter considerable difficulties. The correlation values remain weak, with deep learning–based SR IQA methods also failing to deliver consistent results, since the SROCC scores fluctuate between 0.2375 and 0.7443. The highest performance is obtained with HyperIQA under the crop-based configuration, which achieves a SROCC of 0.7869, a PLCC of 0.7935, a KROCC of 0.6014, and an RMSE of 0.1510. These outcomes underline the complexity of the CVIU dataset, which remains a challenging benchmark for most IQA and SRIQA methods.

The QADS dataset, by contrast, produces much stronger performances across both traditional IQA measures and deep learning–based methods. This observation indicates that QADS is more suitable than CVIU for both evaluation and training, since it provides a more reliable correlation between predicted quality and perceptual ground truth. The best-performing method is TReS under the resizing-based configuration, which records a SROCC of 0.9199, a PLCC of 0.9200, a KROCC of 0.7491, and an RMSE of 0.1228. Interestingly, even IQA methods not explicitly designed for super-resolution prove effective in this dataset, as they are capable of detecting and modeling super-resolution artifacts with high accuracy.

The RealSRQ dataset presents a very different scenario. Most of the methods applied on this dataset produce scores close to zero, reflecting the substantial difficulty of assessing perceptual quality under real-world super-resolution conditions. Only a few deep learning models demonstrate moderate effectiveness, with Paq-2Piq (crop) achieving a SROCC of 0.6517, a PLCC of 0.7009, a KROCC of 0.4818, and an RMSE of 0.6684. The overall performance distribution highlights that RealSRQ is considerably more complex and demanding than QADS or SISAR, making it the most challenging dataset within this comparative study.

The SISAR dataset yields stronger performances than the previous benchmarks. Deep learning–based methods such as HyperIQA and CONTRIQUE achieve high correlations, with values frequently exceeding 0.7, and maintain relatively low RMSE. HyperIQA (crop) provides the best performance, with a SROCC of 0.9418, a PLCC of 0.9384, a KROCC of 0.7837, and an RMSE of 0.0755. Even traditional metrics, such as BRISQUE, reach a SROCC of 0.7169, which is comparatively high for non-deep-learning methods. These results indicate that SISAR, due to its more structured nature, constitutes a favorable dataset for evaluating both classical and deep learning–based IQA approaches.

Taken together, the four datasets demonstrate the diversity of challenges in super-resolution quality assessment. CVIU-2017 highlights the limitations of both traditional and modern methods in handling complex data. QADS emerges as a more consistent and training-friendly dataset, producing strong correlations across most approaches. RealSRQ, with its near-random correlations for many methods, underscores the difficulty of real-world evaluation and the need for more robust models. SISAR, on the other hand, validates the effectiveness of deep learning–based metrics and even allows traditional measures to achieve moderate success. Overall, the comparison shows that deep learning approaches generally outperform traditional ones, particularly on structured datasets such as QADS and SISAR, while all methods continue to struggle on more challenging databases such as RealSRQ.

The comparative analysis further reveals that IQA and SRIQA methods do not exhibit stability across datasets, as their performance varies considerably depending on the characteristics of the benchmark. A method that achieves strong correlations on one database may perform poorly on another, indicating a lack of generalization capacity. Moreover, the evaluation protocol has a noticeable impact on the outcomes: certain metrics consistently produce better scores when applied on cropped image patches, while others achieve superior performance when using resized images. This divergence highlights the sensitivity of existing approaches to both dataset properties and preprocessing strategies, and it underscores the necessity of considering multiple evaluation settings to obtain a reliable and comprehensive assessment of model robustness.

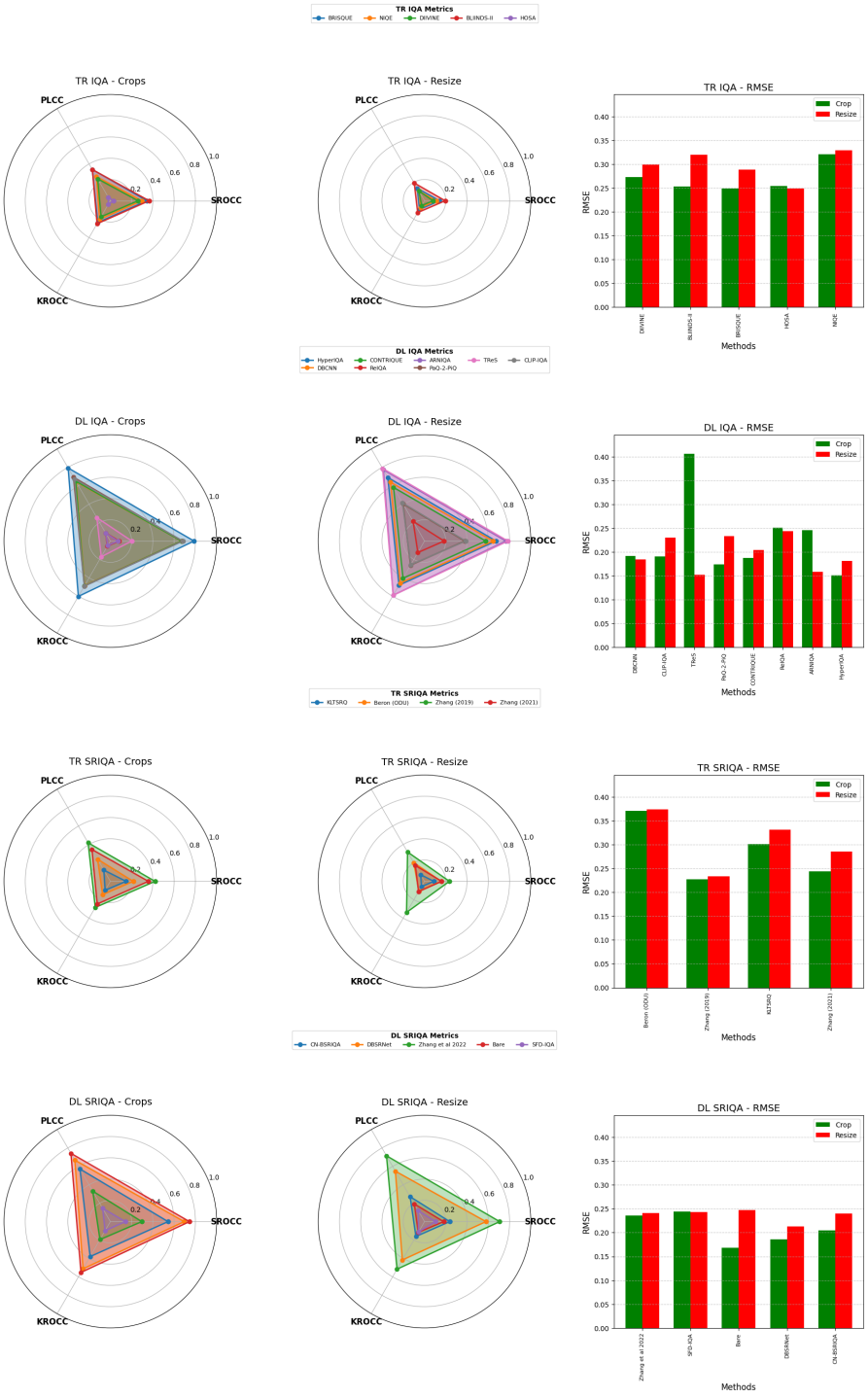


Figure 3.1: *Performance comparison of IQA and SR-IQA metrics on the CVIU-2017 dataset.*

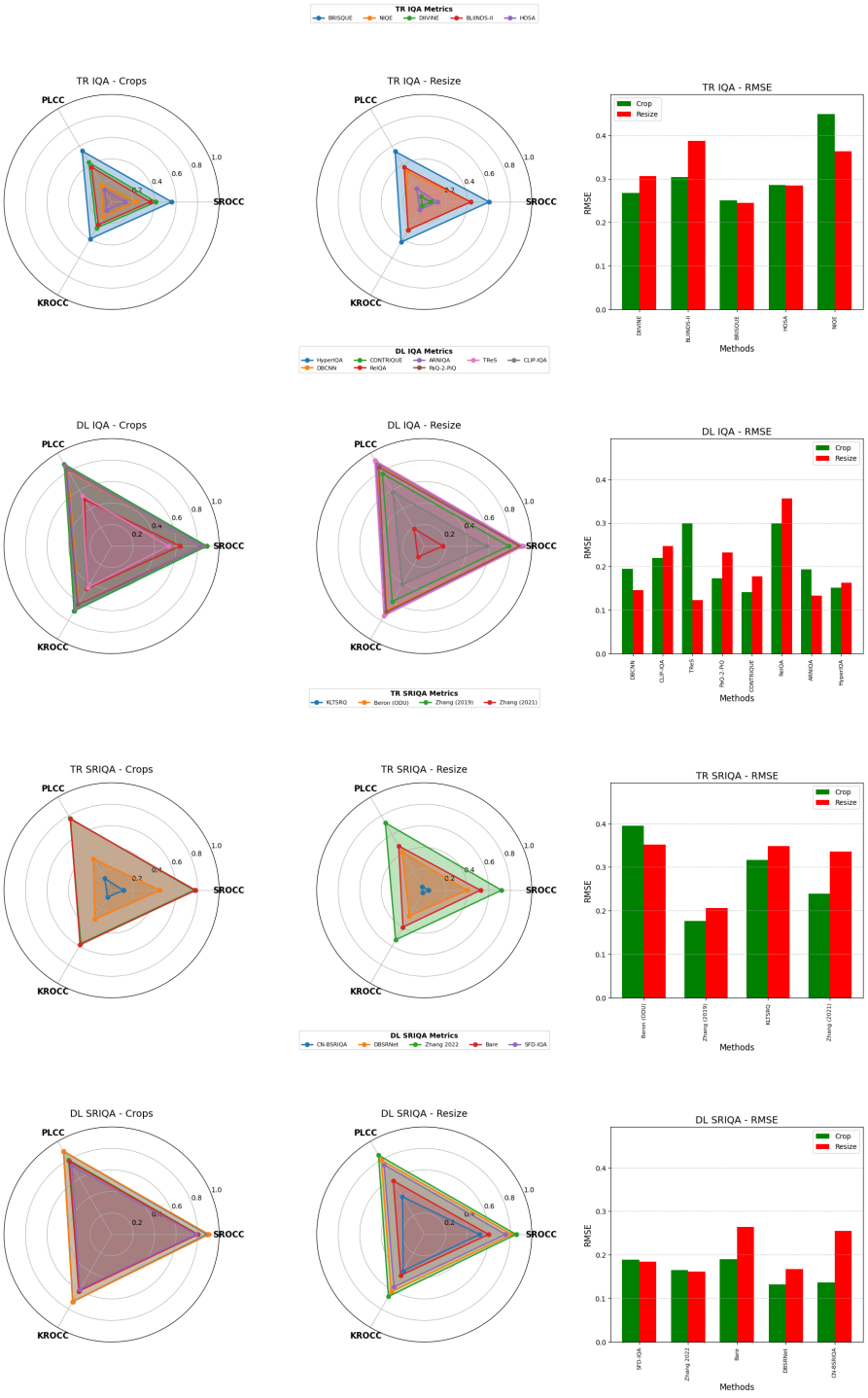


Figure 3.2: *Performance comparison of IQA and SR-IQA metrics on the QADS dataset.*

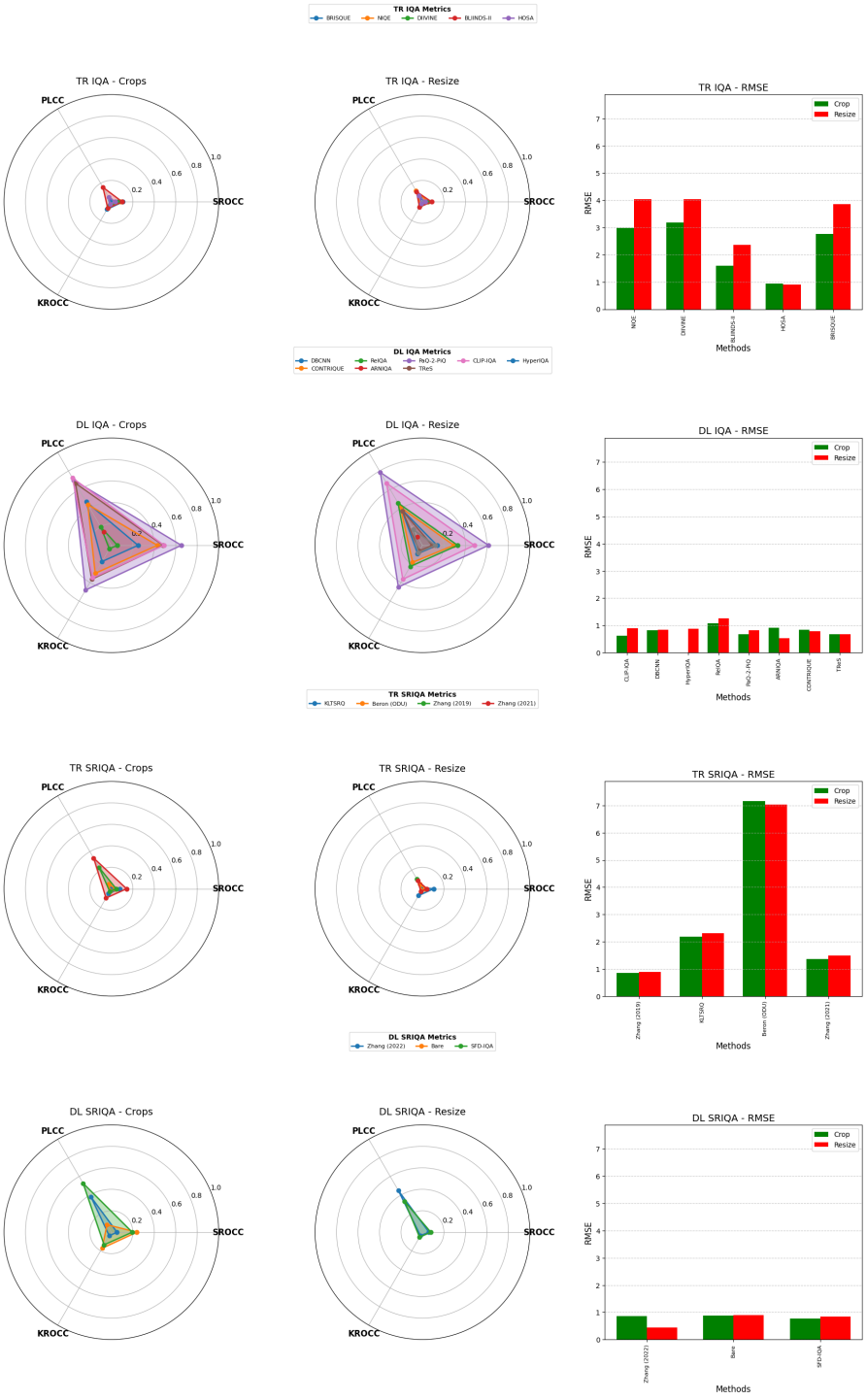


Figure 3.3: *Performance comparison of IQA and SR-IQA metrics on the RealSRQ dataset.*

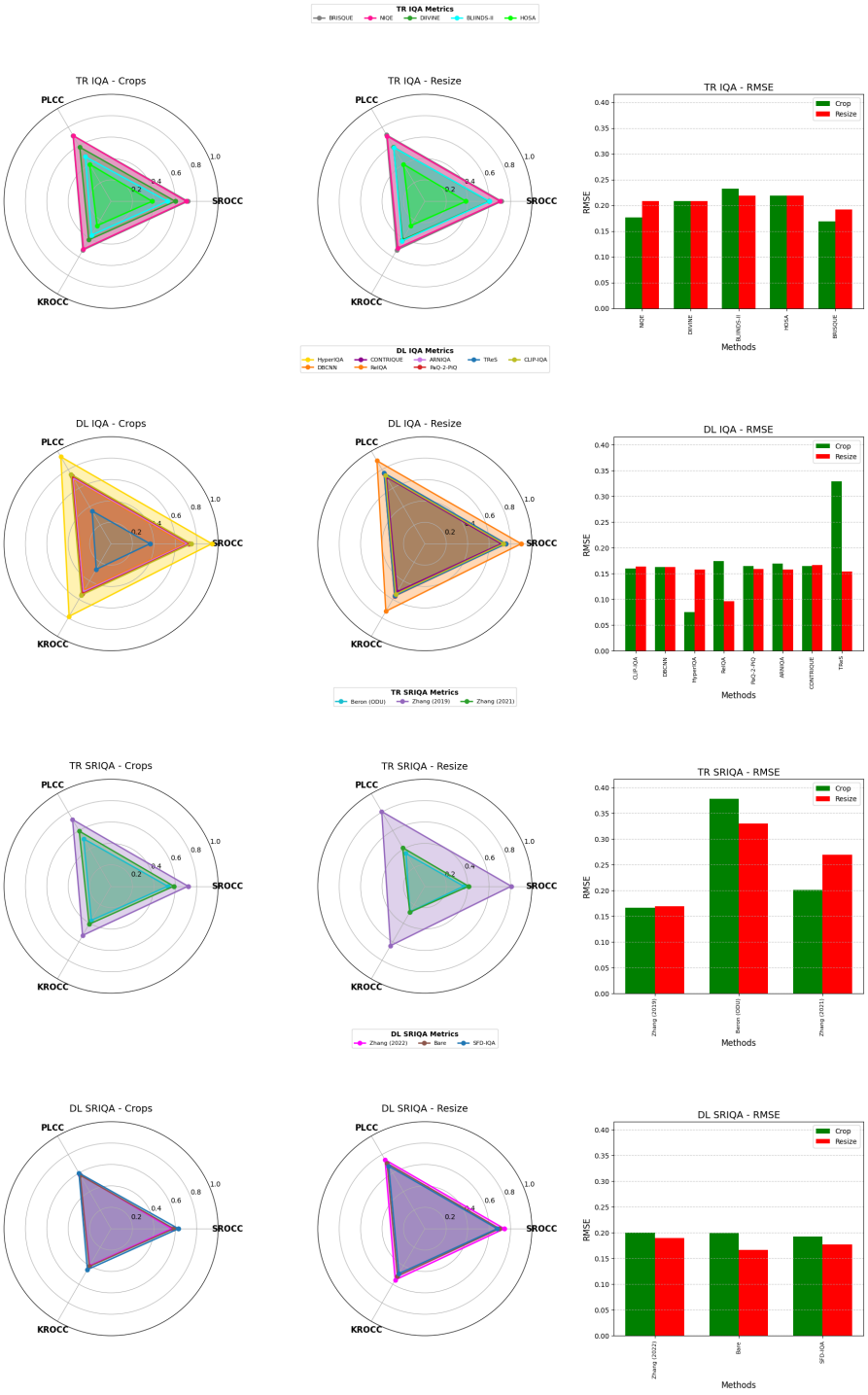


Figure 3.4: *Performance comparison of IQA and SR-IQA metrics on the SISAR dataset.*

To further evaluate the behaviour of different NR-IQA methods on the CVIU dataset, Figures 3.5 and 3.6 show the scatter plots comparing the predicted quality scores against the ground-truth subjective scores. Each scatter plot corresponds to a different metric. Ideally, for a reliable quality assessment method, the data points should closely align along the diagonal line, which represents a perfect correlation between predicted and subjective scores.

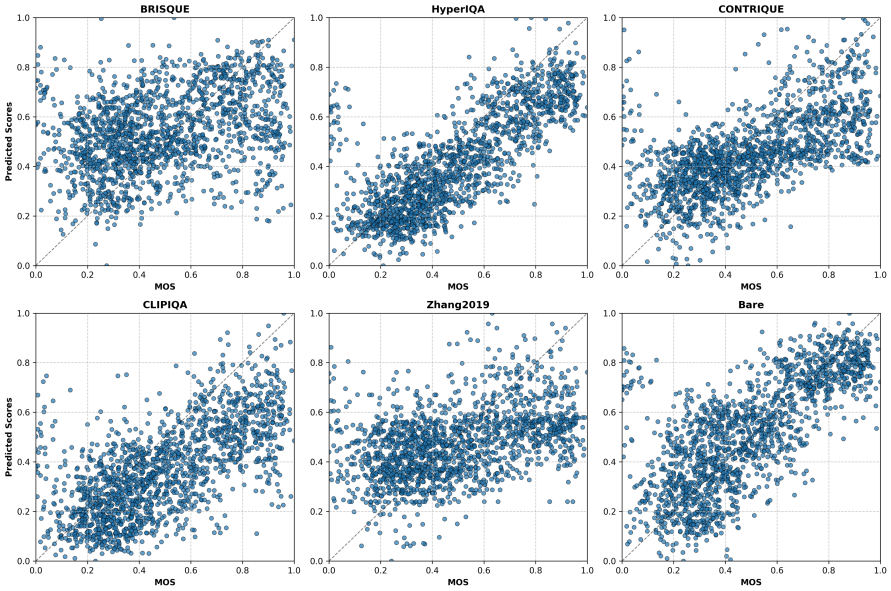


Figure 3.5: *Example of scatter plots showing the correlation between predicted NR-IQA scores and subjective scores on the CVIU dataset*

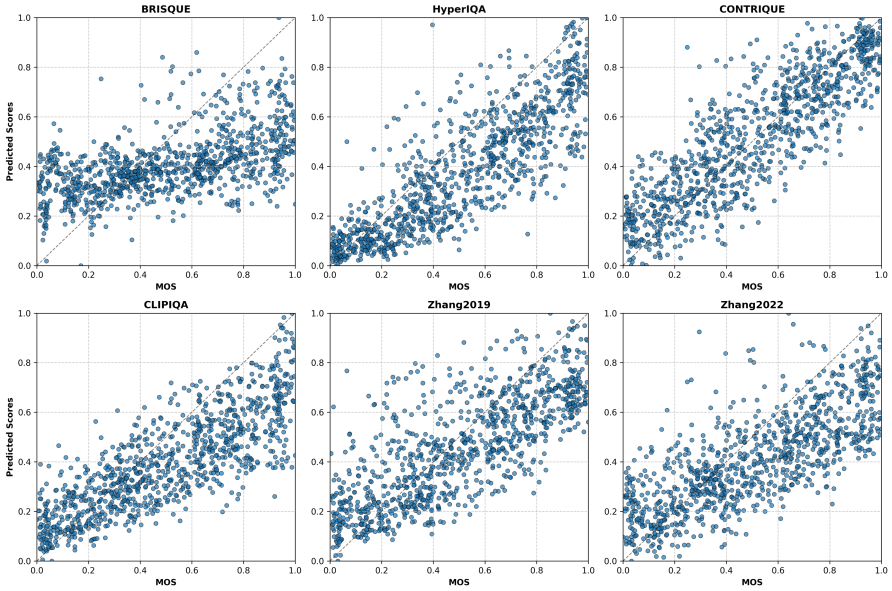


Figure 3.6: *Example of scatter plots showing the correlation between predicted NR-IQA scores and subjective scores on the QADS dataset*

However, as illustrated in the scatter plots, none of the methods exhibit a perfect alignment with the diagonal. While some metrics show a clearer trend and stronger correlation (with data points clustering more tightly around the diagonal), others present a higher degree of dispersion, suggesting weaker consistency with human subjective perception. This lack of alignment indicates that the predictive ability of these NR-IQA metrics is limited and that they may fail to capture certain perceptual aspects of image quality specific to the CVIU dataset.

These observations highlight the challenges in designing robust NR-IQA methods. They further emphasize the need for models that can better generalize across diverse datasets and more effectively approximate subjective human judgments of quality.

In contrast to the results obtained on the CVIU dataset, the scatter plots for the QADS dataset reveal a slightly improved alignment between the predicted scores and the ground-truth subjective quality values. In particular, the Contrique-based method demonstrates a clearer trend along the diagonal line, indicating a stronger correlation with human perception compared to other approaches.

Although the alignment is not perfect, the reduction in dispersion suggests that Contrique is able to capture some of the perceptual factors reflected in QADS more effectively than in CVIU. This partial alignment highlights its potential to generalize better under certain conditions and confirms that dataset characteristics play a crucial

role in evaluating the robustness of NR-IQA models.

3.1.2. Statistical significance analysis

Table 3.1: *T-test results for the statistical comparison of image quality assessment methods on the CVIU-2017 database (SROCC).*

	BRISQUE	NIQE	DIIVINE	BLIINDS-II	HOSA	HyperIQA	DBCNN	CONTRIQUE	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	KLTSRQ	Beron(ODU)	Zhang (2019)	Zhang (2021)	CN-BSRIQA	DBSRNet	Zhang (2022)	Bare	SFD-IQA
BRISQUE																							
NIQE																							
DIIVINE																							
BLIINDS-II																							
HOSA																							
HyperIQA																							
DBCNN																							
CONTRIQUE																							
ReIQA																							
ARNIQA																							
NIMA																							
PaQ-2-PiQ																							
TReS																							
CLIP-IQA																							
KLTSRQ																							
Beron (ODU)																							
Zhang (2019)																							
Zhang (2021)																							
CN-BSRIQA																							
DBSRNet																							
Zhang (2022)																							
Bare																							
SFD-IQA																							

Table 3.2: *T-test results for the statistical comparison of image quality assessment methods on the QADS database (SROCC).*

	BRISQUE	NIQE	DIIVINE	BLIINDS-II	HOSA	HyperIQA	DBCNN	CONTRIQUE	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	KLTSRQ	Beron(ODU)	Zhang (2019)	Zhang (2021)	CN-BSRIQA	DBSRNet	Zhang (2022)	Bare	SFD-IQA
BRISQUE																							
NIQE																							
DIIVINE																							
BLIINDS-II																							
HOSA																							
HyperIQA																							
DBCNN																							
CONTRIQUE																							
ReIQA																							
ARNIQA																							
NIMA																							
PaQ-2-PiQ																							
TReS																							
CLIP-IQA																							
KLTSRQ																							
Beron (ODU)																							
Zhang (2019)																							
Zhang (2021)																							
CN-BSRIQA																							
DBSRNet																							
Zhang (2022)																							
Bare																							
SFD-IQA																							

In addition to the correlation analysis and to rigorously validate the performance of the considred IQA methods, we conduct a statistical significance analysis using a repeated cross-validation procedure with 50 random splits and evaluations on image crops. This is motivated by the fact that average correlation scores, such as SROCC, is insufficient, since variations due to the random partitioning of training and test sets may influence the outcomes.

Specifically, for each pair of IQA methods, the 50 SROCC scores were compared using a two-tailed *"t-test"*, where the null hypothesis stated that ***no statistically significant difference exists between the two mean performances***, and this hypothesis is rejected at a significance level of $p < 0.05$. The results of these pairwise comparisons are summarized in the matrices shown in Tables 3.1 and 3.2 for the CVIU-2017 and QADS databases respectively, where green cells indicate statistically significant differences, red cells denote non-significant differences, and white cells represent no difference.

The analysis of these matrices highlights several consistent trends: recent methods such as SFD-IQA, Bare, DBSRNet and Zhang *et al.* (2022) demonstrate statistically

significant superiority over most earlier approaches, while groups of classical algorithms like DIIVINE, BLIINDS-II and HOSA often yield indistinguishable results, reflecting shared conceptual limitations that deep learning-based methods have overcome. Furthermore, the consistency of trends across both databases strengthens the external validity of the conclusions. A particularly instructive example is the comparison between NIQE and TReS, where despite a substantial mean SROCC gap of 0.2985, the *"t-test"* reveals no statistically significant difference, underscoring that the apparent advantage is offset by high variance in the results across the 50 splits. This paradox illustrates the importance of performance stability: statistical testing considers not only mean differences but also variance, and high instability leads to overlapping performance distributions that prevent systematic superiority from being established. Consequently, these findings emphasize the limitations of relying on a single average score and highlight the necessity of statistical tests, as a method with a high mean performance but large variance may be unreliable in real-world conditions.

3.1.3. Computational complexity analysis

The table 3.3 compares several Image Quality Assessment (IQA) methods based on the number of parameters, computational cost (FLOPs), and inference time. The analysis of Table reveals important insights into the relationship between model complexity and performance in Image Quality Assessment (IQA). Traditional methods such as BRISQUE, NIQE, and DIVINE are characterized by the absence of trainable parameters and extremely low computational cost, resulting in very short inference times (e.g., less than 0.01 seconds). While these approaches are attractive for scenarios where efficiency is the primary requirement, their reliance on hand-crafted features limits their ability to capture complex distortions, leading to reduced prediction accuracy compared to modern deep learning-based approaches.

In contrast, very large deep neural networks, such as Zhang (2021, 2022) and DBSRNet, involve hundreds of millions of parameters and high FLOPs (exceeding 200 GFLOPs in some cases), which translates into inference times ranging from several seconds to nearly ten seconds per image. These models tend to achieve superior accuracy due to their representational power, but their heavy computational demands restrict their use to offline analysis or research contexts where real-time execution is not critical.

Between these two extremes, a number of lightweight deep learning models, including NIMA, PaQ-2-PiQ, and CN-BRISQUE, offer a more favorable balance. With parameter counts typically under 15M and inference times below 0.3 seconds, they provide significantly improved accuracy compared to traditional methods while remaining computationally

efficient. This makes them attractive candidates for real-time or resource-constrained applications such as mobile devices and online image processing systems.

A further category is represented by intermediate models such as HyperIQA, ARNIQA, DBCNN, CONTRIQUE, and ReIQA. These models, with parameter counts in the range of 20–70M and FLOPs between 20–200G, present moderate inference times (often between 0.7 and 2.5 seconds). They provide a strong compromise, combining robust predictive accuracy with a computational cost that remains acceptable for many practical applications.

Finally, recent approaches such as SFD-IQA and Bare demonstrate a trend toward more efficient architectures that achieve competitive accuracy with relatively low complexity. For example, SFD-IQA achieves inference in just 0.157 seconds with only 7.75 GFLOPs, suggesting that architectural optimizations and efficient design principles can substantially improve the trade-off between performance and efficiency.

Table 3.3: *Comparison of various Image Quality Assessment (IQA) models in terms of number of parameters (in millions), computational complexity (measured by FLOPs in GigaFLOPs), and inference time (in seconds).*

Metric	# Params (M)	FLOPs (G)	Inference time (s)
BRISQUE	-	-	0.041
NIQE	-	-	0.084
DIVIINE	-	-	4.5698
BLIINDS-II	-	-	6.9229
HOSA	1.02	3.0424	0.6707
HyperIQA	26.89	68.81	0.9496
DBCNN	15.31	264.05	1.7958
CONTRIQUE	29.02	89.45	0.9803
ReIQA	47.54	16.44	0.2269
ARNIQA	28.51	16.44	0.2203
NIMA	3.24	1.15	0.0237
PaQ-2-PiQ	11.70	3.64	0.0572
TReS	11.32	38.01	0.3461
CLIP-IQA	59.23	17.66	0.8291
KLTSRQ	-	-	0.1519
Beron (ODU)	-	-	0.3352
Zhang (2019)	-	-	0.7143
Zhang (2021)	138.38	15.47	0.2499
CN-BSRIQA	0.97	1.42	0.0676
DBSRNet	25.56	4.15	0.2735
Zhang (2022)	93.78	247.79	3.9820
Bare	1.15	2.72	0.0625
SFD-IQA	0.39	7.75	0.1577

3.1.4. Performance analysis of the proposed SR-IQA method

The proposed SR-IQA method achieves the best correlation results on both the QADS and CVIU-2017 datasets, as shown in Figure 3.7. On QADS, it reaches state-of-the-art performance with SROCC = 0.9523, PLCC = 0.9447, KROCC = 0.8095, and RMSE = 0.1172. On CVIU-2017, it also outperforms competing approaches, obtaining SROCC = 0.8066, PLCC = 0.8068, KROCC = 0.6435, and RMSE = 0.1535. To further validate robustness and generalization, additional experiments were conducted on the RealSRQ dataset. For the RealSRQ dataset, the proposed method ranks among the top-performing approaches, achieving SROCC = 0.7260, PLCC = 0.7300, KROCC = 0.5547, and RMSE = 0.7915. These results confirm both the superiority of the proposed approach on QADS and CVIU-2017 and its strong competitiveness on RealSRQ.

To further analyze prediction behavior, Figure 3.8 presents scatter plots of MOS versus predicted scores. In both QADS and CVIU-2017, the data points exhibit a clear alignment with the diagonal, reflecting strong consistency with human perceptual judgments. On QADS, the points are tightly clustered around the diagonal, showing excellent agreement with MOS values. On CVIU-2017, although a slightly higher dispersion is observed, the scatter plot of the proposed method shows a stronger alignment with the diagonal compared to other top-performing methods, confirming that the predicted scores remain highly consistent with subjective evaluations.

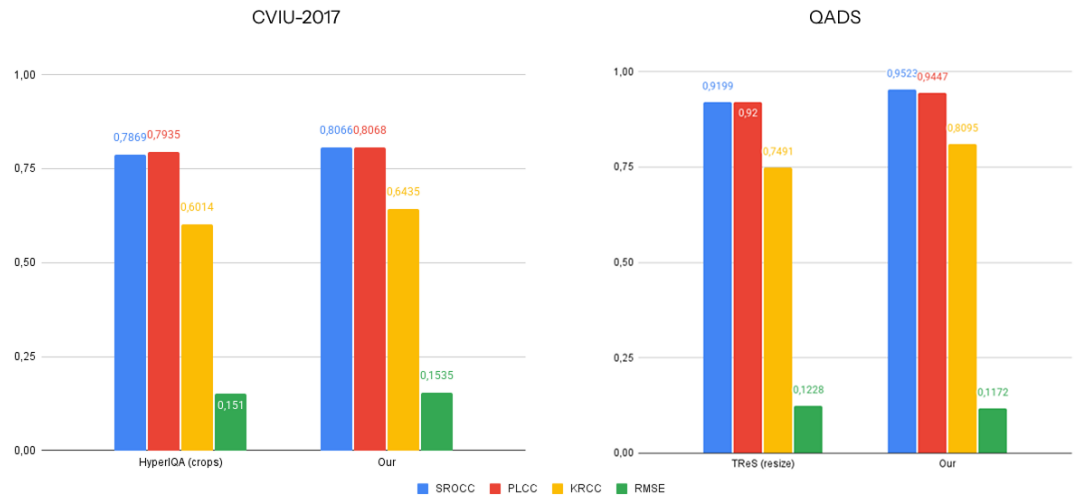


Figure 3.7: *Bar plot comparing the correlation performance of the best existing methods and the proposed SR-NR-IQA on the CVIU-2017 and QADS datasets.*

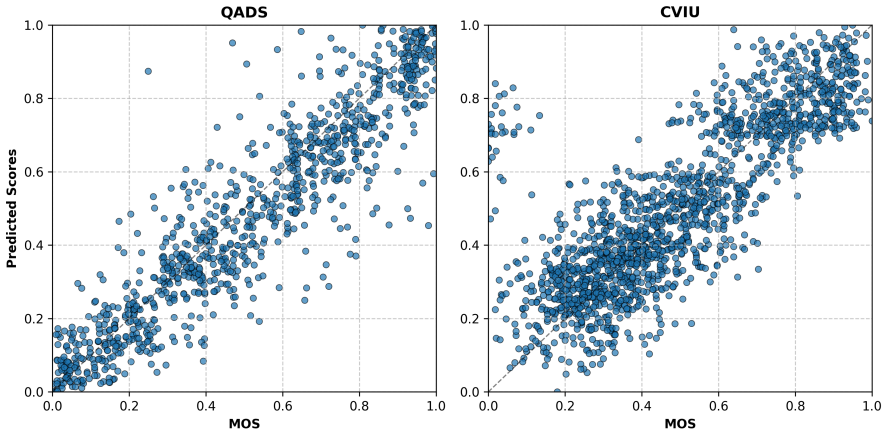


Figure 3.8: *Scatter plot of the proposed SR-NR-IQA method on the QADS and CVIU-2017 datasets.*

3.1.5. Overall Interpretation of the Benchmark Results

Traditional IQA methods, such as BRISQUE, NIQE, DIIVINE, and BLIINDS-II, generally exhibit low correlation with subjective scores (MOS). This limitation stems from the fact that these approaches are not based on human perceptual data but rely instead on natural scene statistics extracted from images, leading to a mismatch between the objective scores they produce and actual human visual perception. In contrast, recent deep learning-based approaches—such as HyperIQA, DBCNN, CONTRIQUE, and PaQ-2-PiQ—are trained directly on subjective judgments. As a result, they are better able to learn visual features relevant to human perception and achieve significantly higher correlations with MOS scores. Furthermore, a distinction must be made between general-purpose IQA methods and those specifically designed for super-resolution image quality assessment (SR-IQA). While general IQA models can detect common visual artifacts such as blur, noise, and compression, SR-IQA methods utilize architectures that are tailored to identify artifacts specific to super-resolution, particularly errors in structure and texture that appear in fine image details. Although SR-IQA techniques have shown improved performance on dedicated datasets, achieving high and generalizable correlation with MOS scores remains a major challenge due to the diversity of artifacts and the inherently subjective nature of visual quality assessment.

In this context, image preprocessing strategies also play a critical role. Resizing images, a common step to standardize input dimensions for models, may alter the original

structure and proportions of the image, potentially introducing or amplifying visual artifacts—especially those related to super-resolution as can be seen in Figure 2.2. While this method allows for global assessment of the image and may benefit some models, it can negatively impact others, particularly those sensitive to structural distortions or fine details, ultimately resulting in poor correlation with MOS scores. Conversely, the cropping strategy offers several advantages. It allows for effective dataset augmentation without artificially modifying the content, enabling more stable training and reducing the risk of overfitting. Moreover, by performing localized assessment through a sliding window approach, it captures regional details that are often critical for human perception. This technique adapts well to deep learning models while preserving the integrity of the original visual content. Altogether, these strategies contribute to the development of models that produce predictive scores more closely aligned with human perception, thereby improving the consistency between predicted scores and subjective MOS ratings.

3.2. Interpretability Results of SR Models

The interpretability analysis is conducted on a diverse set of representative super-resolution models. The selection includes convolutional architectures such as EDSR, CARN, SAN, and RCAN; GAN-based models including SRGAN and RRDBNet (as employed in ESRGAN); as well as the transformer-based approach SwinIR. These models incorporate different mechanisms, ranging from residual learning and channel attention to adversarial training and shifted-window transformers. This diversity enables a comprehensive evaluation of interpretability, allowing us to investigate how architectural choices influence both the reconstruction quality and the visual evidence exploited by the models.

Figure 3.9 compares several super-resolution (SR) models by visualizing their ability to restore a textual region from a low-resolution (LR) image. Each model generates an SR version with a highlighted region of interest, accompanied by a saliency map indicating the most influential visual areas for the model. The saliency maps show that models like RCAN, SAN, and RRDBNet focus more accurately on key characters, reflecting better interpretability and reconstruction precision. In contrast, models such as CARN and SwinIR exhibit more diffuse attention, suggesting less targeted focus. The joint analysis of SR images and saliency maps thus allows for the evaluation of not only the visual quality of the reconstructions but also the relevance of the regions exploited by the models to produce these results.

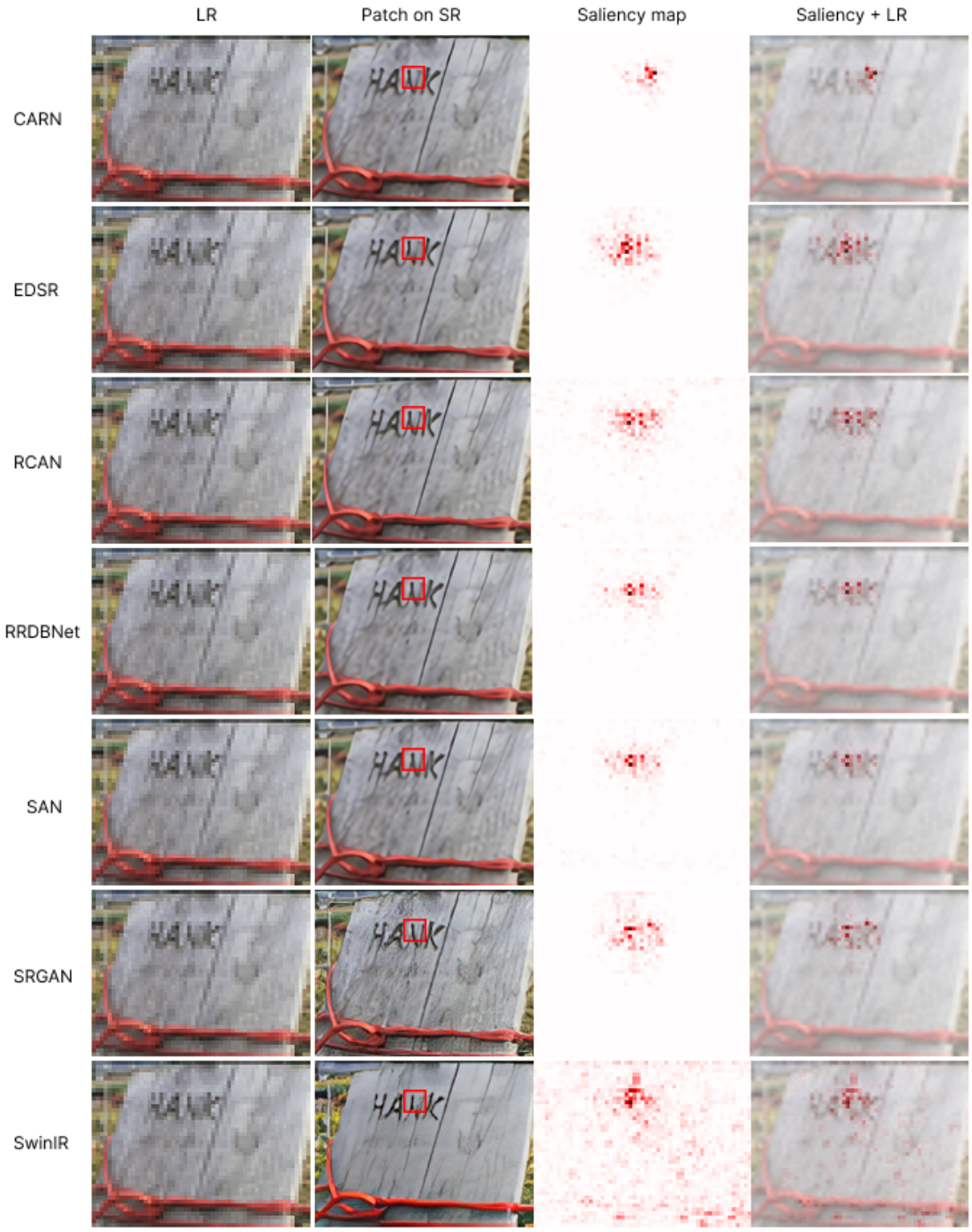


Figure 3.9: *Comparison of SR models with saliency map visualization on a text region*

Building upon this qualitative observation, the figure 3.10 presents the Diffusion Index

(DI) values for the same SR models, providing a quantitative measure of the extent to which each model focuses on discriminative features during reconstruction. A higher DI reflects a sharper and more targeted concentration on salient details, complementing the interpretability insights derived from the saliency maps. Consistent with the visual analysis, SwinIR achieves the highest DI (39.02), confirming its strong ability to capture text-specific details, while RCAN and SAN also show robust performance with DI scores of 25.08 and 19.93, respectively. By contrast, CARN (3.28) and EDSR (5.3) record substantially lower values, which align with their more diffuse and less precise attention patterns. The integration of qualitative saliency visualization and quantitative DI measurement therefore, provides a more comprehensive framework for assessing both the interpretability and the effectiveness of SR models.

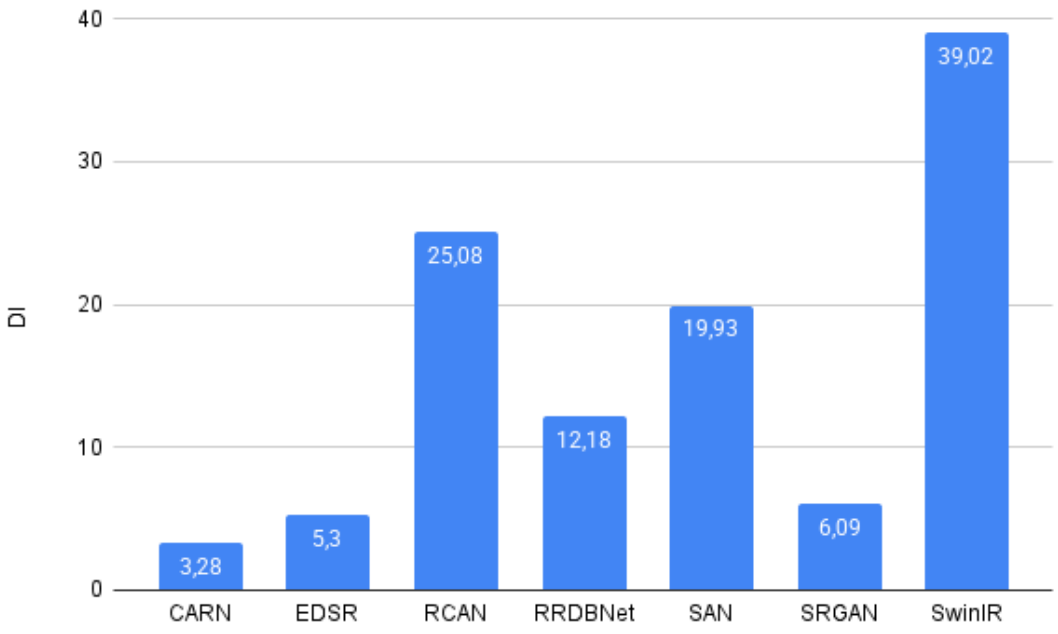


Figure 3.10: *diffusion Index (DI) scores of various SR models on textual region analysis.*

Figure 3.11 presents the saliency maps generated by the GL-AM module across several super-resolution architectures, providing a visual analysis of how each model allocates attention to critical regions in the image. Although GL-AM is explicitly designed to enhance attention by integrating global contextual information with local details, the comparison of saliency maps across different architectures reveals only minor variations.

Across all models, the attention consistently concentrates on key regions such as textual elements and surrounding textures, with remarkably similar spatial patterns. This visual consistency suggests that GL-AM imposes a dominant and uniform attention mechanism that overrides architecture-specific differences, effectively standardizing the focus of the models. Consequently, it becomes challenging to attribute differences in reconstruction performance solely to the underlying network architecture, as the primary influence on attention behavior appears to stem from the GL-AM module itself.

The quantitative analysis presented in Figure 3.12 supports this observation, showing that the Diffusion Index (DI) values for all architectures are closely clustered. The limited variance in DI confirms that GL-AM generates a highly stable and reproducible attention distribution, minimizing the effect of architectural variations. This convergence between qualitative saliency visualizations and quantitative DI metrics highlights a key implication: while GL-AM significantly improves the robustness and reliability of attention across models, it also constrains the natural diversity of attention strategies that different architectures might exhibit.

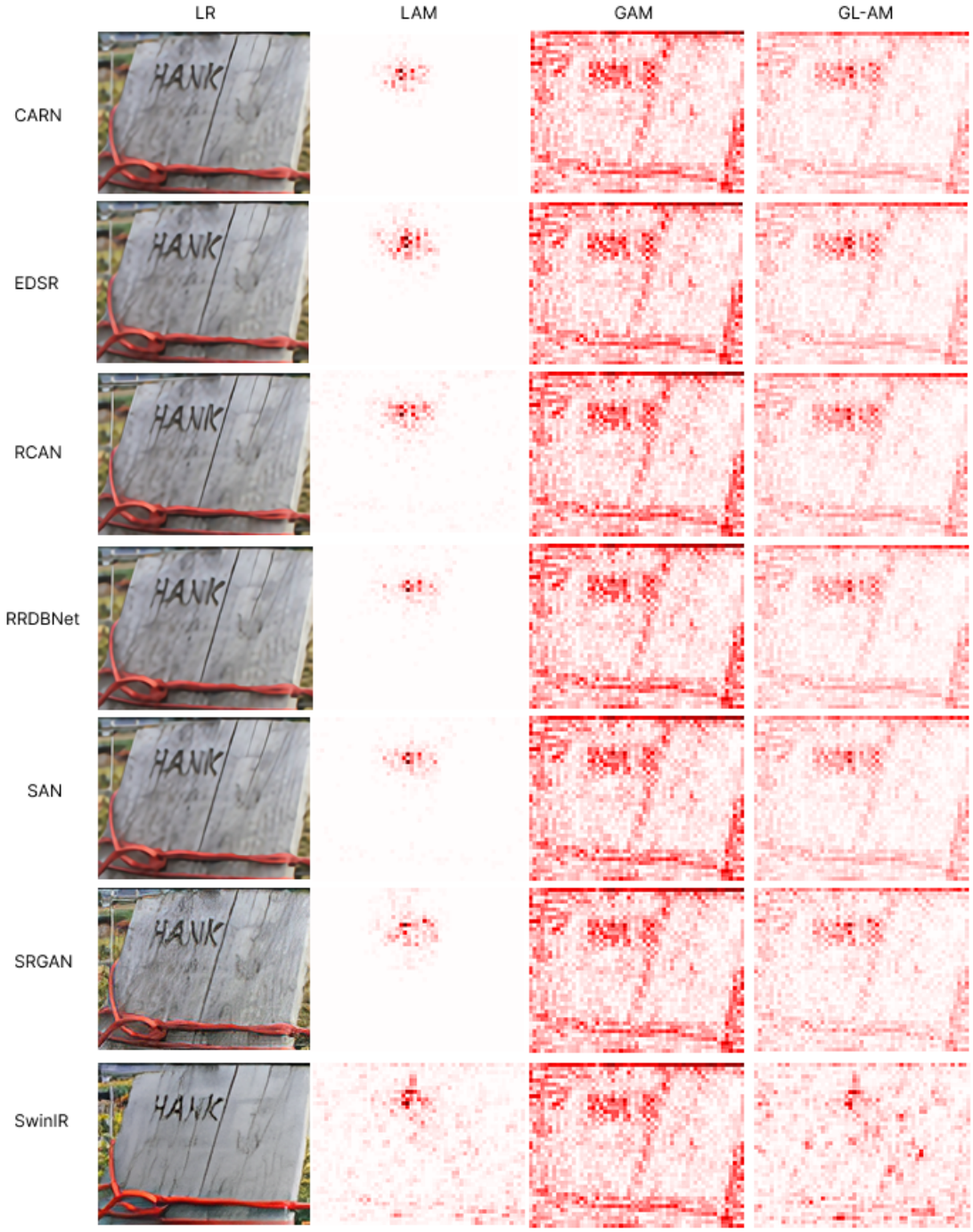


Figure 3.11: *GL-AM attention maps across different super-resolution models, showing effective integration of local details and global context for improved image reconstruction.*

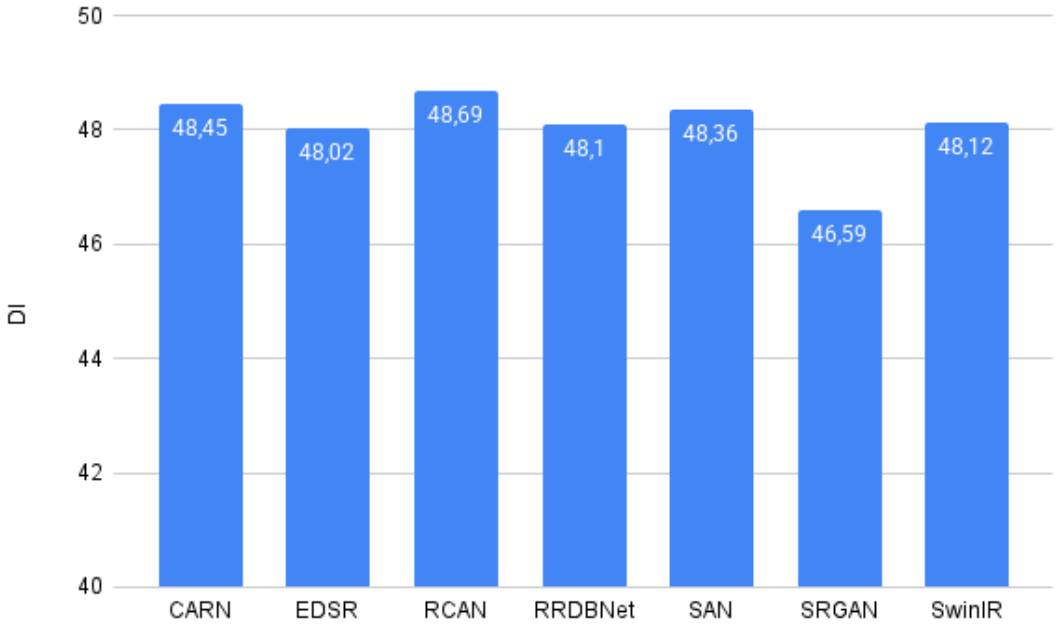


Figure 3.12: *Diffusion Index (DI) Scores of Various SR Models with GL-AM.*

In Figure 3.13, the Relevance Attention Map (RAM) is generated by applying a variety of realistic degradation operators, including blur, noise, compression, and other distortions, in order to simulate diverse and challenging image restoration scenarios. The RAM highlights which regions of the low-resolution input the super-resolution models rely on when reconstructing high-frequency details under these complex conditions. Compared to the Local Attention Map (LAM), which is computed primarily under a single blur degradation and tends to emphasize a limited set of key areas, RAM reveals that the models engage with a broader spectrum of image features. Specifically, when additional distortions are introduced, the attention distribution spreads across larger spatial regions, suggesting that the models adaptively leverage contextual information from both the central and peripheral areas of the image to compensate for various artifacts. This expanded focus indicates a more sophisticated feature utilization strategy, where the models do not rely solely on localized high-frequency cues but incorporate global and mid-level contextual information to achieve robust super-resolution. The comparison between RAM and LAM underscores an important insight: while LAM provides a simplified view of attention under controlled conditions, RAM exposes the adaptive and dynamic nature

of the models' attention mechanisms in realistic scenarios, highlighting their capacity to integrate multiple sources of information when confronted with diverse degradations. Consequently, RAM not only provides a more comprehensive understanding of model behavior under real-world conditions but also emphasizes the limitations of evaluating attention patterns using only simplified or single-degradation setups.

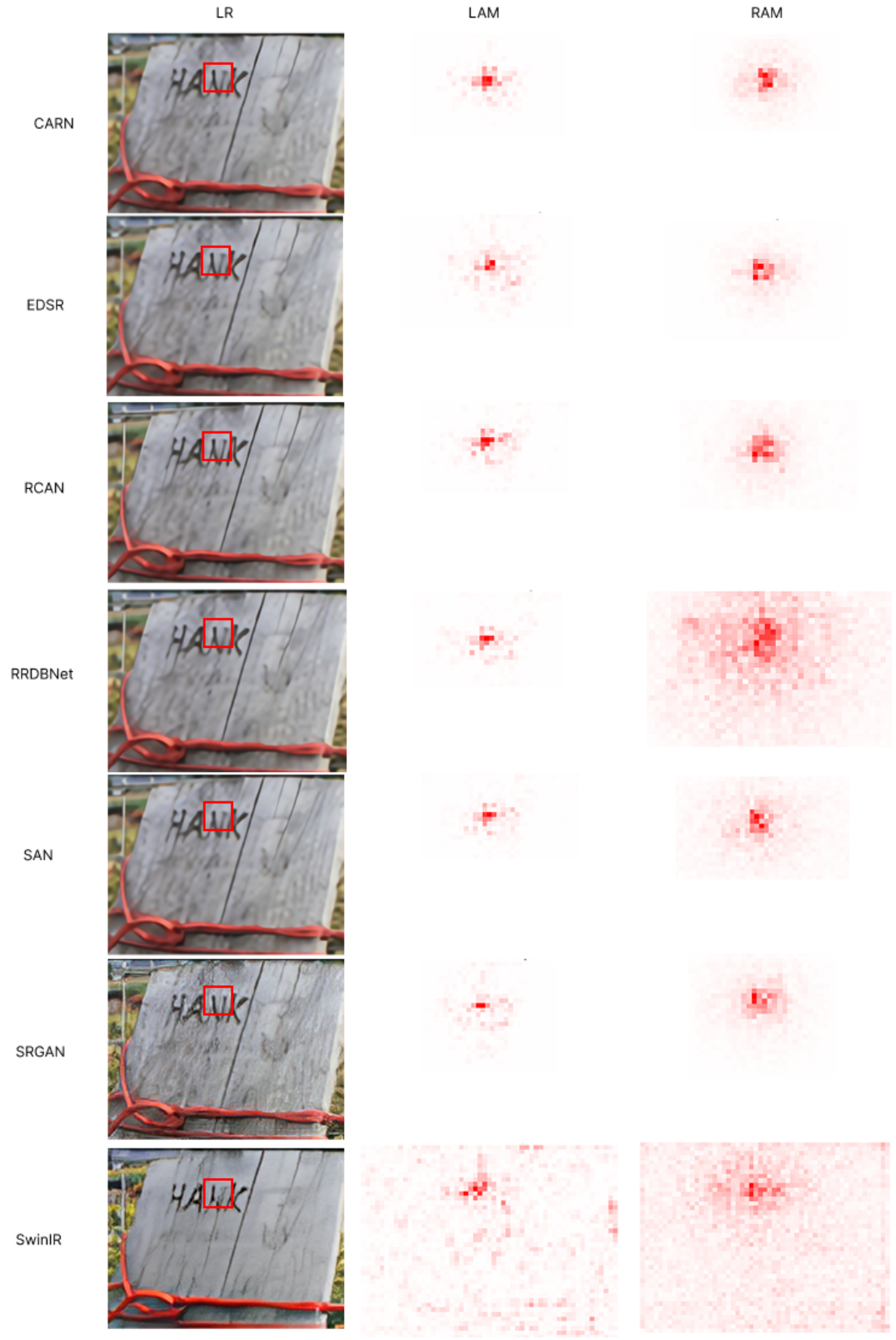


Figure 3.13: *Comparison of Local Attribution Map (LAM) and Real Attribution Map (RAM) for different super-resolution models.*

Figure 3.14 presents a comprehensive comparison of no-reference (NR-IQA) and full-reference (FR-IQA) image quality assessment metrics on the QADS dataset, incorporating both the SROCC and PLCC correlation coefficients alongside the Diffusion Index (DI) as a measure of attention dispersion and feature utilization. The FR-IQA metrics, whose performance scores are reported in the Local Attribution Maps study, consistently achieve higher correlations with the mean opinion scores, confirming their reliability and accuracy when a pristine reference image is available. This high correlation reflects the inherent advantage of FR-IQA methods, which can directly quantify deviations from the reference and thus more precisely capture perceptual distortions. By contrast, NR-IQA methods on QADS exhibit considerably weaker correlations, even when evaluated with the additional context provided by the DI. This disparity highlights the fundamental challenge of predicting perceptual quality without a reference, as NR-IQA metrics must infer image fidelity and artifact impact solely from the degraded input.

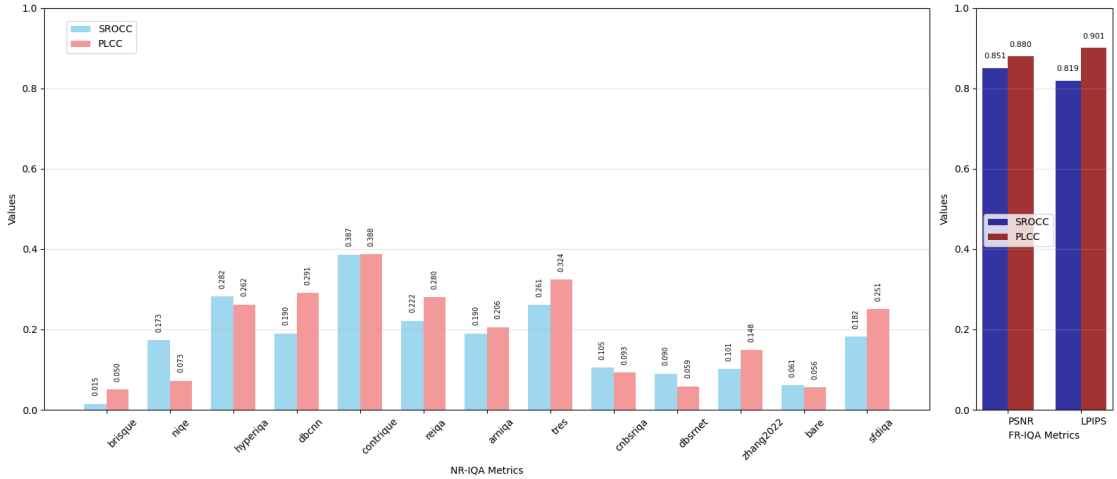


Figure 3.14: *Comparison of NR-IQA metrics on the QADS dataset and FR-IQA metrics with diffusion index based on SROCC and PLCC correlations.*

3.2.1. Analysis of the Contributions of Interpretability

The integration of interpretability has provided valuable insights into the internal mechanisms of SR models. Attribution maps have made it possible to identify the models' areas of attention, revealing potential biases, implicit optimization strategies, and

limitations in generalization. These contributions reinforce the idea that interpretability is not merely a post-hoc analysis tool, but a strategic lever for model development, debugging, and continuous improvement. Despite that, LAM argues that methods using a large amount of information, such as SwinIR, achieve better results. In contrast, methods relying on limited information often fail to accurately reconstruct textures. However, this claim is not always valid. For instance, Figure 3.9 shows that SwinIR produces a very smooth image with many distortions, while SRGAN manages to recover more structural and textural details, despite using less information. Furthermore, LAM states that the model selects important pixels that strongly influence the result, but it does not specify whether these pixels are actually the most relevant.

The diffusion index used for quantitative interpretation is based solely on the importance assigned to pixels, without ensuring their true relevance. In addition, LAM uses only blur as a criterion for evaluating reconstruction quality, which is a limitation: blur alone is not sufficient to assess the fidelity or structural consistency of the generated textures. It would therefore be more appropriate to consider other types of distortions for a more comprehensive evaluation. However, LAM also does not provide a clear explanation of how, or based on what information, models like SRGAN manage to generate such detailed outputs despite having limited data, and it remains uncertain whether the added information can truly be trusted.

3.3. Conclusion

In this chapter, we evaluated super-resolution models from two complementary perspectives: the perceived image quality using no-reference image quality assessment (NR-IQA) metrics, and the interpretability of the models through attribution analysis.

The NR-IQA evaluation demonstrated that certain metrics can effectively reflect image quality without reference images. Their performance, however, depends on both the type of SR model and the nature of distortions. Comparative analyses using tables and graphs highlighted variations across models and helped identify the best-performing approaches. These results also emphasize that the robustness and generalization capability of NR-IQA methods are crucial for reliable evaluation.

The interpretability analysis, particularly via attribution maps, provided deeper insight into the internal functioning of SR models. Visualizing attention areas revealed how models reconstruct details, highlighted potential biases, and allowed comparisons between different network types (CNNs, GANs, Transformers). These findings confirm that interpretability

is a valuable tool for guiding model improvement and debugging.

Despite these insights, certain limitations remain, including the diversity of distortions and the number of models studied, which call for further research to extend and validate the present findings. Overall, the combination of NR-IQA evaluation and interpretability analysis provides a comprehensive framework for assessing and understanding super-resolution models.

General conclusion

This thesis focused on the evaluation of no-reference image quality assessment (NR-IQA) methods applied to SR, as well as on the exploration of interpretability in SR models through explainable AI techniques. It offered a critical overview of existing approaches, presented in-depth analyses of their performance and limitations, and opened up concrete avenues for improving the reliability and transparency of SR systems.

The comparative benchmark of NR-IQA models revealed a clear superiority of deep learning-based approaches over traditional methods relying on natural scene statistics. These newer models, trained on subjective human judgments, better capture visual features relevant to human perception, resulting in stronger correlations with MOS scores. However, even state-of-the-art general-purpose NR-IQA models struggle to detect artifacts specific to super-resolution. In contrast, SR-specific IQA models show promising improvements but still face challenges due to the diversity of artifacts and the inherently subjective nature of perceived quality.

Moreover, image preprocessing strategies proved critical: resizing can degrade the structural fidelity of images, while cropping allows for more effective data augmentation and better preservation of local details, which are often key to human perception.

Finally, the integration of interpretability methods provided valuable insights into the internal behavior of SR models. Attribution maps helped identify attention regions, revealing biases, implicit optimization strategies, and generalization limitations. This confirmed that interpretability is not merely a post-hoc diagnostic tool, but a strategic asset for model development and continuous improvement.

Limitations of the Study

Despite significant advancements, NR-IQA continues to face several inherent limitations. One of the primary challenges lies in accurately predicting perceived quality based solely on the distorted image. This intrinsic complexity makes the development of reliable NR-IQA models highly demanding. Furthermore, although numerous algorithms have been proposed, the field remains largely underexplored, with many fundamental issues still unresolved [80]. The process of extracting meaningful features and effectively mapping them to subjective quality scores remains a major obstacle, particularly for general-purpose

applications. Additionally, the performance of existing NR-IQA methods often varies across different benchmark datasets, raising concerns about their generalizability and robustness.

High-resolution image datasets also present significant challenges due to their limited availability and insufficient size for training deep learning models effectively. The lack of diverse artifacts further restricts robust learning. The small size of these datasets is primarily due to the nature of subjective testing, which is time-consuming, costly, and requires numerous participants under controlled conditions. Expert assessments, while more accurate, add further complexity to the data collection process.

Moreover, even state-of-the-art NR-IQA models struggle with detecting SR-specific artifacts, which vary widely and are highly context-dependent. Preprocessing strategies such as resizing and cropping introduce additional complexity, sometimes degrading structural fidelity or biasing the evaluation process.

The interpretation of super-resolution models remains a major challenge. Existing interpretability approaches struggle to adapt to the complexity of degradations and the diversity of structures, textures, and patterns in input data. The regions highlighted by attribution methods do not always reflect the true decision-making processes of the network. Instead, they are often influenced by architectural biases or optimization artifacts, undermining both trust and reliability. Currently, there is no universally accepted method for rigorously interpreting SR models, and quantitative assessment of explainability remains an open and unresolved challenge.

Implications and Future Directions

This work opens up several research perspectives. A major priority is the design of new NR-IQA models specifically tailored for SR, grounded in perceptual principles that align more closely with human vision. Establishing large and diverse SR datasets will also be essential to enable realistic and comprehensive evaluations, covering a broad spectrum of SR techniques, resolution factors, and artifact types.

Developing perceptually grounded SR-NR-IQA metrics is another key direction, as existing approaches often fail to capture subtle qualities that are critical for real-world usage. Such metrics would help bridge the gap between algorithmic predictions and human expectations, leading to more reliable and user-centric evaluations.

Future research must also advance interpretability methods by moving beyond post-hoc visualization tools toward more quantitative, faithful, and trustworthy frameworks. These frameworks should reflect the actual reasoning of SR models and be validated against human perceptual relevance. Achieving this would enable more transparent, accountable,

and robust deployment of SR systems in high-stakes applications.

Finally, an innovative direction lies in integrating interpretability into the training process itself. Embedding explainability constraints during learning could foster models that are not only performant but also inherently more transparent. Such an approach would pave the way for SR technologies to be applied confidently in domains such as medical imaging, surveillance, and cultural heritage preservation—where both performance and explainability are critical requirements.

Annexes

Table 3.4: *Performance comparison of IQA and SR-IQA metrics on the CVIU-2017 dataset.*

Type	metric	SRCC	PLCC	KRCC	RMSE
Traditional IQA	BRISQUE (Full)	0,3303	0,3253	0,2760	0,2721
	BRISQUE (Crops)	0,3443	0,3404	0,2366	0,2488
	BRISQUE (Resize)	0,1415	0,1353	0,0949	0,2884
	NIQE (Full)	0,3348	0,3197	0,2315	0,2780
	NIQE (crop)	0,2963	0,2622	0,2024	0,3204
	NIQE (Resize)	0,1131	0,1045	0,0749	0,3287
	DIIVINE	0,2810	0,2639	0,1889	0,2733
	DIIVINE (crops)	0,2613	0,2378	0,1739	0,2726
	DIIVINE (Resize)	0,0803	0,1056	0,0529	0,2987
	BLIINDS-II (Full)	0,3601	0,3406	0,2474	0,2526
	BLIINDS-II (crops)	0,3647	0,3384	0,2498	0,2525
	BLIINDS-II (Resize)	0,1943	0,1928	0,1284	0,3194
	HOSA (Full)	0,4196	0,4262	0,2874	0,2195
	HOSA (crops)	0,0326	0,0368	0,0394	0,2540
	HOSA (Resize)	0,0448	0,0352	0,0306	0,2492

Table 3.4 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	HyperIQA (Crops)	**0,7869 (± 0,0254)**	**0,7935 (± 0,0245)**	**0,6014 (± 0,0213)**	**0,1510 (± 0,0100)**
	HyperIQA (Resize)	0.6727 (± 0,0141)	0.6904 (± 0,0146)	0.4796 (± 0,0171)	0.1818 (± 0,0143)
	HyperIQA (Fine-tuning)	0.5858 (± 0.0856)	0.5787 (± 0.0803)	0.4169 (± 0.0654)	0.2380 (± 0.0081)
	DBCNN (Crops)	0.6693 (± 0.0701)	0.6433 (± 0.0707)	0.4825 (± 0.0546)	0.1913 (± 0.0173)
	DBCNN (Resize)	0,6455 (± 0,0506)	0,6387 (± 0,0526)	0,4562 (± 0,0352)	0,1850 (± 0,0133)
	DBCNN (Fine-tuning)	0.5218 (± 0.0450)	0.5195 (± 0.0497)	0.3629 (± 0.0322)	0.2156 (± 0.0085)
	CONTRIQUE (Crops)	0.6691 (± 0.0338)	0.6501 (± 0.0497)	0.4861 (± 0.0225)	0.1880 (± 0.0175)
	CONTRIQUE (Resize)	0,5756 (± 0,0334)	0,5826 (± 0,0423)	0,4046 (± 0,0265)	0,2043 (± 0,0121)
	ReIQA (Crops)	0.0840 (± 0.0629)	0.0883 (± 0.0657)	0.0554 (± 0.0411)	0.2506 (± 0.0119)
	ReIQA (Resize)	0,1843 (± 0,0505)	0,2135 (± 0,2156)	0,1249 (± 0,0327)	0,2436 (± 0,0084)
Deep Learning IQA	ARNIQA (Crops)	0.0734 (± 0.0731)	0.0863 (± 0.0723)	0.0496 (± 0.0481)	0.2458 (± 0.0054)
	ARNIQA (Resize)	0,7654 (± 0,0311)	0,7736 (± 0,0299)	0,5885 (± 0,0280)	0,1584 (± 0,0128)
	ARNIQA (Fine-tuning)	0.6335 (± 0.0334)	0.6360 (± 0.0332)	0.4513 (± 0.0252)	0.1959 (± 0.0103)
	NIMA (Crops)	0,6342 (± 0,0309)	0,6352 (± 0,0336)	0,4667 (± 0,0284)	0,2085 (± 0,0596)
	NIMA-koniq (Fine-tuning)	0.2041 (± 0.1922)	0.2621 (± 0.1499)	0.1370 (± 0.1219)	0.2709 (± 0.0457)
	NIMA-spac (Fine-tuning)	0.2714 (± 0.1256)	0.3284 (± 0.1087)	0.1773 (± 0.0861)	0.2282 (± 0.0150)
	NIMA-vgg16-ava (Fine-tuning)	0.2629 (± 0.1031)	0.3205 (± 0.0814)	0.1755 (± 0.0674)	0.2275 (± 0.0276)
	PaQ-2-PiQ (Crops)	0.6786 (± 0.0320)	0.6936 (± 0.0351)	0.4890 (± 0.0237)	0.1740 (± 0.0153)
	PaQ-2-PiQ (Resize)	0,3772 (± 0,0470)	0,4037 (± 0,0476)	0,2633 (± 0,0365)	0,2338 (± 0,0034)

Table 3.4 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	PaQ-2-PiQ (Fine-tuning)	0.3013 (± 0.1029)	0.3382 (± 0.1342)	0.0742 (± 0.0347)	0.2310 (± 0.0219)
	TReS (Crops)	0.2022 (± 0.0612)	0.2543 (± 0.0309)	0.1707 (± 0.0398)	0.4056 (± 0.0262)
	TReS (Resize)	0.7848 (± 0.0203)	0.7884 (± 0.0190)	0.5895 (± 0.0097)	0.1518 (± 0.0093)
	CLIP-IQA (Crops)	0.6815 (± 0.0350)	0.6746 (± 0.0389)	0.4884 (± 0.0316)	0.1912 (± 0.0127)
	CLIP-IQA (Resize)	0.3862 (± 0.0754)	0.4137 (± 0.0697)	0.2572 (± 0.0533	0.2298 (± 0.0137)
	CLIP-IQA (Fine-tuning)	0.2203 (± 0.1534)	0.3243 (± 0.1315)	0.1386 (± 0.1028)	0.2429 (± 0.0174)
	CLIP-IQA+_rn50_512 (Fine-tuning)	0.6160 (± 0.0401)	0.6113 (± 0.0430)	0.4307 (± 0.0332)	0.1959 (± 0.0120)
	CLIP-IQA+_vitL14_512 (Fine-tuning)	0.2679 (± 0.0354)	0.2744 (± 0.0369)	0.1838 (± 0.0238)	0.2441 (± 0.0098)
	KLTSRQ (Full)	0.1242	0,0979	0,0818	0,2996
	KLTSRQ (Crops)	0,1481	0,1219	0,0976	0,3017
	KLTSRQ (Resize)	0,0884	0,0724	0,0580	0,3311
	Beron et al.(ODU) (Full)	0,2064	0,2387	0,138	0,3631
	Beron et al.(ODU) (Crops)	0,2152	0,2400	0,1443	0,3713
	Beron et al.(ODU) (Resize)	0,1586	0,2021	0,1047	0,3740
Traditional SR-IQA	Zhang (2019) (Full)	0,4176	0,4234	0,2812	0,2229
	Zhang (2019) (Crops)	0,4204	0,4186	0,2838	0,2271
	Zhang (2019) (Resize)	0,2344	0,3186	0,3378	0,2339
	Zhang (2021) (Crops)	0.3577	0.3464	0.2464	0.2441

Table 3.4 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	Zhang (2021) (Resize)	0.1616	0.1719	0.1123	0.2861
Deep Learning SR-IQA	CN-BSRIQA (Crops)	0,5426	0,5718	0,3801	0,2045
	CN-BSRIQA (Resize)	0,2375	0,2707	0,1613	0,2396
	DBSRNet	0,5778 (± 0.0467)	0,5478 (± 0.0389)	0,4168 (± 0.0379)	0,2132 (± 0.0076)
	Zhang (2022) (Crops)	0.2961 (± 0.0364)	0.3319 (± 0.0327)	0.1948 (± 0.0265)	0.2362 (± 0.0319)
	Zhang (2022) (Resize)	0.7070 (± 0.0319)	0.7150 (± 0.0423)	0.5163 (± 0.0255)	0.2409 (± 0.0460)
	Bare (Crops)	0.7443 (± 0.0409)	0.7414 (± 0.0319)	0.5518 (± 0.0343)	0.1678 (± 0.0106)
	Bare (Resize)	0.1805 (± 0.1403)	0.1922 (± 0.1544)	0.1177 (± 0.0934)	0.2470 (± 0.0068)
	SFD-IQA (Crops)	0.1414 (± 0.0515)	0.1452 (± 0.0467)	0.0985 (± 0.0339)	0.2438 (± 0.0055)
	SFD-IQA (Resize)	0.1042 (± 0.0561)	0.1196 (±0.0596)	0.1110 (± 0.0366)	0.2430 (± 0.0063)

Table 3.5: *Performance comparison of IQA and SR-IQA metrics on the QADS dataset.*

Type	metric	SRCC	PLCC	KRCC	RMSE
Traditional IQA	BRISQUE (Full)	0,5455	0,5241	0.3827	0.2401
	BRISQUE (Crops)	0.5605	0.5472	0.3970	0.2501
	BRISQUE (Resize)	0.6036	0.5418	0.431	0.2447
	NIQE (Full)	0.3943	0.3268	0.2768	0.4581
	NIQE (crop)	0.2305	0.1898	0.1575	0.4489
	NIQE (Resize)	0.4300	0.3267	0.3049	0.3635
	DIIVINE	0.3983	0.3983	0.2691	0.2682
	DIIVINE (crops)	0.4131	0.4246	0.2800	0.2672

Table 3.5 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	DIIVINE (Resize)	0.0616	0.0570	0.0415	0.3065
	BLIINDS-II (Full)	0.3659	0.3817	0.2492	0.2917
	BLIINDS-II (crops)	0.3585	0.3753	0.2439	0.3046
	BLIINDS-II (Resize)	0.4313	0.3736	0.3002	0.3873
	HOSA (Full)	0.3336	0.3656	0.2277	0.2655
	HOSA (crops)	0.1254	0.1281	0.0878	0.2861
	HOSA (Resize)	0.1214	0.1444	0.0831	0.2844
	HyperIQA (Crops)	0.8839 (± 0.0287)	0.8638 (± 0.0296)	0.6962 (± 0.0385)	0.1519 (± 0.0329)
	HyperIQA (Resize)	0,9015 (± 0,0432)	0,8983 (± 0,0554)	0,7212 (± 0,0558)	0,1633 (± 0,0407)
	HyperIQA (Fine-tuning)	0.7941 (± 0.0649)	0.7858 (± 0.0723)	0.5897 (± 0.0724)	0.1922 (± 0.0311)
	DBCNN (Crops)	0.8187 (± 0.0116)	0.8062 (± 0.0149)	0.6071 (± 0.0165)	0.1951 (± 0.0164)
	DBCNN (Resize)	0,8779 (± 0,0275)	0,8718 (± 0,0317)	0,6883 (± 0,0336)	0,1455 (± 0,0168)
	DBCNN (Fine-tuning)	0.7911 (± 0.0374)	0.7806 (± 0.0336)	0.6014 (± 0.0435)	0.1784 (± 0.0127)
	CONTRIQUE (Crops)	0.8851 (± 0.0235)	0.8810 (± 0.0239)	0.6950 (± 0.0291)	0.1415 (± 0.0163)
	CONTRIQUE (Resize)	0,7870 (± 0,0365)	0,7808 (± 0,0366)	0,5957 (± 0,0361)	0,1778 (± 0,0189)
	ReIQA (Crops)	0.6386 (± 0.1758)	0.5008 (± 0.1566)	0.4566 (± 0.1408)	0.2997 (± 0.1183)
	ReIQA (Resize)	0,1739 (± 0,0860)	0,1869 (± 0,0939)	0,1171 (± 0,0556)	0,3568 (± 0,0616)
Deep Learning IQA	ARNIQA (Crops)	0.8192 (± 0.0454)	0.8233 (± 0.0561)	0.6172 (± 0.0469)	0.1938 (± 0.0268)
	ARNIQA (Resize)	0,8999 (± 0,0148)	0,8951 (± 0,0148)	0,7221 (± 0,0162)	0,1337 (± 0,0142)
	ARNIQA (Fine-tuning)	0.7161 (± 0.1176)	0.7238 (± 0.1165)	0.5160 (± 0.1014)	0.2078 (± 0.0192)

Table 3.5 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	NIMA (Crops)	0,8521 (± 0.0610)	0,8483 (± 0.0610)	0,671 (± 0.0743)	0,162 (± 0.2449)
	NIMA-koniq (Fine-tuning)	0.5339 (± 0.2515)	0.2536 (± 0.3007)	0.4223 (± .1975)	0.3696 (± 1.6659)
	NIMA-spac (Fine-tuning)	0.6983 (± 0.0698)	0.7160 (± 0.1084)	0.5402 (± 0.0715)	0.2201 (± 0.0387)
	NIMA-vgg16-ava (Fine-tuning)	0.6881 (± 0.3580)	0.6089 (± 0.3751)	0.5264 (± 0.3101)	0.2173 (± 0.0529)
	PaQ-2-PiQ (Crops)	0.8599 (± 0.0481)	0.8527 (± 0.0479)	0.6344 (± 0.0532)	0.1735 (± 0.0613)
	PaQ-2-PiQ (Resize)	0.8857 (± 0.0381)	0.8467 (± 0.0531)	0.7059 (± 0.0510)	0.2323 (± 0.0134)
	PaQ-2-PiQ (Fine-tuning)	0.7196 (± 0.0438)	0.7193 (± 0.0388)	0.5448 (± 0.0345)	0.2075 (± 0.0423)
	TReS (Crops)	0.5290 (± 0.3230)	0.5423 (± 0.2508)	0.4439 (± 0.2525)	0.2991 (± 0.1404)
	TReS (Resize)	**0,9199 (± 0.0342)**	**0,9200 (± 0.0335)**	**0,7491 (± 0.0511)**	**0,1228 (± 0.0213)**
	CLIP-IQA (Crops)	0.8577 (± 0.0305)	0.8631 (± 0.0315)	0.6675 (± 0.0420)	0.2193 (± 0.0575)
	CLIP-IQA (Resize)	0.5800 (± 0.1386)	0.5804 (± 0.1719)	0.4107 (± 0.1156)	0.2472 (± 0.0278)
	CLIP-IQA (Fine-tuning)	0.7748 (± 0.0850)	0.7647 (± 0.1391)	0.6000 (± 0.0797)	0.1891 (± 0.0309)
	CLIP-IQA+_rn50_512 (Fine-tuning)	0.6983 (± 0.0681)	0.7160 (± 0.1084)	0.5402 (± 0.0715)	0.2201 (± 0.0387)
	CLIP-IQA+_vitL14_512 (Fine-tuning)	0.6863 (± 0.4083)	0.6663 (± 0.4817)	0.5218 (± 0.3543)	0.2838 (± 0.0557)
	KLTSRQ (Full)	0.0743	0.1107	0.0500	0,3285
	KLTSRQ (Crops)	0,1105	0,1276	0,0752	0,317
	KLTSRQ (Resize)	0,0397	0,0406	0,0282	0,3489
	Beron et al.(ODU) (Full)	0,5690	0,5476	0,4002	0,3114

Table 3.5 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
Traditional SR-IQA	Beron et al.(ODU) (Crops)	0,4501	0,3376	0,3096	0,3952
	Beron et al.(ODU) (Resize)	0,3975	0,4094	0,2762	0,3515
	Zhang (2019) (Full)	0,7673	0,7800	0,5776	0,1740
	Zhang (2019) (Crops)	0,7688	0,7719	0,5716	0,1763
	Zhang (2019) (Resize)	0,7194	0,7244	0,5324	0,2057
	Zhang (2021) (Crops)	0.7822	0.7651	0.5850	0.2388
	Zhang (2021) (Resize)	0.5218	0.4755	0.3962	0.3358
Deep Learning SR-IQA	CN-BSRIQA (Crops)	0,8924 (± 0,0112)	0,8922 (±0,0079)	0,7218 (±0,0182)	0,1366 (± 0,0086)
	CN-BSRIQA (Resize)	0.5155	0.4047	0.3912	0.2546
	DBSRNet (Resize)	0.8149 (± 0.0606)	0.7981 (± 0.0522)	0.6160 (± 0.0682)	0.1678 (± 0.0197)
	Zhang (2022) (Crops)	0.8049 (± 0.0050)	0.8003 (± 0.0753)	0.6016 (± 0.0694)	0.1646 (± 0.0245)
	Zhang (2022) (Resize)	0.8543 (± 0.0526)	0.8515 (± 0.0467)	0.6668 (± 0.0480)	0.1616 (± 0.0187)
	Bare (Crops)	0.8000 (± 0.0582)	0.7837 (± .0715)	0.6098 (± 0.0565)	0.1896 (± 0.0233)
	Bare (Resize)	0,5984 (± 0.0476)	0,5752 (± 0.0534)	0,4398 (± 0.0358)	0,2638 (± 0.0157)
	SFD-IQA (Crops)	0.7842 (± 0.0748)	0.7366 (± 0.0899)	0.5883 (± 0.0743)	0.1890 (± 0.0322)
	SFD-IQA (Resize)	0.7568 (± 0.1334)	0.7485 (± 0.1338)	0.5639 (± 0.1165)	0.1844 (± 0.0326)

Table 3.6: *Performance comparison of IQA and SR-IQA metrics on the RealSRQ dataset.*

Type	metric	SRCC	PLCC	KRCC	RMSE
Traditional IQA	BRISQUE (Full)	0.0740	0.0239	0.0575	3.1412
	BRISQUE (Crops)	0.1066	0.0207	0.0804	2.7605
	BRISQUE (Resize)	0.0285	0.0443	0.0175	3.8526
	NIQE (Full)	0.1219	0.0212	0.0865	1.7196
	NIQE (crop)	0.0894	-0.0036	0.0653	2.9867
	NIQE (Resize)	0.0540	0.1187	0.0361	4.0365
	DIIVINE	0.0044	-0.0642	0.0051	2.6733
	DIIVINE (crops)	0.0646	-0.0406	0.0480	3.1771
	DIIVINE (Resize)	0.0509	-0.0148	0.0351	4.0444
	BLIINDS-II (Full)	0.1723	0.1756	0.1151	2.5171
	BLIINDS-II (crops)	0.1014	0.1568	0.0680	1.6049
	BLIINDS-II (Resize)	0.0854	0.1101	0.0556	2.3638
	HOSA (Full)	0.0927	0.1264	0.0618	0.9186
	HOSA (crops)	0.0369	0.0502	0.0219	0.9404
	HOSA (Resize)	0.0285	0.0652	0.0181	0.8985
	HyperIQA (Crops)	0.1060 (± 0.0897)	0.2866 (± 0.0791)	0.0652 (± 0.0632)	0.8471 (± 0.0227)
	HyperIQA (Resize)	0,1363 (± 0,0366)	0,3887 (± 0,1422)	0,0906 (± 0,0241)	0,8701 (± 0,0465)
	DBCNN (Crops)	0.2486 (± 0.0489)	0.4683 (± 0.0692)	0.1747 (± 0.0354)	0.8210 (± 0.0232)
	DBCNN (Resize)	0,2927 (± 0,0256)	0,4097 (± 0,0457)	0,1830 (± 0,0174)	0,8375 (± 0,0250)
	CONTRIQUE (Crops)	0.4282 (± 0.2171)	0.4392 (± 0.2235)	0.3015 (± 0.1529)	0.8356 (± 0.0677)
	CONTRIQUE (Resize)	0,3281 (± 0,0560)	0,4564 (± 0,0735)	0,2283 (± 0,0456)	0,7805 (± 0,0440)
	ReIQA (Crops)	0.0562 (± 0.0855)	0.1977 (± 0.0437)	0.0379 (± 0.0615)	1.0842 (± 2.287)

Table 3.6 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	ReIQA (Resize)	-0,0006 (± 0,0377)	0,0925 (± 0,0397)	-0,0009 (± 0,2347)	1,2572 (± 0,0235)
Deep Learning IQA	ARNIQA (Crops)	-0.0157 (± 0.0482)	0.1442 (± 0.1240)	-0.0393 (± 0.0334)	0.9224 (± 0.0889)
	ARNIQA (Resize)	0,6152 (± 0,0502)	0,7850 (± 0,0571)	0,4491 (± 0,0417)	0,5387 (± 0,0923)
	NIMA (Crops)	0.1051 (± 0.1222)	0.1628 (± 0.1495)	0.0713 (± 0.0819)	0.7458 (± 0.0571)
	PaQ-2-PiQ (Crops)	**0.6517 (± 0.0639)**	**0.7009 (± 0.0777)**	**0.4818 (± 0.0530)**	**0.6684 (± 0.0649)**
	PaQ-2-PiQ (Resize)	0.09098 (± 0.0928)	0.3707 (± 0.1661)	0.0614 (± 0.0626)	0.8288 (± 0.0418)
	TReS (Crops)	0.4828 (± 0.0609)	0.6682 (± 0.0819)	0.3656 (± 0.0480)	0.6795 (± 0.0640)
	TReS (Resize)	0.4828 (± 0.0609)	0.6682 (± 0.0819)	0.3655 (± 0.0480)	0.6795 (± 0.0640)
	CLIP-IQA (Crops)	0.4907 (± 0.0429)	0.7255 (± 0.0819)	0.3499 (± 0.0324)	0.6195 (± 0.0752)
	CLIP-IQA (Resize)	0.1176 (± 0.0384)	0.1684 (± 0.0571)	0.0790 (± 0.0255)	0.8972 (± 0.0294)
	KLTSRQ (Full)	0.0644	0.0079	0.0446	5,1502
	KLTSRQ (Crops)	0,0785	0,0034	0,0531	2,1841
	KLTSRQ (Resize)	0,1045	0,0115	0,0734	2,3038
	Beron et al.(ODU) (Full)	0,0038	0,0581	0,0019	5,5860
	Beron et al.(ODU) (Crops)	0,0384	0,0504	0,0219	7,1673
	Beron et al.(ODU) (Resize)	0,0253	0,0521	0,016	7,0365
Traditional SR-IQA	Zhang (2019) (Full)	0.0236	0.2649	0.0132	0.8543
	Zhang (2019) (Crops)	0.0451	0.2283	0.0295	0.8483
	Zhang (2019) (Resize)	-0,0465	0,1034	-0,0318	0,8947

Table 3.6 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	Zhang (2021) (Crops)	0.1422	0.3297	0.0995	1.3536
	Zhang (2021) (Resize)	0.0423	0.0874	0.0294	1.4891
Deep Learning SR-IQA	DBSRNet (Resize)	0.0537 (± 0.1106)	0.2685 (± 0.0746)	0.0352 (± 0.0753)	0.8712 (± 0.0310)
	Zhang (2022) (Crops)	0.0524 (± 0.0540)	0.3773 (± 0.1180)	0.0392 (± 0.0377)	0.8544 (± 0.0359)
	Zhang (2022) (Resize)	0,0609 (±)	0,4479 (±)	0,0375 (±)	0,4459 (±)
	Bare (Crops)	0.2376 (± 0.2006)	0.0807 (± 0.3549)	0.1722 (± 0.1427)	0.8841 (± 0.1038)
	Bare (Resize)	-0.0991 (± 0.0442)	-0.0015 (± 0.0462)	-0.0678 (± 0.0306)	0.9068 (± 0.0282)
	SFD-IQA (Crops)	0.1947 (± 0.0291)	0.5267 (± 0.0606)	0.1359 (± 0.0212)	0.7711 (± 0.0179)
	SFD-IQA (Resize)	0.0781 (± 0.0405)	0.3312 (± 0.0773)	0.0535 (± 0.0289)	0.8494 (± 0.0343)

Table 3.7: *Performance comparison of IQA and SR-IQA metrics on the SISAR dataset.*

Type	metric	SRCC	PLCC	KRCC	RMSE
Traditional IQA	BRISQUE (Full)	0,7008	0,6956	0,5101	0,2164
	BRISQUE (Crops)	0,7167	0,7099	0,5260	0,1689
	BRISQUE (Resize)	0,7160	0,7161	0,5266	0,1916
	NIQE (Full)	0,6959	0,7017	0,5083	0,1844
	NIQE (crop)	0,7038	0,7077	0,5160	0,1761
	NIQE (Resize)	0,7031	0,7044	0,5051	0,2081
	DIIVINE (crops)	0.4740	0.4765	0.3216	0.2228
	DIIVINE (Resize)	0.5993	0.5843	0.4167	0.2078
	BLIINDS-II (Full)	0,5198	0,4922	0,3597	0,2255

Table 3.7 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	BLIINDS-II (crops)	0,5284	0,4808	0,3663	0,2321
	BLIINDS-II (Resize)	0.6029	0.5814	0.4283	0.2190
	HOSA (Full)	0,3814	0,3798	0,2582	0,2234
	HOSA (crops)	0,3814	0,3798	0,2582	0,2234
	HOSA (Resize)	0.3837	0.4003	0.2633	0.2189
	HyperIQA (Crops)	**0.9418 (± 0.1134)**	**0.9384 (± 0.0933)**	**0.7837 (± 0.1270)**	**0.0755 (± 0.0478)**
	HyperIQA (Resize)	0,75474 (± 0,0219)	0,75564 (± 0,0202)	0,5612 (± 0,0097)	0,15732 (± 0,0022)
	DBCNN (Crops)	0.7256 (± 0.0234)	0.7325 (± 0.0223)	0.5440 (± 0.0220)	0.1626 (± 0.0071)
	DBCNN (Resize)	0,7271 (± 0,0207)	0,7256 (± 0,0192)	0,5390 (± 0,0198)	0,1623 (± 0,0053)
	CONTRIQUE (Crops)	0.7208 (± 0.0144)	0.7241 (± 0.0135)	0.5292 (± 0.0146)	0.1641 (± 0.0044)
	CONTRIQUE (Resize)	0,7133 (± 0,0183)	0,7212 (± 0,0163)	0,5180 (± 0,0162)	0,1665 (± 0,0053)
	ReIQA (Crops)	0.6916 (± 0.0279)	0.6975 (± 0.0490)	0.5021 (± 0.0262)	0,1741 (± 0.0191)
	ReIQA (Resize)	0,9000 (± 0,085)	0,8955 (± 0,0859)	0,7285 (± 0,0930)	0,0965 (± 0,0437)
Deep Learning IQA	ARNIQA (Crops)	0.6917 (± 0.0214)	0.6913 (± 0.0219)	0.4963 (± 0.0203)	0.1743 (± 0.0048)
	ARNIQA (Resize)	0,7539 (± 0,0066)	0,7525 (± 0,0070)	0,5563 (± 0,0070)	0,1576 (± 0,0020)
	NIMA (Crops)	0.7558 (± 0.0089)	0.7599 (± 0.0095)	0.5592 (± 0.0098)	0,15687 (± 0.0283)
	PaQ-2-PiQ (Crops)	0.7331 (± 0.0306)	0.7335 (± 0.0294)	0.5389 (± 0.0292)	0.1648 (± 0.0067)
	PaQ-2-PiQ (Resize)	0.7546 (± 0.0200)	0.7598 (± 0.0205)	0.5640 (± 0.0196)	0.1590 (± 0.0053)
	TReS (Crops)	0.3635 (± 0.0782)	0.3539 (± 0.0730)	0.2761 (± 0.0587)	0.3294 (± 0.0173)
	TReS (Resize)	0.7585 (± 0.0163)	0.7636 (± 0.0157)	0.5608 (± 0.0143)	0.1543 (± 0.0046)

Table 3.7 (continued)

Type	metric	SRCC	PLCC	KRCC	RMSE
	CLIP-IQA (Crops)	0.7468 (± 0.0217)	0.7493 (± 0.0215)	0.5536 (± 0.0215)	0.1597 (± 0.0070)
	CLIP-IQA (Resize)	0.7339 (± 0.0218)	0.7365 (± 0.0204)	0.5424 (± 0.0200)	0.1631 (± 0.0056)
Traditional SR-IQA	KLTSRQ (Full)	0,2456	0,2404	0,1640	0,3207
	Beron et al.(ODU) (Crops)	0.5295	0.5123	0.3694	0.3778
	Beron et al.(ODU) (Resize)	0,3635	0,3539	0,2761	0,3294
	Zhang (2019) (Full)	0.7193	0.7195	0.5310	0.1664
	Zhang (2019) (Crops)	0,7193	0,7195	0,5310	0,1664
	Zhang (2019) (Resize)	0.8077	0.8060	0.6418	0.1693
	Zhang (2021) (Crops)	0.5859	0.5950	0.4103	0.2002
	Zhang (2021) (Resize)	0.4086	0.4161	0.2807	0.2686
Deep Learning SR-IQA	DBSRNet (Resize)	0.7324 (± 0.0155)	0.7340 (± 0.0129)	0.5412 (± 0.0124)	0.1610 (± 0.0052)
	Zhang (2022) (Crops)	0.5669 (± 0.0354)	0.5830 (± 0.0314)	0.4001 (± 0.0259)	0.2002 (± 0.0033)
	Zhang (2022) (Resize)	0.7424 (± 0.0218)	0.7444 (± 0.0214)	0.5537 (± 0.0197)	0.1900 (± 0.0090)
	Bare (Crops)	0.5862 (± 0.0083)	0.5777 (± 0.0117)	0.4079 (± 0.0068)	0.1996 (± 0.0017
	Bare (Resize)	0.6967 (± 0.0286)	0.7049 (± 0.0264)	0.5122 (± 0.0257)	0.1661 (± 0.0071)
	SFD-IQA (Crops)	0.6295 (± 0.0406)	0.6004 (± 0.0496)	0.4421 (± 0.0331)	0.1929 (± 0.0106)
	SFD-IQA (Resize)	0.6752 (± 0.0324)	0.6760 (± 0.0326)	0.4854 (± 0.0312)	0.1770 (± 0.0080)

Table 3.8: *T-test results for the statistical comparison of image quality assessment methods on the CVIU-2017 database (PLCC).*

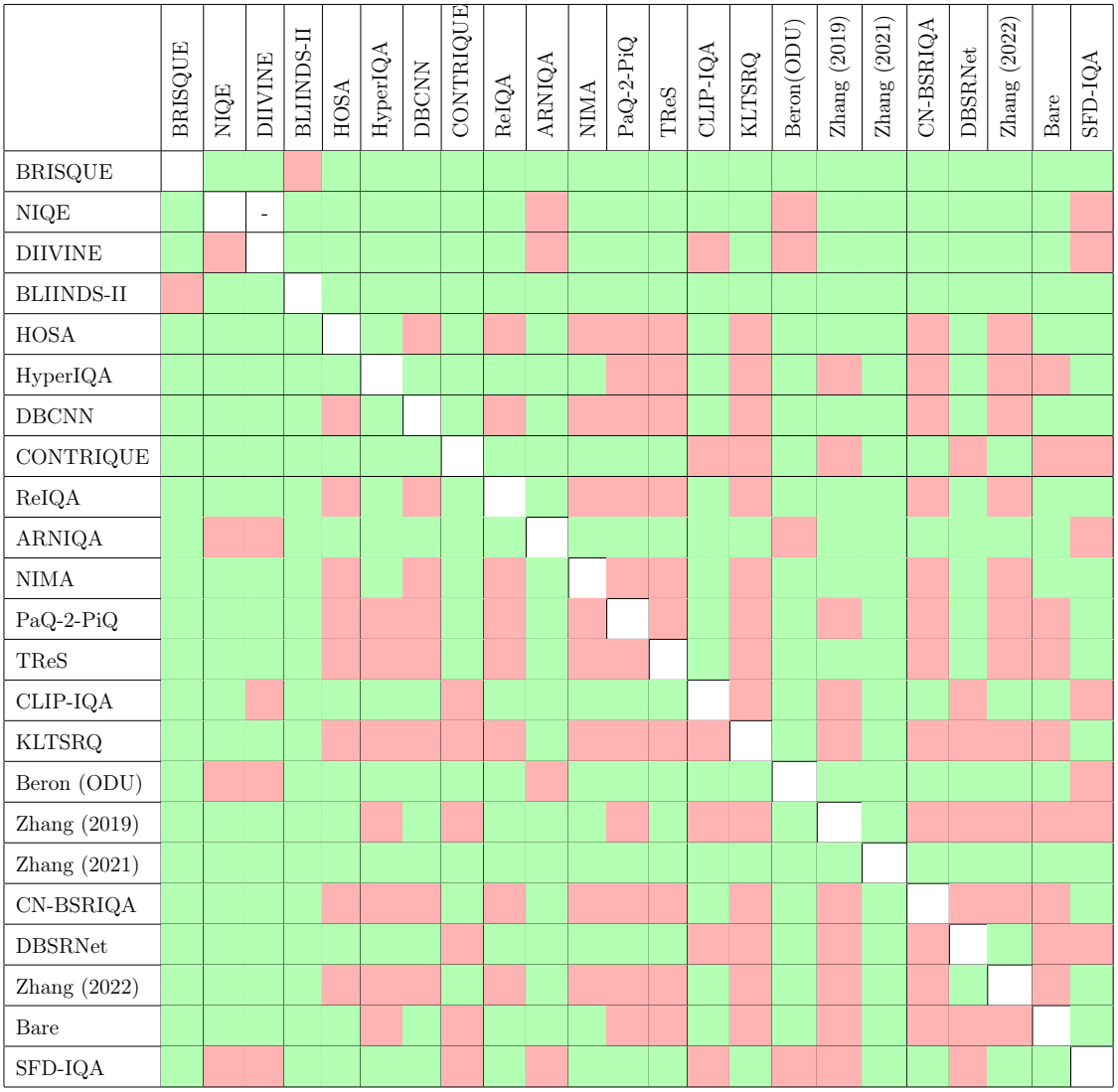


Table 3.9: *T-test results for the statistical comparison of image quality assessment methods on the CVIU-2017 database (KROCC).*

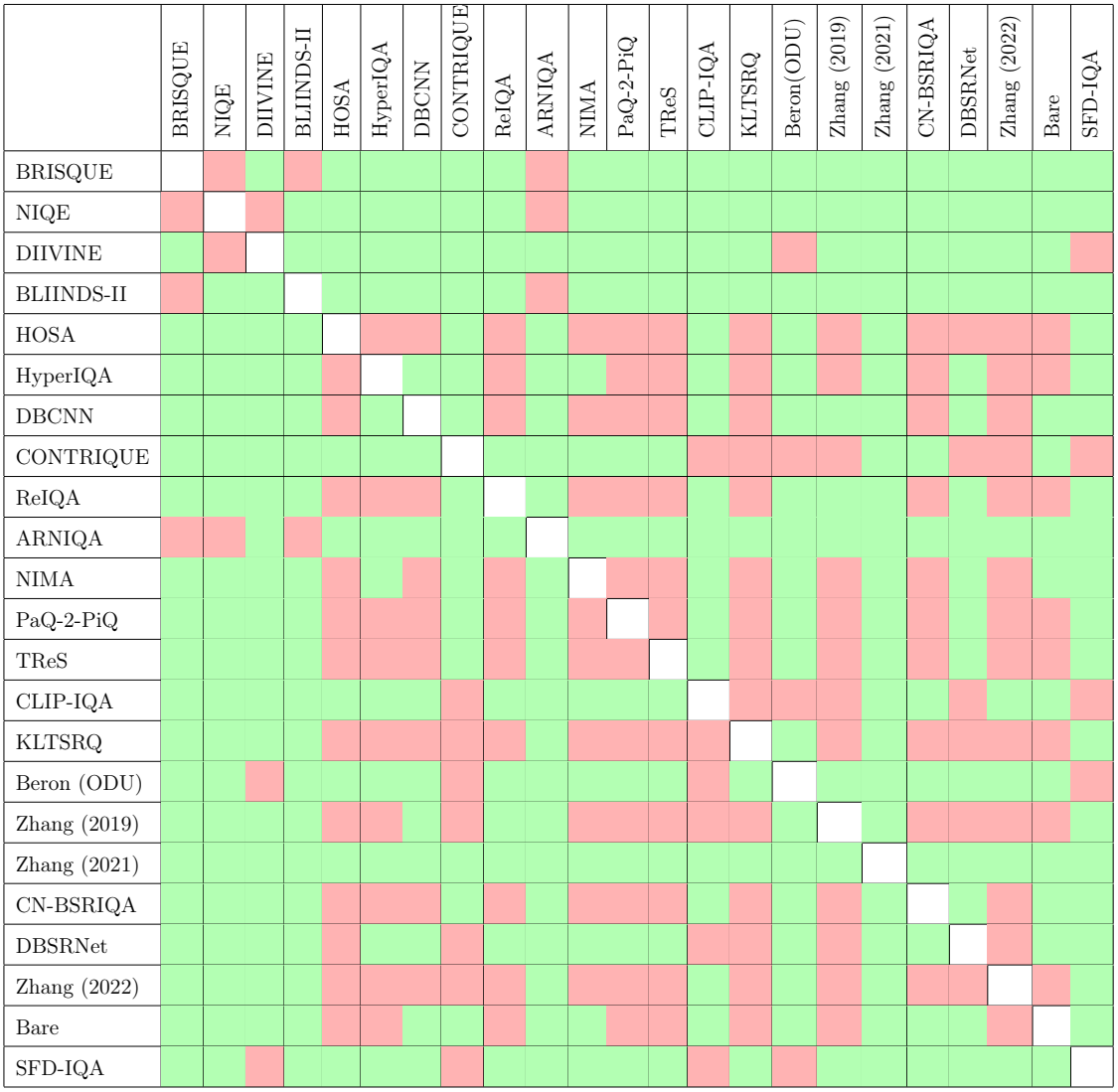


Table 3.10: *T-test results for the statistical comparison of image quality assessment methods on the CVIU-2017 database (RMSE).*

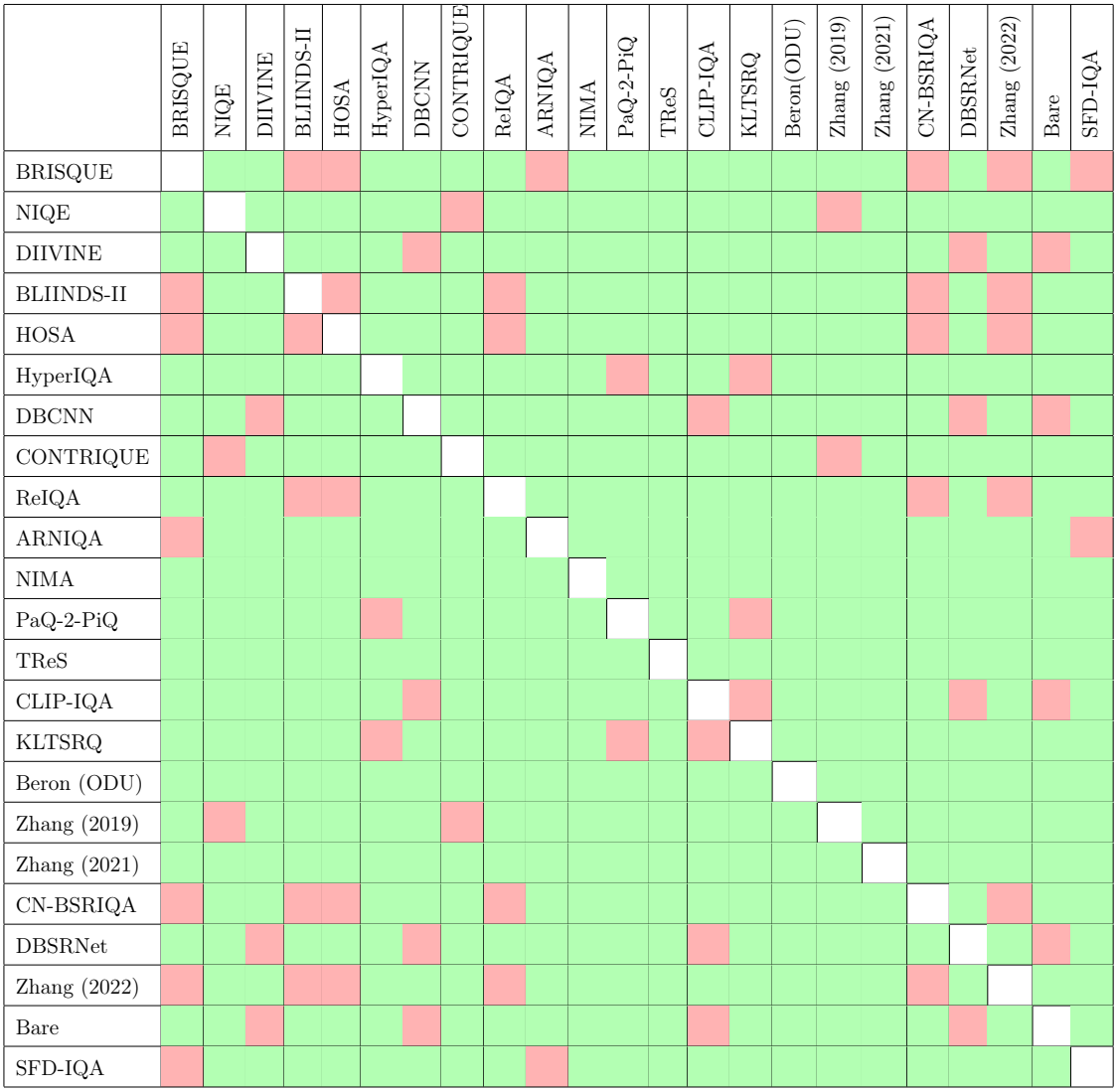


Table 3.11: *T-test results for the statistical comparison of image quality assessment methods on the QADS database (PLCC).*

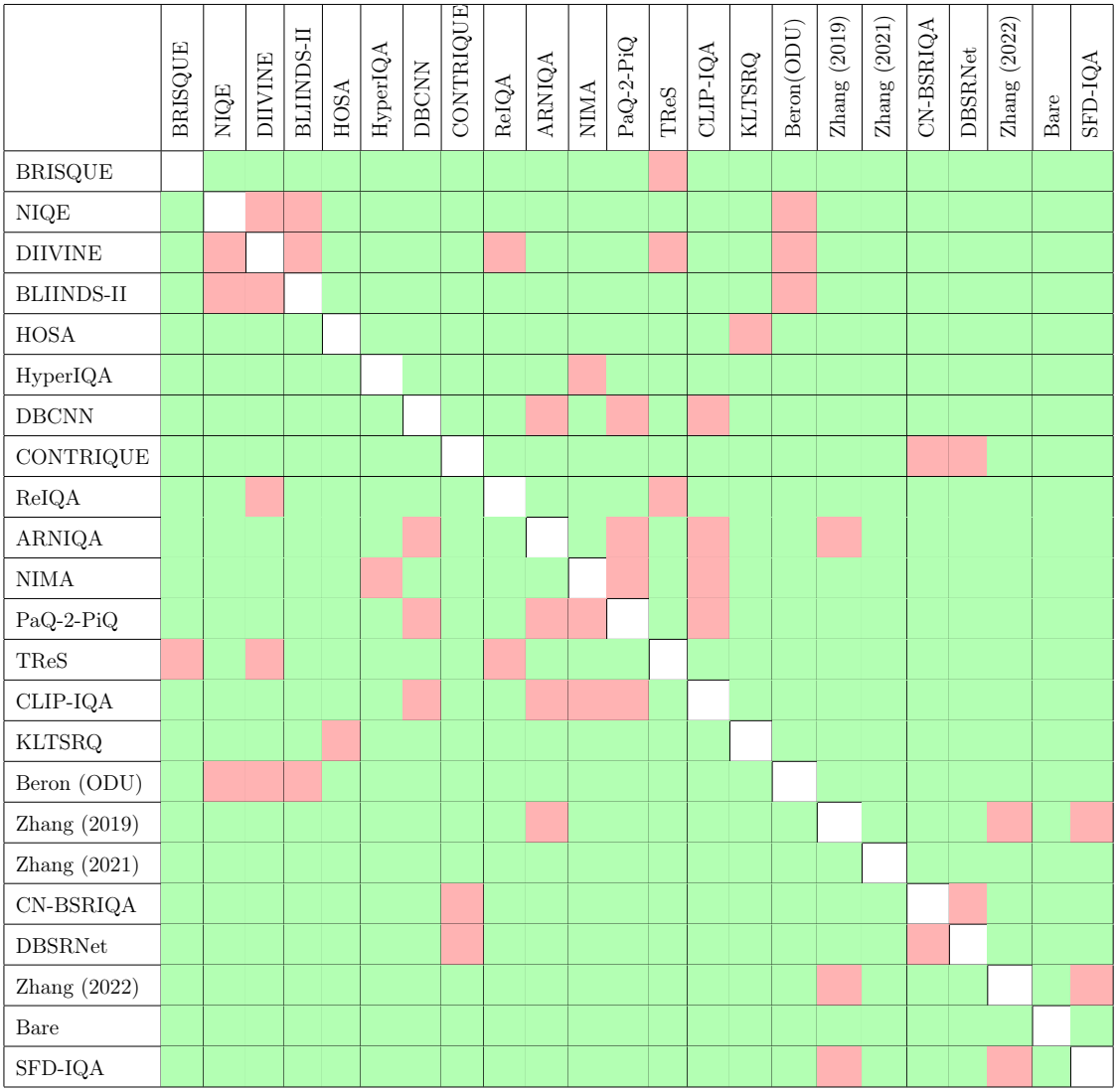


Table 3.12: *T-test results for the statistical comparison of image quality assessment methods on the QADS database (KROCC).*

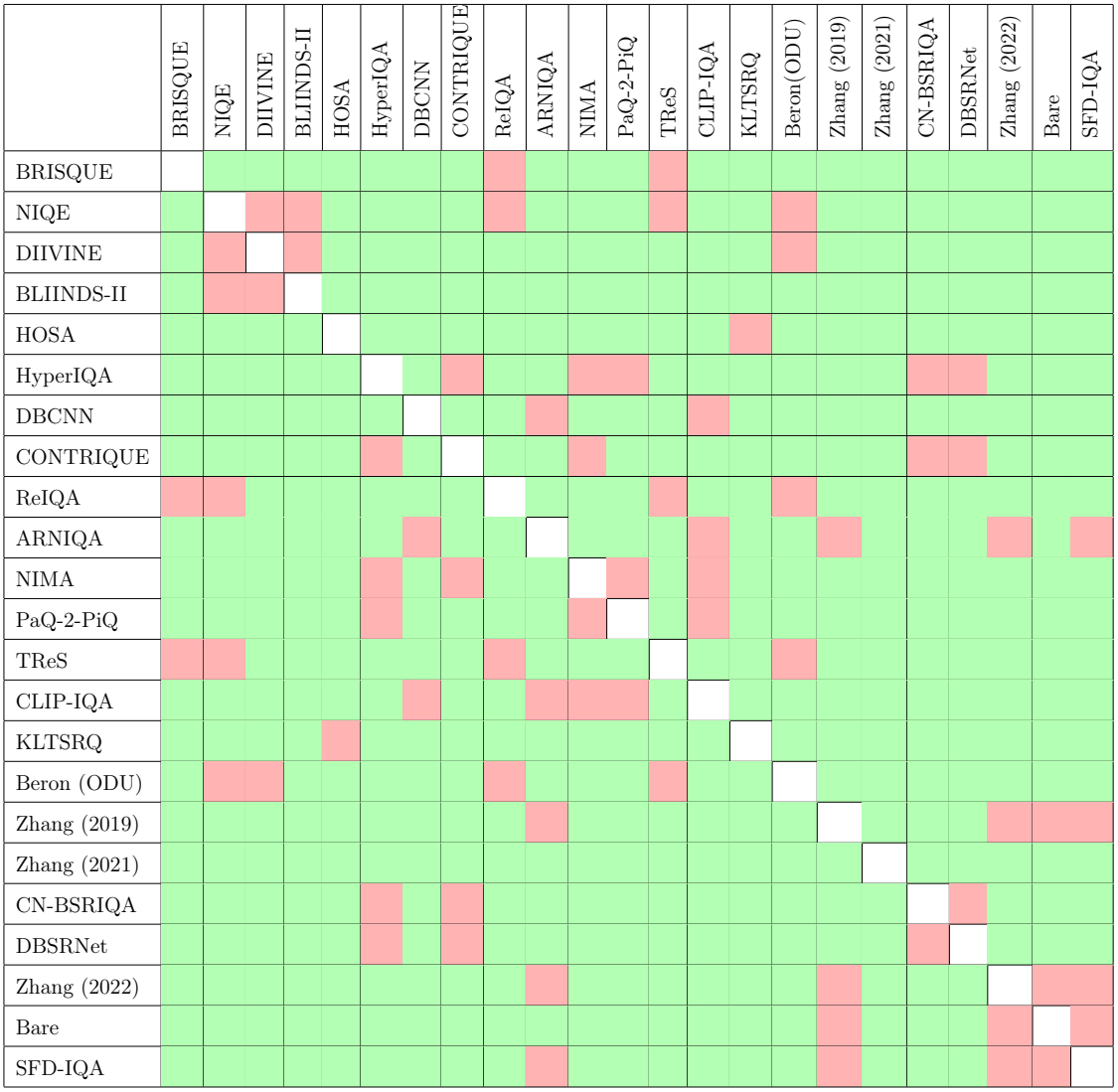


Table 3.13: *T-test results for the statistical comparison of image quality assessment methods on the QADS database (RMSE).*

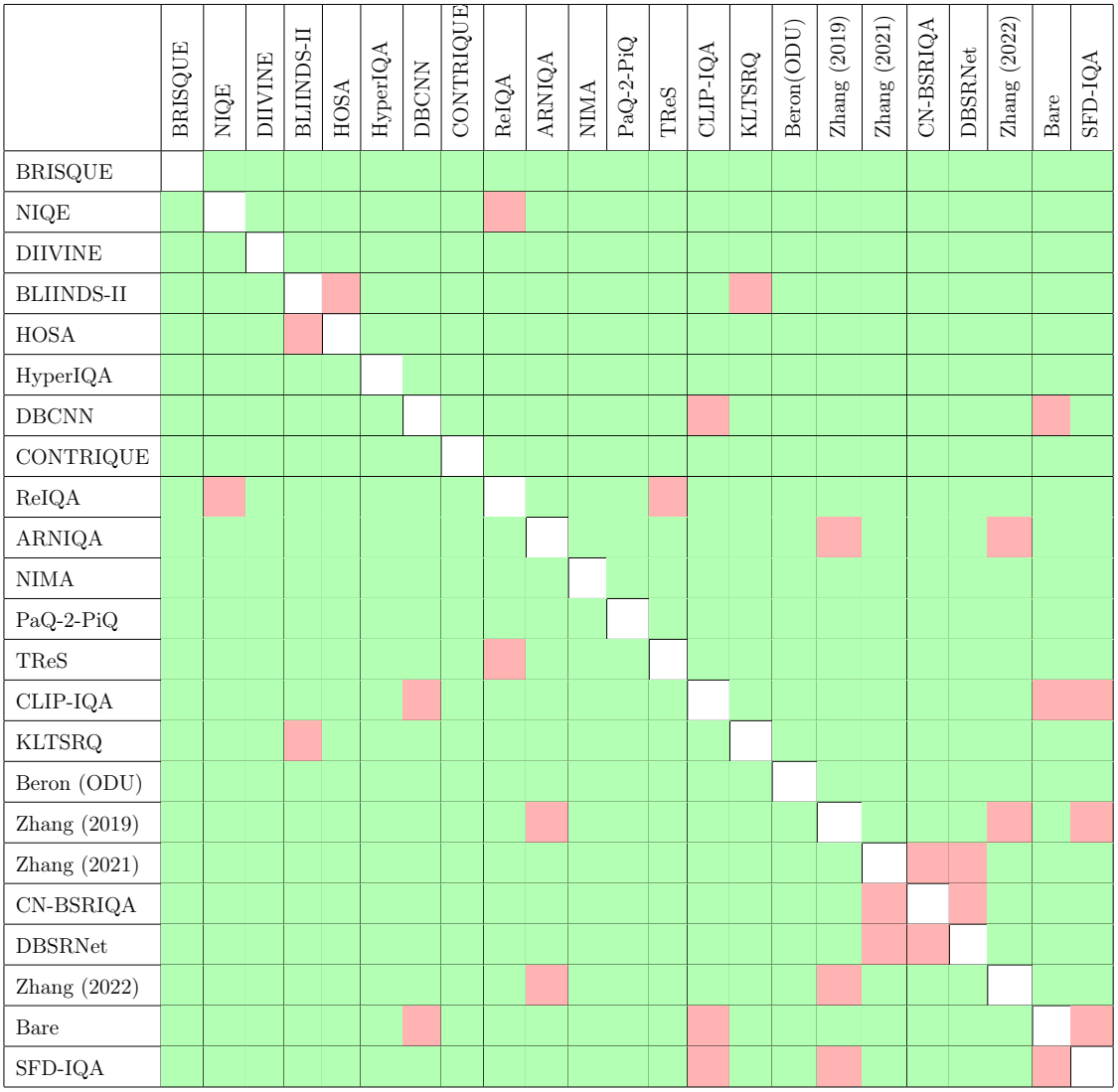


Table 3.14: *T-test results for the statistical comparison of image quality assessment methods on the RealSRQ database (SROCC).*

	BRISQUE	NIQE	DIIVINE	BLIINDS-II	HOSA	HyperIQA	DBCNN	CONTRIQUE	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	KLTSRQ	Beron(ODU)	Zhang (2019)	Zhang (2021)	Zhang (2022)	Bare	SFD-IQA
BRISQUE																					
NIQE																					
DIIVINE																					
BLIINDS-II																					
HOSA																					
HyperIQA																					
DBCNN																					
CONTRIQUE																					
ReIQA																					
ARNIQA																					
NIMA																					
PaQ-2-PiQ																					
TReS																					
CLIP-IQA																					
KLTSRQ																					
Beron (ODU)																					
Zhang (2019)																					
Zhang (2021)																					
Zhang (2022)																					
Bare																					
SFD-IQA																					

Table 3.15: *T-test results for the statistical comparison of image quality assessment methods on the RealSRQ database (PLCC).*

	BRISQUE	NIQE	DIIVINE	BLIINDS-II	HOSA	HyperIQA	DBCNN	CONTRIQUE	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	KLTSRQ	Beron(ODU)	Zhang (2019)	Zhang (2021)	Zhang (2022)	Bare	SFD-IQA
BRISQUE																					
NIQE																					
DIIVINE																					
BLIINDS-II																					
HOSA																					
HyperIQA																					
DBCNN																					
CONTRIQUE																					
ReIQA																					
ARNIQA																					
NIMA																					
PaQ-2-PiQ																					
TReS																					
CLIP-IQA																					
KLTSRQ																					
Beron (ODU)																					
Zhang (2019)																					
Zhang (2021)																					
Zhang (2022)																					
Bare																					
SFD-IQA																					

Table 3.16: *T-test results for the statistical comparison of image quality assessment methods on the RealSRQ database (KROCC).*

	BRISQUE	NIQE	DIIVINE	BLIINDS-II	HOSA	HyperIQA	DBCNN	CONTRIQUE	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	KLTSRQ	Beron(ODU)	Zhang (2019)	Zhang (2021)	Zhang (2022)	Bare	SFD-IQA
BRISQUE																					
NIQE																					
DIIVINE																					
BLIINDS-II																					
HOSA																					
HyperIQA																					
DBCNN																					
CONTRIQUE																					
ReIQA																					
ARNIQA																					
NIMA																					
PaQ-2-PiQ																					
TReS																					
CLIP-IQA																					
KLTSRQ																					
Beron (ODU)																					
Zhang (2019)																					
Zhang (2021)																					
Zhang (2022)																					
Bare																					
SFD-IQA																					

Table 3.17: *T-test results for the statistical comparison of image quality assessment methods on the RealSRQ database (RMSE).*

	BRISQUE	NIQE	DIIVINE	BLIINDS-II	HOSA	HyperIQA	DBCNN	CONTRIQUE	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	KLTSRQ	Beron(ODU)	Zhang (2019)	Zhang (2021)	Zhang (2022)	Bare	SFD-IQA
BRISQUE																					
NIQE																					
DIIVINE																					
BLIINDS-II																					
HOSA																					
HyperIQA																					
DBCNN																					
CONTRIQUE																					
ReIQA																					
ARNIQA																					
NIMA																					
PaQ-2-PiQ																					
TReS																					
CLIP-IQA																					
KLTSRQ																					
Beron (ODU)																					
Zhang (2019)																					
Zhang (2021)																					
Zhang (2022)																					
Bare																					
SFD-IQA																					

Table 3.18: *T-test results for the statistical comparison of image quality assessment methods on the SISAR database (SROCC).*

	BRISQUE	NIQE	HyperIQA	DBCNN	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	Zhang (2021)	DBSRNet	Zhang (2022)	Bare	SFD-IQA
BRISQUE															
NIQE															
HyperIQA															
DBCNN															
ReIQA															
ARNIQA															
NIMA															
PaQ-2-PiQ															
TReS															
CLIP-IQA															
Zhang (2021)															
DBSRNet															
Zhang (2022)															
Bare															
SFD-IQA															

Table 3.19: *T-test results for the statistical comparison of image quality assessment methods on the SISAR database (PLCC).*

	BRISQUE	NIQE	HyperIQA	DBCNN	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	Zhang (2021)	DBSRNet	Zhang (2022)	Bare	SFD-IQA
BRISQUE															
NIQE															
HyperIQA															
DBCNN															
ReIQA															
ARNIQA															
NIMA															
PaQ-2-PiQ															
TReS															
CLIP-IQA															
Zhang (2021)															
DBSRNet															
Zhang (2022)															
Bare															
SFD-IQA															

Table 3.20: *T-test results for the statistical comparison of image quality assessment methods on the SISAR database (KRCC).*

	BRISQUE	NIQE	HyperIQA	DBCNN	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	Zhang (2021)	DBSRNet	Zhang (2022)	Bare	SFD-IQA
BRISQUE															
NIQE															
HyperIQA															
DBCNN															
ReIQA															
ARNIQA															
NIMA															
PaQ-2-PiQ															
TReS															
CLIP-IQA															
Zhang (2021)															
DBSRNet															
Zhang (2022)															
Bare															
SFD-IQA															

Table 3.21: *T-test results for the statistical comparison of image quality assessment methods on the SISAR database (RMSE).*

	BRISQUE	NIQE	HyperIQA	DBCNN	ReIQA	ARNIQA	NIMA	PaQ-2-PiQ	TReS	CLIP-IQA	Zhang (2021)	DBSRNet	Zhang (2022)	Bare	SFD-IQA
BRISQUE															
NIQE															
HyperIQA															
DBCNN															
ReIQA															
ARNIQA															
NIMA															
PaQ-2-PiQ															
TReS															
CLIP-IQA															
Zhang (2021)															
DBSRNet															
Zhang (2022)															
Bare															
SFD-IQA															

Bibliography

[1] M. Farooq, M. N. Dailey, A. Mahmood, J. Moonrinta, and M. Ekpanyapong, "Human face super-resolution on poor quality surveillance video footage," *Neural Computing and Applications*, vol. 33, pp. 13505–13523, 2021.

[2] MRI Questions, "Super-resolution." https://mriquestions.com/super-resolution.html, 2025. Consulté le 30 juin 2025.

[3] Open Data Science, "Enhancing satellite imagery through super-resolution." https://opendatascience.com/enhancing-satellite-imagery-through-super-resolution/, 2023. Consulté le 30 juin 2025.

[4] W. Zhou and Z. Wang, "Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity," in *Proceedings of the 30th ACM international conference on multimedia*, pp. 934–942, 2022.

[5] F. Zhou, W. Sheng, Z. Lu, B. Kang, M. Chen, and G. Qiu, "Super-resolution image visual quality assessment based on structure–texture features," *Signal Processing: Image Communication*, vol. 117, p. 117025, 2023.

[6] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9199–9208, 2021.

[7] J. Gildenblat, "Introduction to grad-cam," 2021. Accessed: 2025-07-28.

[8] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*, pp. 2366–2369, IEEE, 2010.

[9] Y. Qi, M. Lou, Y. Liu, L. Li, Z. Yang, and W. Nie, "Advancing image super-resolution techniques in remote sensing: A comprehensive survey," *arXiv preprint arXiv:2505.23248*, 2025.

[10] C. Otgonbaatar, H. Kim, P.-H. Jeon, S.-H. Jeon, S.-J. Cha, J.-K. Ryu, W. B. Jung, H. Shim, and S. M. Ko, "Super-resolution deep learning image reconstruction: image quality and myocardial homogeneity in coronary computed tomography angiography," *Journal of Cardiovascular Imaging*, vol. 32, no. 1, p. 30, 2024.

[11] K. Ye, I. G. Dorado, M. Raptis, M. Delbracio, I. Zhu, P. Milanfar, and H. Talebi, "Textsr: Diffusion super-resolution with multilingual ocr guidance," *arXiv preprint arXiv:2505.23119*, 2025.

[12] A. Sendjasni and M.-C. Larabi, "Embedding similarity guided license plate super resolution," *Neurocomputing*, vol. 651, p. 130657, 2025.

[13] S. N. Ferdous, A. Dabouei, J. Dawson, and N. M. Nasrabadi, "Super-resolution guided pore detection for fingerprint recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8085–8092, IEEE, 2021.

[14] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, pp. 1423–1468, 2014.

[15] S. Lertrattanapanich and N. K. Bose, "High resolution image formation from low resolution frames using delaunay triangulation," *IEEE transactions on image processing*, vol. 11, no. 12, pp. 1427–1441, 2002.

[16] D. Su and P. Willis, "Image interpolation by pixel-level data-dependent triangulation," in *Computer graphics forum*, vol. 23, pp. 189–201, Wiley Online Library, 2004.

[17] X. Zhao, Y. Su, Y. Dong, J. Wang, and L. Zhai, "Kind of super-resolution method of ccd image based on wavelet and bicubic interpolation," *Application Research of Computers*, vol. 26, no. 6, pp. 2365–2367, 2009.

[18] L. Shu, Q. Zhu, Y. He, W. Chen, and J. Yan, "A survey of super-resolution image quality assessment," *Neurocomputing*, p. 129279, 2024.

[19] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[20] A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 169–172, IEEE, 1992.

[21] A. J. Patti, M. I. Sezan, and A. M. Tekalp, "High resolution standards conversion of low resolution video," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2197–2200, IEEE, 1995.

[22] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 346–360, 2013.

[23] M. N. Bareja and C. K. Modi, "An effective iterative back projection based single image super resolution approach," in *2012 international conference on communication systems and Network technologies*, pp. 95–99, IEEE, 2012.

[24] Z. Tang, M. Deng, C. Xiao, and J. Yu, "Projection onto convex sets super-resolution image reconstruction based on wavelet bi-cubic interpolation," in *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, vol. 1, pp. 351–354, IEEE, 2011.

[25] Q. Yuan, L. Zhang, and H. Shen, "Multiframe super-resolution employing a spatially weighted total variation model," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 3, pp. 379–392, 2011.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

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

[29] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on image processing*, vol. 20, no. 7, pp. 1838–1857, 2011.

[30] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

[31] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[32] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

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

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

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

[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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

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

[39] Y. Li, X. Yang, J. Fu, G. Yue, and W. Zhou, "Deep bi-directional attention network for image super-resolution quality assessment," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2024.

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

[41] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[42] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[43] M. A. Saad, A. C. Bovik, and C. Charrier, "Dct statistics model-based blind image quality assessment," in *2011 18th ieee international conference on image processing*, pp. 3093–3096, IEEE, 2011.

[44] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[45] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.

[46] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[48] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 4149–4161, 2022.

[49] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5846–5855, 2023.

[50] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo, "Arniqa: Learning distortion manifold for image quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024.

[51] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE transactions on image processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[52] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "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*, pp. 3575–3585, 2020.

[53] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 2555–2563, 2023.

[54] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: a real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022.

[55] J. Beron, H. D. Benitez-Restrepo, and A. C. Bovik, "Blind image quality assessment for super resolution via optimal feature selection," *IEEE Access*, vol. 8, pp. 143201–143218, 2020.

[56] K. Zhang, D. Zhu, J. Jing, and X. Gao, "Learning a cascade regression for no-reference super-resolution image quality assessment," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 450–453, IEEE, 2019.

[57] K. Zhang, D. Zhu, J. Li, X. Gao, F. Gao, and J. Lu, "Learning stacking regression for no-reference super-resolution image quality assessment," *Signal Processing*, vol. 178, p. 107771, 2021.

[58] B. Bare, K. Li, B. Yan, B. Feng, and C. Yao, "A deep learning based no-reference image quality assessment model for single-image super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1223–1227, IEEE, 2018.

[59] Z. Zhang, W. Sun, X. Min, W. Zhu, T. Wang, W. Lu, and G. Zhai, "A no-reference deep learning quality assessment method for super-resolution images based on frequency maps," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3170–3174, IEEE, 2022.

[60] M. U. Rehman, I. F. Nizami, M. Majid, F. Ullah, I. Hussain, and K. T. Chong, "Cn-bsriqa: Cascaded network-blind super-resolution image quality assessment," *Alexandria Engineering Journal*, vol. 91, pp. 580–591, 2024.

[61] T. Tang, F. Yang, X. Lin, and W. Li, "Dual-branch network for no-reference super-resolution image quality assessment," *IEEE Signal Processing Letters*, 2025.

[62] G. Dong, X. Liao, M. Li, G. Guo, and C. Ren, "Exploring semantic feature discrimination for perceptual image super-resolution and opinion-unaware no-reference image quality assessment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28176–28187, 2025.

[63] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[65] M. T. Ribeiro, S. Singh, and C. Guestrin, """ 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*, pp. 1135–1144, 2016.

[66] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[67] Z.-Y. Mi and Y.-B. Yang, "Ram: Interpreting real-world image super-resolution in the industry environment," *Pattern Recognition Letters*, vol. 192, pp. 86–92, 2025.

[68] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[69] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, vol. 2, pp. 416–423, IEEE, 2001.

[70] F. Zhou, R. Yao, B. Liu, and G. Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3528–3541, 2019.

[71] D. Varga, "Empirical evaluation of full-reference image quality metrics on mdid database," *arXiv preprint arXiv:1910.01050*, 2019.

[72] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, "Learning-based quality assessment for image super-resolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3570–3581, 2021.

[73] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, pp. 184–199, Springer, 2014.

[74] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[75] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

[76] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 252–268, 2018.

[77] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11065–11074, 2019.

[78] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018.

[79] X. Liu, A. Jacob, W. Song, and A. Liotta, "Interpreting image super-resolution in artificial neural networks from global and local views," in *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pp. 129–136, IEEE, 2024.

[80] S. Xu, S. Jiang, and W. Min, "No-reference/blind image quality assessment: a survey," *IETE Technical Review*, vol. 34, no. 3, pp. 223–245, 2017.