People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

KASDI MERBAH UNIVERSITY - OUARGLA

Faculty of New Technologies of Information and Telecommunication

Department of Computer Science and Information Technology

Master Thesis

Domain: Computer Science

Field: Artificial Intelligence and Data Science

Explainable Artificial Intelligence for Image Quality Assessment

Presented by: Djezzar Moncef and Sadoudi Abdessamad

Evaluation Date : 11/06/2025

Before the Jury :

Benchabana Ayoub	President	UKM Ouargla
Bouanane Khadra	Examiner	UKM Ouargla
Khaldi Belal	Supervisor	UKM Ouargla

Academic year: 2024/2025

Abstract

The "black-box" nature of many Artificial Intelligence (AI) systems used in Image Quality Assessment (IQA) limits their trustworthiness, especially in critical applications like medical imaging. This thesis introduces a novel framework for eXplainable Artificial Intelligence (XAI) tailored to IQA, demonstrated by detecting and explaining foreign objects in medical X-rays.

Our approach integrates a DeepLabV3+ (ResNet50 backbone) semantic segmentation model with gradient-based XAI methods (Grad-CAM, NormGrad) to provide visual explanations. A key contribution is an advanced *Visualization and Scoring Engine* that processes model outputs and saliency maps to derive nuanced image quality scores based on per-object characteristics (size, location, model confidence) and generates Large Language Model (LLM)-based textual summaries of the assessment.

Evaluated on the Object-CXR dataset via 5-Fold Cross-Validation, the framework demonstrated robust performance. Significantly, XAI applied to the segmentation architecture yielded vastly superior explanation localization (e.g., Grad-CAM Pointing Game Accuracy $\approx$ 0.512) compared to classification-based baselines (PGA $\approx 0.1 - 0.2$). Grad-CAM was found to be more efficient and faithful for this IQA task. The LLM summaries effectively enhanced interpretability. This research presents an end-to-end XAI-IQA system, offering a pathway to more transparent and reliable automated image quality assessment. The developed framework and experimental code are publicly available, fostering reproducibility and further research, at `https://github.com/MoncefDj/Explainable-Artificial-Intelligence-for-Image-Quality-Assessment`.

Keywords: Explainable Artificial Intelligence (XAI), Image Quality Assessment (IQA), Semantic Segmentation, Deep Learning, Grad-CAM, NormGrad, Medical Imaging.

ملخص

إن طبيعة "الصندوق الأسود" للعديد من أنظمة الذكاء الاصطناعي (AI) المستخدمة في تقييم جودة الصور (IQA) تحد من موثوقيتها، خاصة في التطبيقات الحساسة مثل التصوير الطبي. تقدم هذه الأطروحة إطارًا مبتكرًا للذكاء الاصطناعي القابل للتفسير (XAI) مُخصصًا لتطبيقات تقييم جودة الصور (IQA)، ويتجلى ذلك من خلال الكشف عن الأجسام الغريبة في صور الأشعة السينية الطبية وتقديم تفسيرات لوجودها.

يدمج منهجنا نموذج تجزئة دلالية (Semantic Segmentation) من نوع DeepLabV3+ (مع بنية أساسية ResNet50) مع أساليب XAI القائمة على التدرج (gradient-based) (مثل Grad-CAM و NormGrad) لتقديم تفسيرات مرئية. تمثل إحدى المساهمات الرئيسية في تطوير محرك متقدم للتصوير والتقييم (Visualization and Scoring Engine) يقوم بمعالجة مخرجات النموذج وخرائط البروز (saliency maps) لاستخلاص درجات دقيقة لجودة الصورة بناءً على خصائص كل كائن (الحجم، الموقع، ثقة النموذج)، كما يقوم بإنشاء ملخصات نصية للتقييم بالاعتماد على نماذج لغوية كبيرة (LLM).

أظهر الإطار أداءً قويًا عند تقييمه على مجموعة بيانات Object-CXR باستخدام التحقق المتقاطع خماسي الطيات (5-Fold Cross-Validation). بشكل ملحوظ، أدى تطبيق XAI على بنية التجزئة إلى تحديد مواقع التفسير بدقة فائقة (على سبيل المثال، دقة Pointing Game Accuracy لـ Grad-CAM تقارب 0.512) مقارنة بالنماذج المرجعية القائمة على التصنيف (≈ 0.1 – 0.2 PGA). وُجد أن Grad-CAM أكثر كفاءة وموثوقية لمهمة IQA هذه. عززت ملخصات LLM قابلية التفسير بشكل فعال. يقدم هذا البحث نظام XAI-IQA متكاملًا (end-to-end)، مما يوفر مسارًا نحو تقييم آلي لجودة الصور أكثر شفافية وموثوقية. الإطار المطور والكود التجريبي متاحان للعموم، مما يعزز قابلية تكرار النتائج والبحث المستقبلي، على الرابط https://github.com/MoncefDj/Explainable-Artificial-Intelligence-for-Image-Quality-Assessment.

الكلمات المفتاحية: الذكاء الاصطناعي القابل للتفسير (XAI)، تقييم جودة الصور (IQA)، التجزئة الدلالية (Semantic Segmentation)، التعلم العميق (Deep Learning)، Grad-CAM، NormGrad، التصوير الطبي.

Acknowledgements

First and foremost, we would like to express our deepest gratitude to Allah, the Almighty, for His endless blessings, guidance, and strength that have made this work possible.

We would also like to express our sincere appreciation to our supervisor, Prof. Khaldi Belal, for his invaluable guidance, consistent support, and constant encouragement throughout this research journey. His insightful feedback and commitment to excellence have been pivotal in the successful completion of this thesis.

We extend our heartfelt appreciation to all our teachers and professors who have guided us throughout our academic journey. Their dedication to education and their passion for their respective fields have profoundly inspired us, motivating us to strive for excellence. Each of them has played a significant role in shaping our knowledge and character.

A special note of thanks goes to our families, whose constant support, patience, and love have been a source of strength. Their belief in us and their sacrifices have enabled us to persevere through the challenges of this journey, and for that, we are eternally grateful.

We are also deeply thankful to our friends for their companionship, understanding, and continuous encouragement. Their presence has brought joy and motivation, making this journey more meaningful.

Finally, to everyone who has contributed to our education, personal growth, and overall success—whether directly or indirectly—we extend our sincerest gratitude. This achievement would not have been possible without your support and encouragement.

List of Tables

List of Figures

Figure **Page**

Table of Contents

Chapter 1

General Introduction

1.1 Background and Motivation

Contemporary society has witnessed the widespread integration of Artificial Intelligence (AI) systems across numerous domains, where these technologies address complex challenges and modernize traditional approaches [1]. From mobile applications performing diverse computational tasks to automotive safety systems preventing collisions, from financial institutions automating investment decisions to healthcare facilities supporting medical professionals in diagnosis and disease detection, AI applications span virtually every sector [2]. Although AI's foundational concepts emerged decades ago, there is now widespread recognition of the critical role played by intelligent systems that possess capabilities for learning, reasoning, and environmental adaptation. These fundamental characteristics enable AI methodologies to reach remarkable performance levels when addressing increasingly sophisticated computational challenges, positioning them as cornerstone technologies for societal advancement.

Data-driven decision-making systems powered by AI depend on substantial datasets to develop reliable predictive models. Traditional Machine Learning (ML) approaches, including linear regression, logistic regression, and Decision Trees (DT), demonstrate limited effectiveness when dealing with real-world scenarios due to their underlying assumptions about data linearity and simplicity [1]. In contrast, authentic data exhibits high degrees of non-linearity and complexity, presenting significant challenges in extracting meaningful knowledge and actionable insights.

The evolution from early interpretable AI systems to contemporary opaque computational frameworks, particularly Deep Neural Networks (DNNs), marks a significant shift in the field. DNNs excel at processing and extracting patterns from highly complex, multi-dimensional datasets. Research has demonstrated that architectures with greater depth generally outperform shallow networks in decision-making tasks [1]. The remarkable performance of Deep Learning (DL) models results from the synergy between sophisticated learning algorithms and their extensive parameter spaces. Contemporary DNNs, featuring hundreds of layers and millions of parameters, represent quintessential examples of complex black-box systems [2].

The architectural complexity of DNN models stems from multiple interconnected design considerations, encompassing activation function selection, input data characteristics and dimensionality, network depth, pooling strategies, connectivity patterns, classification mechanisms, and the implementation of advanced learning methodologies. The training process itself involves additional critical components, including normalization and regularization techniques, parameter update algorithms, objective functions, and final classification layers. This intricate web of design choices creates systems that, unlike more transparent ML approaches such as DT, Fuzzy rule-based systems (FRBSs), or Bayesian networks (BNs), produce decisions that are inherently difficult to interpret and validate, thus perpetuating the black-box challenge [1].

Figure 1.1 demonstrates the opacity characteristic of DNNs relative to more transparent modeling approaches.

Transparent models, exemplified by linear regression and decision trees, provide inherent interpretability through direct examination of their decision-making processes. However, this clarity often requires sacrificing predictive performance, especially when handling complex, non-linear datasets. Intermediate models, such as certain rule-based systems and shallow neural networks, attempt to balance interpretability with performance, offering moderate transparency while maintaining acceptable accuracy levels. Opaque models, predominantly represented by deep neural networks, typically deliver superior accuracy on complex tasks but operate as black boxes, requiring specialized interpretation techniques to understand their decision-making processes. As emphasized by [1], understanding this interpretability-performance spectrum and implementing appropriate XAI methodologies for opaque systems is fundamental to establishing trustworthy AI frameworks.

1.2 The Motivation for Explainable AI

The deployment of AI systems that cannot provide justifiable, legitimate, or comprehensible explanations for their decisions poses significant risks across various domains.

When human users are hesitant to adopt technologies that lack direct interpretability, tractability, and trustworthiness, the importance of explainability becomes paramount. Although there exists an inherent tension between model performance and transparency, enhanced understanding of system behavior can facilitate the identification and correction of model limitations [2].

Growing concerns within the AI research community regarding the black-box problem have intensified following the establishment of guidelines for developing trustworthy AI systems that prioritize safety and reliability. eXplainable Artificial Intelligence (XAI) methodologies strive to develop ML models that achieve an optimal balance between interpretability and accuracy through two primary approaches: (i) constructing transparent or semi-transparent ML models that maintain inherent interpretability while delivering high performance, or (ii) augmenting opaque models with interpretability mechanisms when transparent alternatives cannot achieve acceptable accuracy levels [1].

Model explanations that support and justify outputs are particularly critical in high-stakes applications such as precision medicine, where medical professionals require comprehensive information beyond simple predictions to inform their diagnostic decisions [2]. As illustrated in Figure 1.1, different modeling approaches exhibit varying characteristics regarding interpretability and accuracy, with XAI techniques serving to bridge this gap for opaque models.

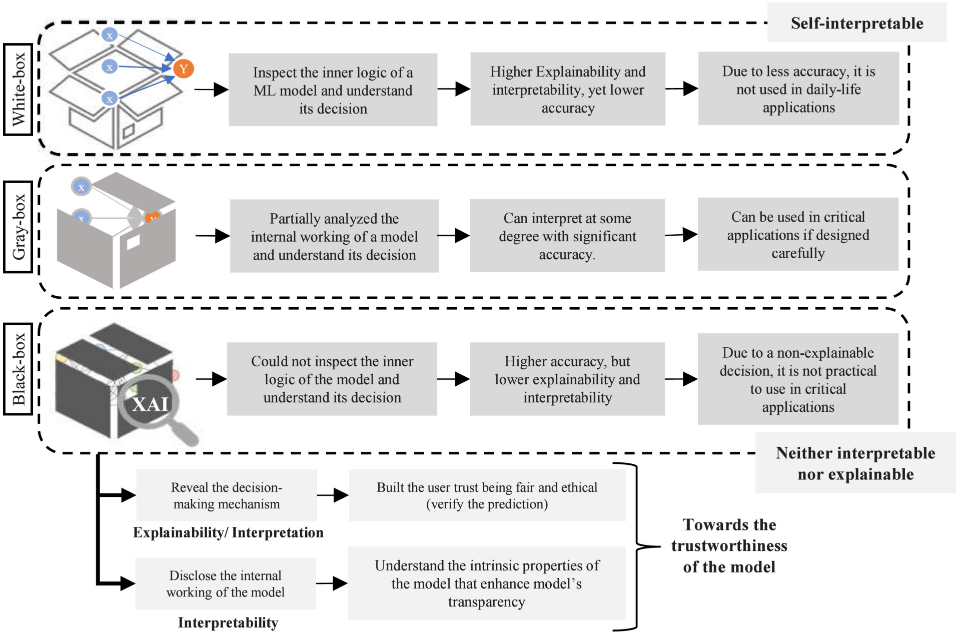

Self-interpretable
White-box
Inspect the inner logic of a ML model and understand its decision
Higher Explainability and interpretability, yet lower accuracy
Due to less accuracy, it is not used in daily-life applications
Gray-box
Partially analyzed the internal working of a model and understand its decision
Can interpret at some degree with significant accuracy.
Can be used in critical applications if designed carefully
Black-box
XAI
Could not inspect the inner logic of the model and understand its decision
Higher accuracy, but lower explainability and interpretability
Due to a non-explainable decision, it is not practical to use in critical applications
Neither interpretable nor explainable
Reveal the decision-making mechanism
Built the user trust being fair and ethical (verify the prediction)
Explainability/ Interpretation
Disclose the internal working of the model
Understand the intrinsic properties of the model that enhance model's transparency
Interpretability
Towards the trustworthiness of the model

Figure 1.1: Comparison of white-box, gray-box, and black-box model characteristics [1].

A significant challenge in establishing unified foundations in this field stems from the inconsistent and interchangeable usage of the terms interpretability and explainability throughout the literature. These concepts possess distinct characteristics that warrant careful differentiation. Interpretability represents an inherent property of a model, describing the extent to which a system's behavior can be understood by human observers. This characteristic is alternatively referred to as transparency [2]. Conversely, explainability constitutes an active property of a model, encompassing deliberate actions or procedures implemented to clarify or elaborate on the system's internal mechanisms and decision-making processes.

The fundamental objectives of XAI encompass the development of ML methodologies that:

1. Generate models with enhanced explainability while preserving high levels of predictive performance.
2. Empower humans to comprehend, appropriately trust, and efficiently collaborate with the next generation of artificially intelligent systems.

Incorporating interpretability as a core design principle during ML model development enhances implementation viability for three fundamental reasons [2]:

- Interpretability promotes fairness in decision-making processes by enabling the detection and subsequent mitigation of bias present in training datasets.
- Interpretability enhances robustness by identifying potential adversarial perturbations that could alter model predictions.
- Interpretability ensures that only relevant variables influence outputs, thereby guaranteeing the existence of meaningful causal relationships in model reasoning.

The growing demand for XAI has experienced substantial growth in recent years, as demonstrated by the increasing volume of research publications in this domain, illustrated in Figure 1.2. Interest in methodologies for explaining AI model behavior has spread throughout the research community, particularly gaining momentum since 2017 [2]. This expanding interest aligns with research priorities established by national governments and regulatory agencies, highlighting the critical importance of responsible AI development practices.

Figure 1.2 demonstrates this trend by presenting the volume of publications incorporating XAI-related terminology in their titles, abstracts, or keywords across recent years. The data, sourced from Scopus through December 10th, 2019 using the search terms displayed in the figure legend, reveals a notable surge in interest beginning approximately in 2017. This pattern supports the observation by [2] that while the requirement for interpretable AI has existed for some time, widespread attention to AI explanation techniques has only recently gained significant traction within the research community.

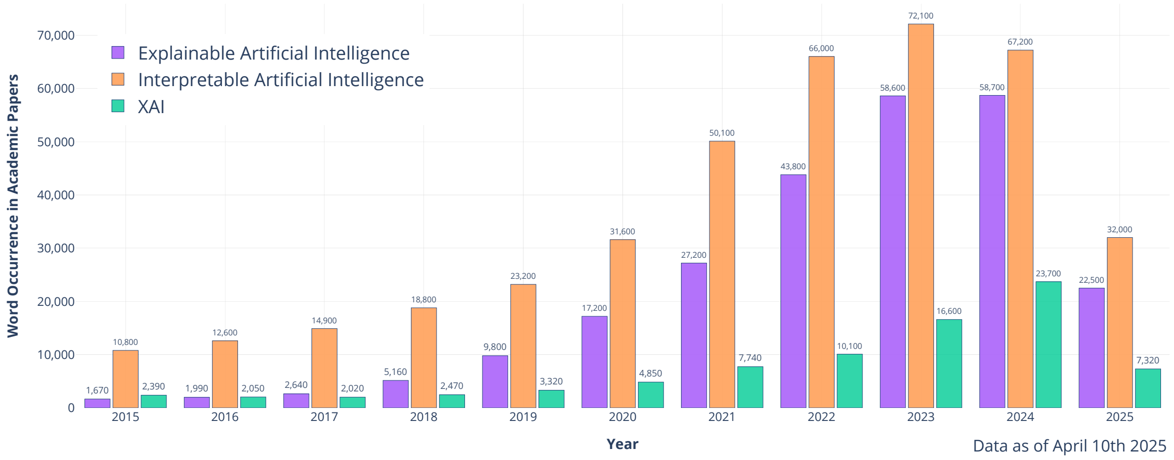


Figure 1.2: Growth in XAI-related publications over recent years, based on Google Scholar data (academic-keyword-occurrence[3]).

1.3 Image Quality Assessment: Fundamentals and Applications

Visual information constitutes one of the most comprehensive data modalities for human information acquisition. Research by Sharma et al. [4] indicates that approximately 57% of human cognitive processing involves visual communication, with roughly 90% of brain-received data being visual in nature. In the contemporary digital landscape, images serve not only as abundant resources but also as fundamental components of countless applications—spanning from medical diagnostic systems to multimedia content development platforms.

Technological advancement has facilitated unprecedented access to vast repositories of visual content through internet platforms for diverse purposes including communication, education, and entertainment. Social media platforms such as Facebook and Instagram process millions of daily image uploads [5], while the recent global pandemic has accelerated the evolution of videoconferencing applications including Zoom, Microsoft Teams, and Skype.

Visual content experiences various forms of degradation throughout its lifecycle, from initial capture and storage to transmission and final display. Image Quality Assessment (IQA) plays a fundamental role in evaluating the perceptual quality of visual content and ensuring reliable operation of dependent applications, including diagnostic systems and content processing pipelines. While conventional IQA metrics (such as PSNR and SSIM) provide computational efficiency and simplicity, they exhibit limitations when addressing complex distortions and frequently fail to correlate with human visual perception [6, 7].

Compression algorithms, particularly JPEG, are extensively employed but may introduce visual artifacts that compromise image quality, especially under high compression ratios. Figure 1.3 demonstrates this phenomenon using an example from [8]. As compression levels intensify (progressing from left to right, corresponding to lower quality settings), distinctive artifacts such as blocking effects and blurring become increasingly apparent, resulting in diminished perceived quality. This degradation emphasizes the necessity for IQA metrics that accurately capture the perceptual impact of compression-induced artifacts.

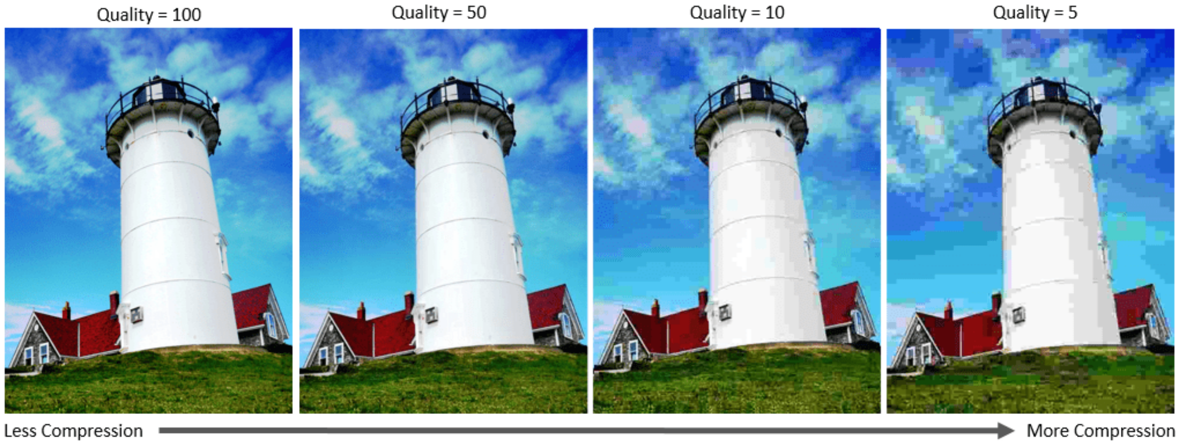


Figure 1.3: Visual impact of increasing JPEG compression levels [8].

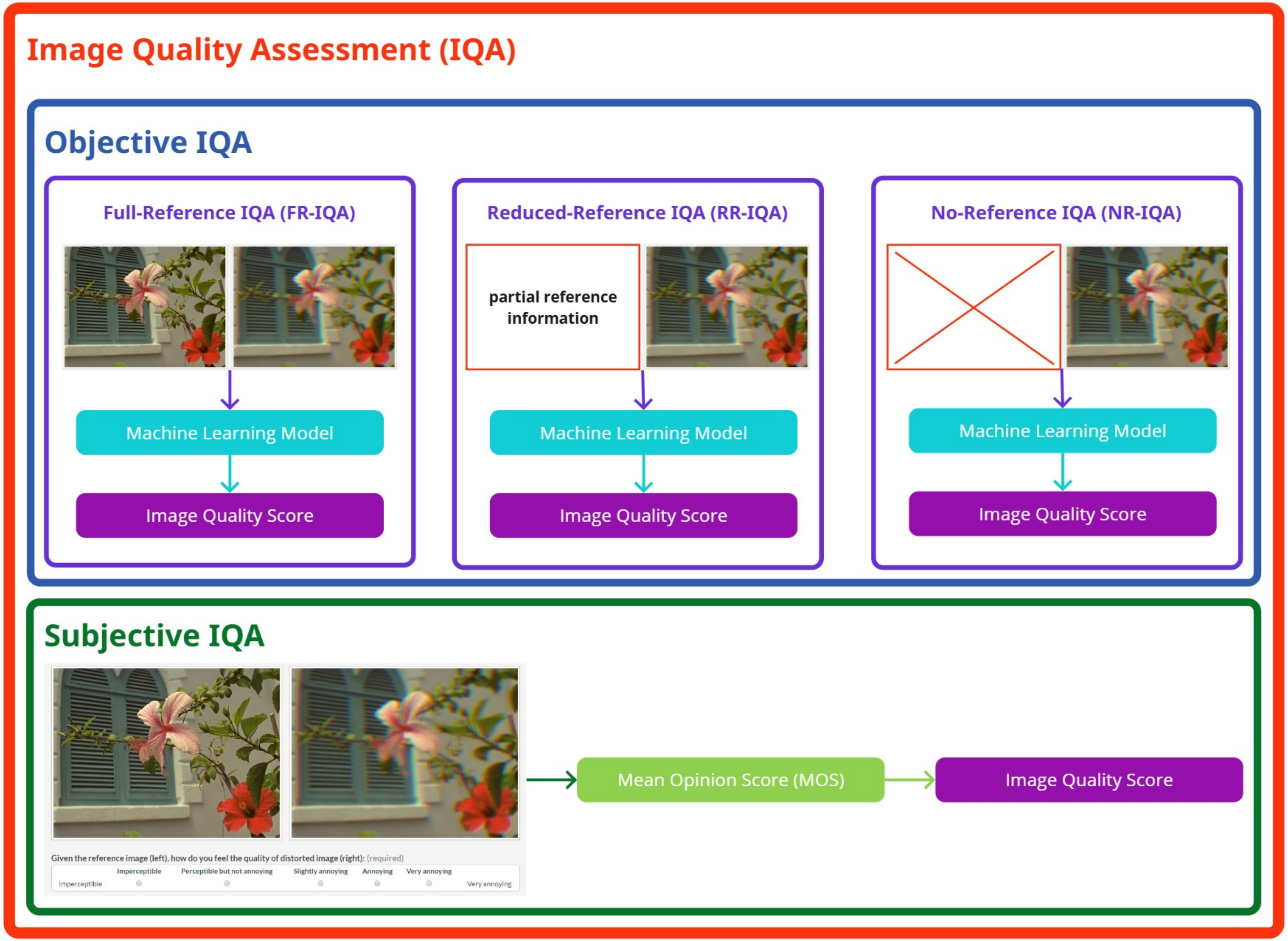
IQA methodologies are categorized into subjective and objective quality evaluation approaches. Subjective quality assessment involves human observers providing quality ratings for images. These evaluations may focus on technical quality aspects or aesthetic preferences, depending on experimental design. The former addresses perceptual distortions including noise, blur, and compression artifacts, while the latter emphasizes visual appeal and artistic merit. This work concentrates on technical quality aspects, which quantify the degree of perceived distortions in images.

Subjective ratings for individual images are typically aggregated across all observers to produce a Mean Opinion Score (MOS), representing the judgment of a statistically representative observer. Subjective quality assessment remains the most reliable method for measuring perceptual image quality due to its foundation in human perception. However, this approach presents several limitations: the need for large observer populations to establish reliable MOS values, extensive preparation and participant recruitment requirements, result reproducibility challenges, and other practical constraints. Consequently, subjective quality assessment is impractical for routine evaluation in image processing algorithms. To address these limitations, objective quality assessment methods have been developed to automatically predict human-perceived image quality.

Objective quality assessment models, also known as image quality metrics (IQMs), provide computational solutions for automated image quality measurement. IQMs are conventionally organized into three distinct frameworks:

- **Full-Reference IQA (FR-IQA):** Evaluates a degraded image by comparison with a pristine reference. When an undistorted reference version of the degraded image is available and accessible, FR-IQA methods estimate image quality through comparative analysis with the reference.
- **Reduced-Reference IQA (RR-IQA):** Utilizes partial reference information for quality evaluation. When limited information such as distortion characteristics or original image histograms is available, RR-IQA methods are employed.
- **No-Reference IQA (NR-IQA):** Assesses image quality without any reference image, frequently employing deep learning for feature extraction [6, 7, 9]. In most practical scenarios, reference data is completely unavailable or non-existent, making NR-IQA methods the appropriate solution for quality assessment. Due to the absence of reference images, NR-IQA is also termed Blind IQA (BIQA).

Figure 1.4 illustrates the comprehensive IQA process, utilizing a sample image from [10]. The figure differentiates between subjective and objective methodologies. Subjective IQA involves human observers evaluating image quality, typically aggregated into a Mean Opinion Score (MOS), representing the gold standard for perceptual quality measurement. Objective IQA, in contrast, employs computational approaches. These are classified according to reference image availability: Full-Reference (FR-IQA) methods require complete reference images, Reduced-Reference (RR-IQA) methods utilize partial reference information, and No-Reference (NR-IQA) or Blind IQA (BIQA) methods operate without reference images, often utilizing machine learning models to predict quality scores based exclusively on the distorted image.



Given the reference image (left), how do you feel the quality of distorted image (right): (required)

	Imperceptible	Perceptible but not annoying	Slightly annoying	Annoying	Very annoying	
Imperceptible	○	○	○	○	○	Very annoying

Figure 1.4: Overview of subjective and objective IQA frameworks (sample image from [10]).

FR-IQA and RR-IQA methods typically demonstrate exceptional performance due to their utilization of reference image information. However, practical applications with available reference data are limited. Consequently, these two IQA categories have restricted applicability in real-world image systems. In contrast, NR-IQA methods exhibit broader applicability to image processing systems. Therefore, this work primarily focuses on NR-IQA methodologies.

Deep learning advancements have transformed IQA through the implementation of sophisticated models including convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformers [7, 11, 12]. However, these state-of-the-art approaches often function as "black boxes," constraining the interpretability of their predictions. As opaque ML models are increasingly deployed for critical decision-making in sensitive contexts, demands for transparency from various AI stakeholders continue to escalate.

1.4 Explainable AI for Image Quality Assessment

As deep learning continues to revolutionize image quality assessment, the challenge of interpretability becomes increasingly pronounced. The black-box nature of advanced IQA models—particularly those based on CNNs, GANs, and transformers—creates a paradoxical situation: while these models achieve remarkable performance in predicting image quality, they offer little insight into how or why they arrive at their assessments. This opacity is particularly problematic in domains where quality assessment directly impacts critical decisions, such as medical diagnostics or content authentication systems. Explainable AI (XAI) methodologies applied to IQA aim to address this tension by rendering quality assessment models more transparent without sacrificing their predictive power. By integrating XAI techniques with IQA frameworks, researchers seek to develop systems that not only accurately evaluate image quality but also provide meaningful explanations for their evaluations. These explanations can range from highlighting distortion-relevant regions to elucidating the relationship between visual features and quality scores, thereby fostering trust, facilitating debugging, and enabling domain experts to validate algorithmic decisions. This section explores the intersection of XAI and IQA, examining both established approaches and emerging methodologies in this rapidly evolving field.

XAI approaches for IQA can be categorized based on their implementation strategy, which generally falls into the following categories:

- **Model-Specific Explainable Methods:** These incorporate interpretability directly into the IQA model architecture, enabling inherent explanations through transparent design [2].
- **Post-hoc Explainability Techniques:** These methods generate explanations after model inference, typically through visualization techniques such as saliency maps or attribution methods [13, 14].
- **Hybrid Approaches:** These combine elements of both strategies, leveraging the strengths of inherent and post-hoc explanations to provide comprehensive insights [1].

In the context of IQA, XAI techniques help in several ways:

- **Identify image regions that predominantly influence quality scores.** By generating visual explanations such as saliency or discrepancy maps, XAI highlights which parts of the image contribute most to the predicted quality score. For example, an explainable IQA system can process a distorted input image to produce not only a quantitative quality score but also a visual explanation, such as a discrepancy map. As illustrated in Figure 1.5 using images from [11], this map highlights the specific regions within the image (e.g., areas with blocking artifacts or blur) that most significantly contributed to the assigned quality score, thus pinpointing the locations of perceived distortions.
- **Understand the role of various deep features in the prediction process.** Through interpretability methods, researchers can discover how convolutional layers or attention mechanisms weigh different image attributes (e.g., texture, color, sharpness). This insight can guide more targeted improvements to model design and training strategies.
- **Provide domain experts (e.g., radiologists) with insights into the model's decisions.** In fields such as medical imaging, transparency is essential for trust and accountability. By revealing the rationale behind quality scores, XAI techniques allow experts to verify the model's performance and detect potential misjudgments, thereby enhancing clinical confidence [13–15].

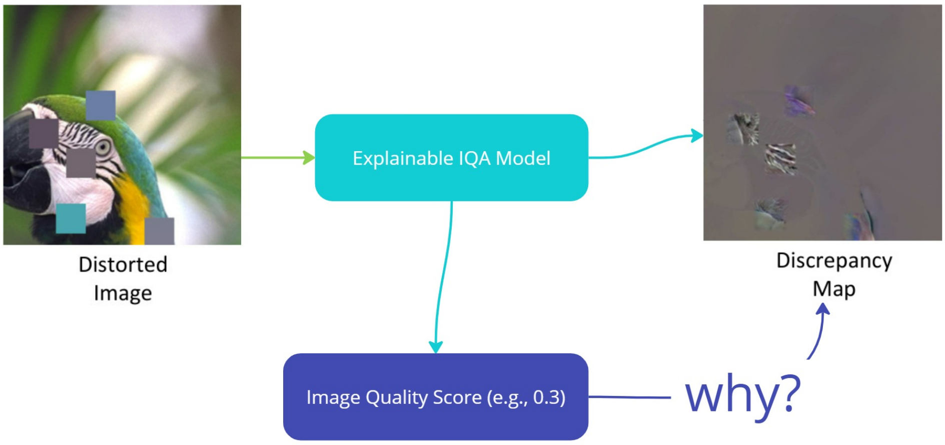


Figure 1.5: An explainable IQA system generating a quality score and a discrepancy map highlighting distortion-relevant regions (sample image and discrepancy map from[11].)

Figure 1.6 illustrates a conceptual pipeline for integrating XAI within an IQA system. The process begins with a raw image input, which is fed into a feature extraction module, typically a deep learning model. These features then guide an attention mechanism, which helps the model focus on relevant areas for quality assessment. The attention mechanism influences both the final image quality prediction and the generation of an explanatory saliency map via an explainability module. This saliency map provides a visual explanation for the predicted score, ultimately enhancing user interpretation and trust in the systemś output.

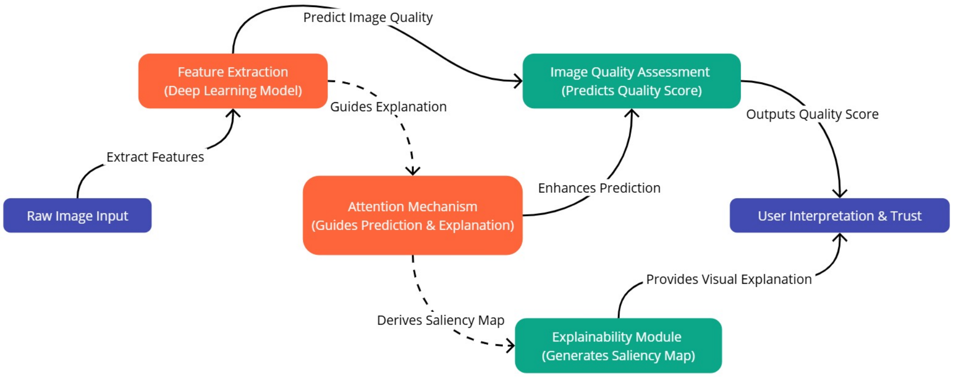

Predict Image Quality
Feature Extraction (Deep Learning Model)
Image Quality Assessment (Predicts Quality Score)
Guides Explanation
Outputs Quality Score
Extract Features
Raw Image Input
Attention Mechanism (Guides Prediction & Explanation)
Enhances Prediction
User Interpretation & Trust
Provides Visual Explanation
Derives Saliency Map
Explainability Module (Generates Saliency Map)

Figure 1.6: Conceptual pipeline integrating feature extraction, attention, quality prediction, and explainability for IQA.

1.5 Real-World Applications

The integration of Explainable Artificial Intelligence (XAI) with Image Quality Assessment (IQA) has transformative potential across diverse fields. As illustrated in Figure 1.7, explainable IQA systems are increasingly pivotal in applications ranging from healthcare to AI-generated content and multimedia. This section outlines key application areas, highlighting both established and emerging use cases.

1.5.1 Healthcare

In healthcare, the interpretability of IQA systems is critical for ensuring reliable diagnostic outcomes and fostering trust among clinicians. The integration of XAI techniques in medical imaging analysis has emerged as a cornerstone of responsible AI implementation [16, 17]. Beyond conventional medical imaging analysis, XAI-enhanced IQA supports several advanced applications:

- **Medical Imaging Analysis:** Visual explanation techniques, such as saliency maps and attribution methods, help highlight diagnostically relevant regions to reduce misinterpretations [18, 19]. For instance, NormGrad-based approaches have demonstrated exceptional performance in localizing image quality issues with a Pointing Game accuracy of 0.862 on foreign objects in chest X-rays [14, 18].

- **Telemedicine & Remote Diagnosis:** High-quality, interpretable imaging enables accurate assessments in remote settings, ensuring that quality issues are identified early. This is particularly relevant in resource-limited environments where in-person specialist evaluation may be unavailable [16].

- **Surgery Assistance & Robotic Guidance:** Real-time feedback from explainable IQA systems can assist surgeons and guide robotic systems during complex procedures, enhancing precision and reducing potential complications [17].

- **Pathology & Histopathology:** In microscopic imaging, interpretable quality assessments help pathologists detect subtle artifacts and maintain diagnostic accuracy, particularly crucial in cancer detection and classification [18, 19].

The application of XAI in healthcare extends beyond technical improvements, addressing crucial ethical and legal considerations. As noted by Mienye et al. [16], XAI has the potential to transform healthcare by making AI-driven medical decisions more transparent, reliable, and ethically compliant. This aligns with the growing emphasis on responsible AI development in the medical domain, where the balance between technological advancement and patient safety is paramount [19].

1.5.2 AI-Generated Content

The surge in AI-generated imagery necessitates robust evaluation mechanisms that combine technical rigor with aesthetic considerations. The proliferation of generative models has introduced new challenges in image quality assessment, particularly regarding the evaluation of synthetic content that may not align with traditional distortion paradigms [20]. In this domain, explainable IQA systems offer:

- **Image Synthesis & Enhancement:** Frameworks like SF-IQA provide an integration of quality and similarity metrics, ensuring that synthetic images are both technically sound and visually appealing [21]. This novel approach addresses the unique challenges of AI-generated image quality assessment by incorporating both local and global-level features for comprehensive evaluation.

- **Video Upscaling & Restoration:** Interpretable quality metrics guide the enhancement and restoration processes, ensuring consistency and clarity in video content. This is particularly relevant for applications such as film restoration and content preservation [6].

- **GAN-Based Quality Assessment:** Dedicated techniques assess the outputs of generative adversarial networks, offering insights into the generative process and its impact on image quality. This addresses the unique challenges posed by GAN-generated content, which may exhibit artifacts not present in traditionally distorted images [20, 21].

- **Deepfake Detection & Authentication:** By revealing underlying cues in synthetic images, XAI methods contribute to robust detection and authentication of manipulated media [20]. This application has gained significant attention due to the potential societal impact of convincing deepfakes in misinformation campaigns.

The evolution of AI-generated content evaluation has shifted from traditional distortion-based models to more sophisticated approaches that consider both perceptual quality and semantic coherence. Yu et al. [21] highlight that prevalent issues with AI-generated images include not only quality concerns but also misalignment between generated images and their corresponding textual prompts. This underscores the need for multi-dimensional assessment frameworks that can evaluate both technical quality and semantic fidelity.

1.5.3 Multimedia & Entertainment

In multimedia applications, maintaining optimal visual quality is essential for enhancing user experience across diverse viewing conditions. Explainable IQA systems enable dynamic adaptation across a range of scenarios, bridging the gap between technical performance and user perception [6, 22]:

- **Streaming Optimization:** Real-time explainable feedback allows adaptive streaming algorithms to adjust image quality in response to network fluctuations, enhancing user experience while optimizing bandwidth usage [6]. This application has gained prominence with the proliferation of video streaming services and the increasing demand for high-quality content delivery.
- **Compression Algorithms:** Quality assessments inform the design of compression techniques that preserve critical visual details while minimizing data loss. Explainable approaches help identify perceptually significant regions that warrant higher bit allocation during compression [6, 22].
- **AR & VR Quality Control:** In immersive environments, explainable IQA ensures that augmented and virtual reality content meets high quality standards, addressing the unique perceptual challenges of stereoscopic and 360-degree content [22].
- **Game Rendering & Evaluation:** Transparent quality metrics help fine-tune rendering processes, leading to smoother and more visually compelling gaming experiences. This application is particularly relevant for real-time rendering scenarios where computational resources must be allocated efficiently [22].

Recent advancements in no-reference IQA for multimedia applications have leveraged vision-language correspondence within a multitask learning framework [22]. This approach recognizes the importance of semantic understanding in quality assessment, particularly for content where context and meaning significantly influence perceived quality.

As a conclusion the integration of XAI with IQA represents a crucial step toward more transparent, reliable, and user-centered visual assessment systems. By providing insights into the decision-making processes of complex algorithms, explainable IQA enhances trust and facilitates meaningful human-AI collaboration across diverse application domains.

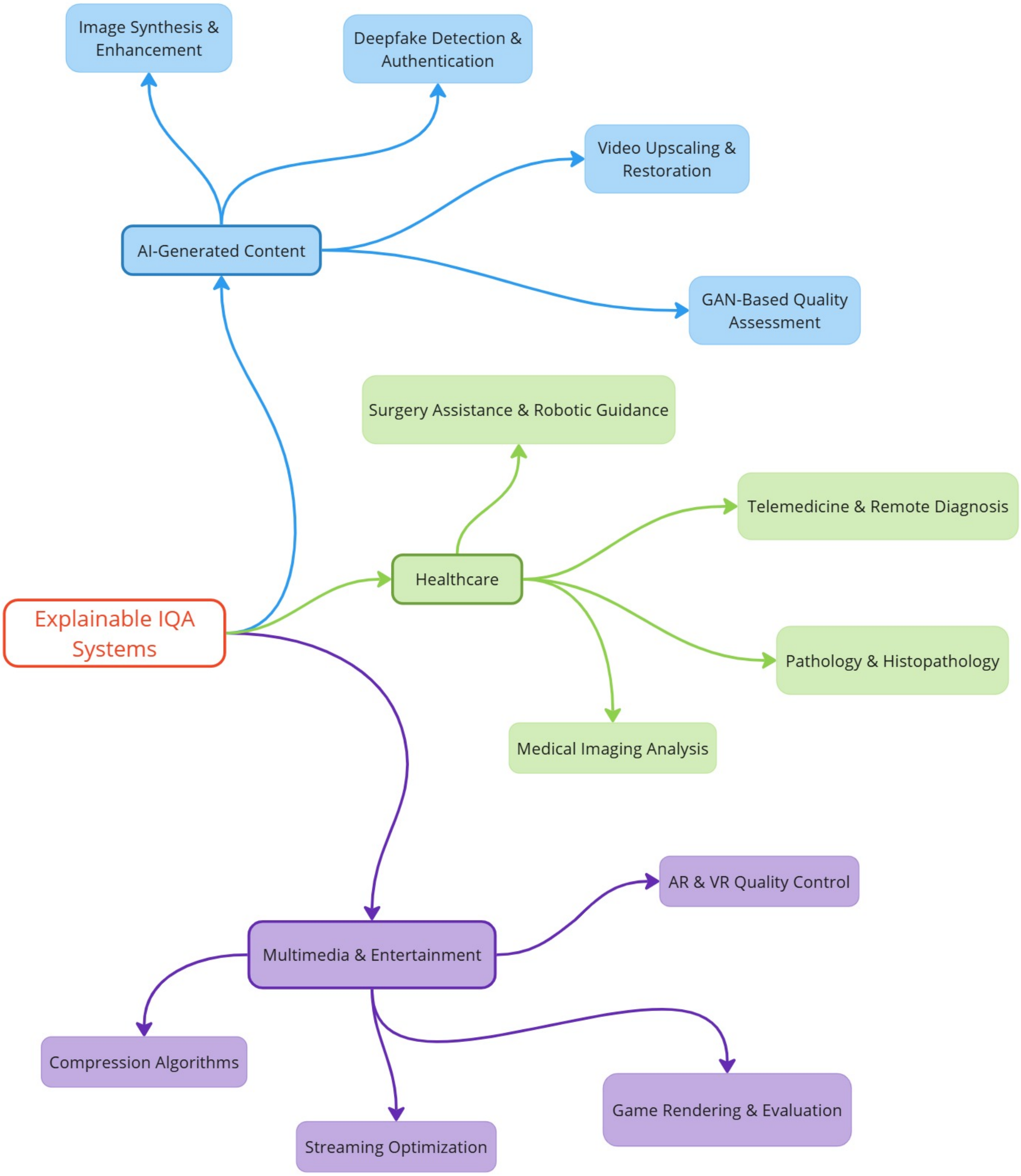


Image Synthesis & Enhancement
Deepfake Detection & Authentication
Video Upscaling & Restoration
AI-Generated Content
GAN-Based Quality Assessment
Surgery Assistance & Robotic Guidance
Telemedicine & Remote Diagnosis
Healthcare
Explainable IQA Systems
Pathology & Histopathology
Medical Imaging Analysis
AR & VR Quality Control
Multimedia & Entertainment
Compression Algorithms
Game Rendering & Evaluation
Streaming Optimization

Figure 1.7: Key application domains for explainable IQA: healthcare, AI-generated content, and multimedia.

As noted by Tjoa and Guan [17], the black-box nature of deep learning models remains a significant challenge, particularly in high-stakes domains such as healthcare where accountability and transparency are paramount.

As AI continues to permeate critical applications, the emphasis on explainability will only grow stronger, driven by regulatory requirements, ethical considerations, and practical needs for effective human-AI collaboration. The real-world applications discussed in this section illustrate the diverse contexts in which explainable IQA can make a meaningful impact, from enhancing diagnostic accuracy in healthcare to improving the quality of synthetic media and optimizing multimedia experiences.

1.6 Current Challenges in XAI for IQA

Although promising advances have been made in integrating XAI with IQA, several challenges that warrant further investigation remain. Recent research has demonstrated potential in developing methods that provide visual explanations for quality assessments, particularly in healthcare imaging and multimedia applications. However, as both images and assessment models grow in complexity, significant obstacles to achieving truly explainable systems persist. These challenges span technical limitations, methodological gaps, and ethical considerations, requiring interdisciplinary approaches to address effectively. Resolving these issues is essential for deploying trustworthy IQA systems in critical applications where explainability directly impacts user acceptance and safety. The following subsections examine these challenges in detail, highlighting current limitations and potential pathways toward more transparent and reliable explainable IQA frameworks.

- **Interpretability vs. Accuracy Trade-off**

  Achieving an optimal balance between high predictive performance and transparent decision-making remains a significant challenge [2, 15]. Traditional IQA models often prioritize performance metrics over explainability, resulting in black-box solutions that, while accurate, provide little insight into their decision-making processes [17]. Conversely, more interpretable models might sacrifice some accuracy for transparency. This dilemma is particularly pronounced in deep learning-based IQA approaches, where complex architectures with millions of parameters deliver state-of-the-art performance but remain fundamentally opaque [20].

  Recent work by [23] and [18] has attempted to address this challenge in medical imaging contexts, demonstrating that techniques like NormGrad [14] can effectively localize image quality issues without significantly compromising performance. However, generalizing these approaches across diverse domains and distortion types remains an open challenge.

- **Computational Complexity**

  Many XAI techniques introduce additional computational overhead, which can limit their applicability in real-time systems. Gradient-based methods such as NormGrad [14] and other backpropagation-based saliency approaches [13] require significant computational resources, especially for deep networks. This complexity becomes problematic for applications requiring real-time quality assessment, such as video streaming services, autonomous vehicles, or live medical imaging systems.

  Furthermore, the computational demands grow exponentially when dealing with high-resolution images or when assessing multiple images simultaneously [7]. Developing more efficient XAI methods for IQA that maintain explanatory power while reducing computational overhead represents a critical direction for future research.

- **Standardization of Evaluation Metrics**

  There is a pressing need for standardized evaluation metrics that assess both image quality and the quality of explanations. Current IQA databases and evaluation protocols primarily focus on prediction accuracy through metrics like Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) [24], with little consideration for explanation quality.

  The community lacks consensus on how to quantitatively evaluate explanations in IQA contexts. While metrics like the Pointing Game accuracy have been employed in specific studies [18], a comprehensive framework for evaluating explanation quality in IQA remains absent. This standardization would facilitate meaningful comparisons between different XAI approaches for IQA and accelerate progress in the field.

- **Domain-Specific Customization**

  Tailoring XAI methods to specific domains, such as medical imaging, requires interdisciplinary collaboration and extensive validation [16, 25]. Different application domains present unique challenges and requirements for explainability. For instance, medical IQA necessitates explanations that align with clinical terminology and diagnostic relevance [19], while consumer photography may require explanations more focused on aesthetic factors.

Research by [23] highlights the importance of domain-knowledge integration in medical imaging quality assessment, demonstrating that explanations must be clinically meaningful to be valuable to healthcare professionals. Similarly, [25] emphasize the need for intelligibility-enriched IQA approaches that can be tailored to specific application contexts.

- **Handling Diverse Distortion Types**

  Modern imaging systems encounter complex mixtures of distortions rather than isolated, synthetic degradations [26]. Traditional IQA databases often contain artificially distorted images with single, well-defined distortion types [10], which may not reflect real-world scenarios where multiple distortions co-occur. Explainable IQA models must be capable of identifying, differentiating, and explaining these diverse and compound distortion types.

  Recent approaches like TOPIQ [27] have attempted to address this challenge by incorporating semantic understanding into quality assessment, enabling more nuanced explanations of diverse distortion types. However, further research is needed to develop XAI techniques that can effectively explain quality issues in images with heterogeneous, realistic distortions [8].

- **Privacy and Security Concerns**

  As XAI techniques provide insights into model behavior, they may inadvertently expose sensitive information or vulnerabilities in critical systems [2]. For instance, explanations might reveal protected attributes in images or expose details about the underlying model architecture that could be exploited for adversarial attacks.

  This concern is particularly relevant for applications handling sensitive data, such as medical imaging or biometric authentication. Balancing transparency with privacy protection remains a challenging task that requires careful consideration of both technical and ethical dimensions [1].

- **Ethical Implications**

  The development of responsible AI systems requires addressing ethical considerations, including fairness, accountability, and privacy [2]. Explainable IQA models must ensure that their assessments and explanations do not perpetuate biases or unfairly penalize certain image types or content categories.

  Studies have shown that IQA models can exhibit biases based on the datasets they were trained on [28]. For instance, if training data predominantly features certain photography styles or content types, models might unfairly assess images that deviate from these patterns. XAI techniques must be designed to identify and mitigate such biases, ensuring fair and equitable quality assessments across diverse image types and cultural contexts [4].

- **Bridging the Gap Between Machine and Human Perception**

  A fundamental challenge in explainable IQA is aligning machine assessments with human perception [6]. While traditional IQA metrics aim to correlate with human quality judgments, they often fail to capture the nuanced and context-dependent nature of human perception [26].

  Recent approaches leveraging vision-language models [21, 22] and multimodal large language models [20] show promise in better aligning with human perceptual judgments. Integrating these advances with XAI techniques could yield explanations that better reflect human reasoning about image quality, making them more intuitive and valuable to users.

- **Transfer Learning and Generalization**

  Developing XAI-IQA models that generalize well across different datasets, domains, and distortion types remains challenging [7]. While transfer learning approaches have shown promise in traditional IQA [5, 17], extending these methods to maintain explainability across domains introduces additional complexity.

  Transitive transfer learning approaches [7] and meta-learning techniques [5] offer potential pathways for improving generalization while preserving explainability. Future research should focus on developing methods that can effectively transfer not only predictive capabilities but also explanatory mechanisms across different domains and tasks.

- **Real-time Adaptation and Learning**

  As imaging technologies and quality standards evolve, XAI-IQA systems must adapt accordingly. Developing models capable of continuous learning while maintaining explainability represents another significant challenge [1]. The dynamic nature of image quality perception, influenced by evolving technologies and changing user expectations, necessitates adaptive systems that can refine their assessments and explanations over time.

  Recent unsupervised and self-supervised approaches [9, 12] offer promising directions for developing more adaptive IQA systems. Integrating these approaches with XAI techniques could yield systems that not only adapt to new quality standards but also provide evolving explanations that reflect these changing paradigms.

The integration of XAI with IQA presents significant opportunities for advancing transparent, trustworthy, and user-centered image quality assessment. By addressing the challenges outlined above, researchers can develop systems that not only accurately predict image quality but also provide meaningful explanations that enhance user understanding and trust.

Future research directions should focus on developing more efficient and generalizable XAI techniques for IQA, standardizing evaluation protocols for explanation quality, and addressing ethical considerations in explainable quality assessment. Interdisciplinary collaboration between computer vision experts, human-computer interaction researchers, and domain specialists will be essential for advancing this field and realizing its full potential.

As AI-generated content becomes increasingly prevalent [21], explainable IQA will play a crucial role in ensuring the quality and reliability of synthetic images. By developing robust, transparent, and ethical XAI-IQA systems, researchers can contribute to a future where intelligent image quality assessment serves as a trusted tool for enhancing visual communication across diverse applications and domains.

1.7 Problem Statement and Research Objectives

’ While modern IQA methods exhibit impressive performance, a significant gap remains in the interpretability of these systems. In high-stakes domains—such as healthcare—understanding the rationale behind a model’s quality score is as critical as the score itself. In response, this thesis aims to bridge this gap by:

1. Critically reviewing state-of-the-art IQA methodologies and identifying their limitations.

2. Proposing novel XAI frameworks that integrate gradient-based saliency methods (e.g., NormGrad) to enhance the transparency of IQA models.

The remainder of this thesis is organized as follows:

- **Chapter 2:** Provides a comprehensive review of the state-of-the-art in IQA and XAI.
- **Chapter 3:** Details the proposed methodologies, including the design and integration of novel explainability techniques into IQA models.
- **Chapter 4:** Describes the experimental setup, evaluation, analysis of the results, and discusses the findings.
- **Chapter 5:** Concludes the thesis by summarizing the contributions, discussing limitations, and suggesting directions for future research.

Chapter 2

State of the Art

The relentless proliferation of digital imagery across diverse technological landscapes, from everyday social media platforms to critical applications in healthcare and autonomous systems, has underscored the paramount importance of accurately assessing image quality. Image Quality Assessment (IQA) aims to develop computational models that emulate human perception, providing essential feedback for optimizing image processing pipelines and ensuring satisfactory user experiences or reliable diagnostic information. Historically dominated by signal processing techniques, the field of IQA has been revolutionized by the advent of deep learning, which offers powerful tools for feature extraction and quality prediction, particularly in challenging No-Reference (NR) scenarios where pristine reference images are unavailable. However, the very success of these deep learning models, often characterized by their immense complexity and millions of parameters, has simultaneously introduced a significant challenge: their inherent opacity. These "black box" models, while achieving high predictive accuracy, lack transparency, making it difficult for users to understand, trust, and debug their outputs.

This lack of transparency is particularly problematic in high-stakes domains, motivating the rapid development of Explainable Artificial Intelligence (XAI). XAI seeks to bridge the gap between complex AI capabilities and human understanding by providing insights into how models arrive at their decisions. Integrating XAI principles and techniques into the IQA workflow is not merely an academic pursuit but a practical necessity. Explainable IQA promises to enhance trust by revealing the rationale behind quality scores, enable more effective model debugging by localizing perceived artifacts or quality-degrading factors, and facilitate fairer comparisons and development of imaging systems by making the assessment process more scrutable.

This chapter delves into the state of the art at the intersection of Image Quality Assessment and Explainable Artificial Intelligence. It aims to provide a comprehensive overview of the key advancements, methodologies, challenges, and synergistic potentials within these two rapidly evolving fields. We begin by surveying the landscape of modern IQA, covering traditional approaches, the impact of deep learning, the development of benchmark datasets, and innovative training paradigms designed to address data scarcity and improve generalization. Subsequently, we explore the core concepts, motivations, taxonomies, and techniques within XAI, focusing on methods relevant to understanding complex models, particularly those used in computer vision and medical imaging. Following this, we specifically examine the nascent but crucial area of integrative approaches, showcasing studies that explicitly combine IQA functionalities with XAI methods to create transparent and trustworthy quality assessment systems. Finally, the chapter concludes with a synthesis of the current state, identifying key research gaps and outlining promising future directions for the development of truly effective and explainable image quality assessment models.

2.1 Image Quality Assessment: State-of-the-Art Methods

Image Quality Assessment (IQA) has established itself as a cornerstone in the domains of computer vision and image processing, aiming to computationally evaluate the perceptual quality of images in a manner consistent with human subjective judgment. The accurate assessment of image quality is paramount across a multitude of applications, ranging from the optimization of image acquisition, compression, and transmission systems to enhancing user experience in multimedia consumption and ensuring diagnostic reliability in medical imaging. As detailed in [6], the field has undergone significant evolution. It has transitioned from classical image processing techniques reliant on handcrafted features and statistical models to sophisticated deep learning paradigms that leverage vast datasets and complex neural architectures. This evolution reflects a continuous effort to bridge the gap between objective computational metrics and the intricate, often subjective, nature of human visual perception.

A comprehensive performance evaluation by [24] highlighted the state of the field, assessing a large number of Full-Reference (FR), fused FR, and No-Reference (NR) IQA methods across diverse datasets, including both singly and multiply distorted images. Their findings underscored the performance gap between FR and NR methods and pointed towards the potential of rank aggregation-based FR fusion as a robust alternative for annotating large-scale datasets where subjective ratings are infeasible, showcasing the maturity yet ongoing challenges within established IQA methodologies.

The development of large-scale, ecologically valid datasets has been crucial for advancing deep learning approaches in IQA. Addressing the limitations of smaller, artificially distorted datasets, [28] introduced KonIQ-10k, a database comprising over 10,000 images captured "in the wild" with diverse content and authentic distortions, annotated via crowdsourcing. This resource aimed to facilitate the training of more generalizable Blind IQA (BIQA) models by providing data more representative of real-world photographic scenarios.

Building upon the need for extensive training data, particularly for deep learning models which often suffer from overfitting on small IQA datasets, [10] proposed a weakly supervised feature learning approach. They introduced the large-scale KADIS-700k dataset (700,000 artificially distorted images without subjective scores but with objective FR-IQA metric scores) for pre-training and the KADID-10k benchmark (10,125 images with subjective scores) for evaluation, demonstrating that features learned via multi-task regression on objective metrics (DeepFL-IQA) can significantly improve NR-IQA performance by effectively transferring knowledge derived from easily computable FR scores.

The challenge of limited labeled data in IQA has spurred innovations in training methodologies beyond simple supervised learning. [5] introduced MetaIQA, a deep meta-learning approach for NR-IQA designed to learn meta-knowledge shared by humans when evaluating images with various distortions. By training on a variety of distortion-specific tasks using a bi-level gradient descent strategy, the resulting meta-model could be quickly adapted to assess images with unknown distortions using fewer samples, demonstrating improved generalization, particularly for authentic distortions, which existing pre-trained models often struggle with due to task mismatch.

Similarly addressing the domain gap and data scarcity issues, [7] proposed TTL-IQA, a transitive transfer learning framework for NR-IQA. This method introduced an auxiliary domain and task, constructed using a novel generative adversarial network based on distortion translation (DT-GAN), to act as an intermediate bridge between the large-scale source domain (e.g., ImageNet) used for pre-training and the specific target IQA domain. By enhancing multi-domain correlation and refining feature transfer using a Semantic Feature Transfer network (SFTnet), TTL-IQA demonstrated improved performance and generalization ability compared to direct transfer learning.

Unsupervised learning presents another promising avenue for circumventing the reliance on large labeled datasets. [9] proposed Re-IQA, employing a Mixture of Experts approach to independently learn high-level content features and low-level quality-aware features using contrastive learning. This unsupervised method, inspired by MoCo-v2 and incorporating novel image augmentation (including various distortion types) and intra-pair image swapping schemes tailored for quality, generates complementary representations that achieve state-of-the-art performance on diverse IQA databases, highlighting the potential of unsupervised methods to learn perceptually relevant features without subjective scores.

Adversarial learning has also been explored to compensate for the lack of reference images in NR-IQA. [11] developed Hallucinated-IQA, which first generates a "hallucinated" reference image constrained by the distorted input using a quality-aware generative network. The perceptual discrepancy between the distorted image and its hallucinated reference, captured in a discrepancy map, is then used, along with an IQA-discriminator and high-level semantic fusion incorporating implicit ranking relationships, to guide a quality regression network, significantly improving prediction accuracy on standard benchmarks by simulating the human comparative process.

Recent advancements have increasingly incorporated Transformer architectures and explored novel assessment paradigms. [12] introduced TReS, an NR-IQA model that uniquely combines Convolutional Neural Networks (CNNs) for local feature extraction with Transformers for modeling non-local dependencies across multi-scale features. Further enhancing robustness, TReS incorporates a relative ranking loss using an adaptive margin based on subjective scores and enforces self-consistency between predictions for an image and its equivariant transformations (like horizontal flipping), achieving state-of-the-art results on multiple benchmarks by capturing both local details and global context effectively.

Focusing on the intrinsic relationship between image understanding and quality, [25] proposed IE-IQA, emphasizing that image quality encompasses both distortion degree and intelligibility (the degree to which image content can be understood). Their framework uses a bilateral network to integrate intelligibility features, derived from semantic understanding tasks (like classification or detection) via feature selection, with conventional distortion features extracted by a separate backbone. By freezing the intelligibility backbone and training the distortion backbone, they showed that incorporating intelligibility significantly improves the generalization ability of the NR-IQA model across diverse datasets.

The specific domain of AI-Generated Images (AIGIs) presents unique quality assessment challenges, requiring evaluation of not only perceptual quality (faithfulness to reality, absence of artifacts) but also alignment with generating prompts (semantic similarity). [21] introduced SF-IQA, a metric specifically for AIGIs that integrates quality and image-text similarity using score fusion. It employs a multi-layer feature extractor for quality-aware features and a fine-tuned vision-language model based on perceptual-aware alignment priors for similarity assessment, achieving high performance on AIGIQA benchmarks by considering both aspects crucial for generated content evaluation.

Further exploring the potential of advanced AI models for IQA, [20] conducted a comprehensive study on using Multimodal Large Language Models (MLLMs) for IQA. They systematically evaluated nine prompting systems (combining psychophysical methods like single/double/multiple-stimulus with NLP strategies like standard/in-context/chain-of-thought prompting) across different visual attributes and scenarios (FR/NR). Their findings revealed that while closed-source models like GPT-4V show promise, particularly with chain-of-thought prompting, significant challenges remain, especially in fine-grained quality discrimination and multi-image comparison, suggesting MLLMs are not yet a replacement for specialized IQA models despite their versatility.

Innovations in network architecture and feature fusion continue to push the boundaries of IQA efficiency and effectiveness. [27] proposed TOPIQ, advocating a top-down approach inspired by human visual processing, where high-level semantics guide the focus on important local distortion regions. Their CFANet (Coarse-to-Fine Attention Network) uses a novel cross-scale attention (CSA) mechanism, progressively propagating semantic information from coarse to fine feature levels via query-key-value attention, and incorporates gated local pooling (GLP) for efficiency. This top-down semantic guidance achieved competitive performance on FR and NR benchmarks with significantly higher efficiency compared to Transformer-based models using only a ResNet50 backbone.

The intersection of vision and language offers another rich source of information for BIQA, enabling multitask learning frameworks. [22] framed BIQA as a multitask learning problem solvable via vision-language correspondence using CLIP. By describing image attributes (quality level, scene category, distortion type) in a textual template and computing the joint probability based on visual-textual embedding similarity, their LIQE (Language-Image Quality Evaluator) model jointly optimizes for all three tasks using automatically weighted fidelity losses. This approach demonstrated improved performance and better realignment of quality scores across different datasets, verifying that BIQA can benefit from auxiliary knowledge, even from conceptually conflicting tasks like scene classification, when mediated through a shared vision-language space.

Addressing the challenge of reference-free IQA, particularly for degradation and reconstruction artifacts where clean references are often unavailable (e.g., in medical imaging), [8] developed a lightweight, fully convolutional network trained in a self-supervised manner. By training the network simply to predict the JPEG Quality Factor (QF) of randomly compressed image patches, the model implicitly learns to recognize a wide range of degradation artifacts. This "QF Predictor" demonstrated surprising generalization capabilities, accurately measuring the severity of other artifacts like Gaussian blur, noise, and even undersampling artifacts in Magnetic Resonance Imaging (MRI) reconstructions without needing task-specific training or reference images, showcasing the versatility of self-supervised learning driven by a simple proxy task.

2.2 Explainable AI: Techniques and Methodologies

While deep learning models have demonstrated remarkable predictive power in IQA and other domains, their inherent complexity often renders them "black boxes," limiting user trust and hindering adoption, particularly in high-stakes applications like healthcare and autonomous systems. Explainable Artificial Intelligence (XAI) has emerged as a critical field dedicated to developing methods that make AI decisions understandable to humans [2]. The imperative for XAI stems from various needs: justifying algorithmic decisions to ensure fairness and accountability, controlling system behavior by identifying vulnerabilities and flaws, improving models through better understanding of their internal workings, and discovering new knowledge encoded within complex representations [29]. The ultimate goal is to transform opaque models into transparent or at least interpretable systems, fostering trust, accountability, and effective human-AI collaboration by providing insights into the 'how' and 'why' behind AI predictions and actions.

Comprehensive surveys on XAI, such as the work by [29], provide a broad overview of the field, tracing its roots from the explanation facilities in early expert systems to the challenges posed by modern machine learning, especially DNNs. They emphasize the need for techniques that can elucidate the reasoning process without necessarily sacrificing predictive accuracy, categorizing approaches based on dimensions like complexity (inherently interpretable vs. post-hoc explanation), scope (local vs. global interpretability), and model dependency (model-agnostic vs. model-specific methods).

[2] further expanded on this by proposing a detailed taxonomy of XAI methods, meticulously distinguishing between inherently transparent models (like linear regression or decision trees under certain constraints) and post-hoc explainability techniques applied to black boxes (categorized by explanation type, such as text, visual, local, example-based, simplification-based, or feature relevance). Their work crucially underscores the importance of the target audience (e.g., domain experts, data scientists, affected users, regulators) in shaping appropriate explanations and introduces the overarching concept of Responsible AI, where explainability serves as a cornerstone alongside fairness, accountability, and privacy, guiding the ethical deployment of AI systems.

Further elaborating on the path towards trustworthy AI, [1] presented a systematic review focusing on the assessment of explanations themselves, alongside available tools, datasets, and the practical application of XAI methods. They highlight the importance of evaluating explanations based on criteria like faithfulness and plausibility and discuss the challenges in standardizing such evaluations. The survey advocates for tailoring explanation content to specific user types and consolidating the current state of knowledge to guide future research in creating AI systems that are not only powerful but also reliable and accountable.

Understanding what constitutes a "good" explanation is central to XAI and intrinsically involves considering the perspectives and needs of various stakeholders who interact with or are affected by AI systems. [15] directly addressed this by proposing a conceptual model centered on stakeholders' desiderata (their interests, goals, expectations, and needs). The model posits that explainability approaches provide explanatory information, which facilitates stakeholder understanding, ultimately affecting the satisfaction of these desiderata within a specific context. They argue that the success of XAI hinges on this chain and emphasize the crucial role of interdisciplinary collaboration (involving psychology, philosophy, law, sociology, and computer science) to identify relevant desiderata, develop appropriate explainability approaches, and empirically evaluate their effectiveness in fostering the necessary understanding to satisfy those desiderata.

The application of XAI is particularly critical and faces unique challenges in sensitive domains such as healthcare, where algorithmic decisions can directly impact human lives and well-being. [16] provided a survey specifically focused on XAI in healthcare, discussing core concepts, diverse applications (like clinical decision support, diagnostic assistance, and treatment planning), and significant challenges. These challenges include balancing the inherent trade-off between model interpretability and predictive accuracy, seamlessly integrating XAI methods into established clinical workflows, and ensuring strict adherence to rigorous regulatory and ethical standards, all while aiming to make AI-driven medical decisions more transparent, reliable, and trustworthy for both clinicians and patients.

Similarly, [17] offered a survey oriented specifically towards Medical XAI, categorizing various interpretability methods (from inherently interpretable models to post-hoc techniques like saliency mapping and feature attribution) and applying this categorization to the medical context. They highlight the crucial need for clinicians and practitioners to approach AI-generated explanations with caution, understanding their limitations, while simultaneously advocating for data-driven, mathematically grounded, and technically informed medical education to responsibly leverage XAI's potential in improving healthcare delivery and patient outcomes.

Focusing specifically on the rapidly growing use of deep learning in medical image analysis, [19] presented a systematic overview of XAI techniques applied in this subfield. They classified methods based on a practical framework considering model-based vs. post-hoc, model-specific vs. model-agnostic, and global vs. local explanation scopes. Their survey revealed that visual explanations, particularly saliency maps highlighting relevant image regions, are the most common approach used by researchers to interpret deep learning models for tasks like classification and segmentation in medical imaging, while also pointing out the existing gap in the rigorous evaluation and validation of these explanation techniques within the clinical context.

Developing and refining specific XAI techniques, especially those applicable to complex deep learning models used in vision tasks like IQA, remains an active and vital research area. Saliency methods, which aim to identify the input features (pixels or regions) most relevant to a model's output, represent a popular class of post-hoc, often model-specific, techniques. Addressing limitations of earlier gradient-based methods, [14] introduced NormGrad, a principled attribution method derived directly from the gradient computation of convolutional layer weights. By considering the norm (specifically, the Frobenius norm for the outer product) of local gradient contributions—decomposed into activation gradients and input features at each spatial location—NormGrad offers an efficient way to generate fine-grained importance maps throughout the network. This method potentially highlights image regions crucial for model training updates, offering a different perspective than methods focused solely on inference pathways.

Building on this foundation and seeking to unify and clarify the relationships between various saliency approaches, [13] presented a comprehensive framework for backpropagation-based saliency methods, encompassing techniques like simple gradient visualization, linear approximation (gradient × input), and NormGrad variants. Their analysis systematically decomposed these methods into 'extract' (isolating spatial gradient contributions) and 'aggregate' (transforming contributions into a heatmap) phases. This framework facilitated a comparative study, revealing, for instance, why methods like Grad-CAM often fail at layers other than the final convolutional one due to spatial averaging eliminating class-sensitive gradient information. Importantly, they demonstrated that combining saliency maps from multiple network layers often improves performance (e.g., in weak localization tasks) and introduced meta-saliency—a novel technique inspired by meta-learning—which incorporates an inner optimization step to significantly enhance the class-sensitivity of any backpropagation-based saliency method, making the resulting explanations more specific to the target prediction.

2.3 Integrative Approaches (IQA + XAI)

The convergence of Image Quality Assessment (IQA) and Explainable Artificial Intelligence (XAI) represents a significant frontier, particularly crucial in domains where both the accuracy of quality prediction and the transparency of the assessment process are paramount. While IQA focuses on determining what the perceived quality of an image is, potentially summarizing it with a single score, XAI provides insights into why a certain quality score is assigned or where the quality issues contributing to that score are located within the image. This integration is especially vital in high-stakes areas like medical imaging, where understanding the basis for an automated quality assessment—identifying artifacts or confirming the clarity of relevant anatomy—can directly influence diagnostic confidence, workflow efficiency, and patient safety. The development of methods that not only predict quality but also explain their predictions allows users, such as clinicians reviewing scans or engineers developing acquisition protocols, to trust, verify, and potentially debug the automated assessments, moving beyond opaque scoring towards actionable insights.

Applying XAI techniques, particularly saliency methods, to IQA tasks in medical imaging offers a clear pathway to building more trustworthy and useful automated systems. [18] provided an early demonstration of this synergy by developing an explainable system specifically for assessing the quality of Chest X-rays, focusing on the detection of foreign objects (like buttons or clips) which can obscure pathology and degrade diagnostic utility. Their pipeline utilized a ResNet-based classifier to distinguish between images with and without such objects. Crucially, they employed NormGrad [13, 14], an XAI method designed to produce fine-grained saliency maps based on gradient norms, to visually highlight the image regions corresponding to the foreign objects that drove the classifier's decision. Through quantitative evaluation using the Pointing Game metric, which measures the alignment between the saliency map's peak activation and ground-truth bounding boxes, they showed that NormGrad significantly outperformed other common saliency methods like Grad-CAM in accurately localizing these specific quality-degrading elements, thereby providing not just an automated quality check but also a visual explanation for it.

Generalizing this concept beyond Chest X-rays and foreign object detection, [23] extended the application of explainable IQA to Cardiac Magnetic Resonance (CMR) imaging, focusing this time on detecting the inappropriate appearance of the Left Ventricular Outflow Tract (LVOT) in four-chamber views—a quality issue arising from suboptimal scan planning that can hinder atrial assessment. They again leveraged NormGrad within their classification pipeline and, importantly, conducted a more comprehensive and rigorous evaluation of the generated explanations. This involved comparing multiple saliency detectors (Input x Grad, Guided Backpropagation, Grad-CAM, Guided Grad-CAM, and various NormGrad configurations) using not only the repeated Pointing Game for localization accuracy but also incorporating robustness checks through model randomization and reproducibility tests across different network architectures (ResNet vs. EfficientNet). They introduced a Difference of Means (DoM) metric to quantify consistency across architectures. Their results robustly demonstrated the superior performance (higher Pointing Game accuracy) and consistency (lower DoM scores, especially for combined-layer NormGrad) of NormGrad in localizing specific image quality issues within medical scans compared to other baseline saliency techniques. These studies collectively exemplify how integrating specific XAI methods like NormGrad directly into the IQA workflow can transform automated quality assessment from a simple scoring mechanism into an interpretable diagnostic tool, providing localized, visual evidence that underpins quality judgments, thereby enhancing transparency, facilitating error analysis, and ultimately building greater trust in these systems, particularly within the demanding medical domain.

Beyond these direct medical applications, the principles of XAI can enhance or be informed by a broader range of IQA research. Many advanced IQA models, while not always explicitly framed as XAI systems, incorporate mechanisms or produce outputs that lend themselves to explainability. For instance, attention mechanisms inherent in some modern architectures can offer insights. [27], with its Coarse-to-Fine Attention

Network (CFANet) and cross-scale attention (CSA), not only achieves efficient quality prediction but its attention maps can serve as visual explanations for the model's focus on semantically important regions when assessing distortions. Similarly, Transformer-based models like [12], often utilize attention layers whose visualizations can illuminate which image patches most influence the quality score, providing a degree of spatial explainability.

The integration of richer semantic understanding and multimodal information into IQA also opens avenues for more intuitive explanations. [22] leverages CLIP to associate image quality with textual descriptions of attributes like quality level, scene, and distortion type. The learned correspondence itself can form the basis of an explanation, e.g., "this image is high quality because it aligns well with 'sharp, clear photograph'". This approach allows the model to potentially explain quality by linking it to understandable semantic concepts. Expanding on this, [20] explores how MLLMs can be prompted to not only predict quality scores but also generate natural language rationales for their assessments. While still facing challenges in fine-grained discrimination, the potential for MLLMs to provide human-readable justifications represents a promising XAI direction for IQA.

Furthermore, models that explicitly consider image content and intelligibility, such as [25], offer another angle for XAI. By disentangling distortion features from intelligibility features, it becomes potentially feasible to explain a quality score in terms of how well the image content can be understood, or which specific semantic features contribute positively or negatively to the perceived quality. Unsupervised learning approaches, like [9], which learn distinct content and quality-aware features, could also contribute to explainability if these learned features can be mapped to human-interpretable concepts or visually localized, thereby offering insights into what aspects the model deems important for quality without explicit human labeling of these aspects.

Finally, comprehensive evaluations and comparative studies in IQA can also benefit from an XAI perspective. [24] and [6] meticulously benchmark various IQA methods. While these studies identify high-performing algorithms, XAI techniques could be employed to delve deeper, explaining "why" certain models outperform others on specific datasets or distortion types. Such insights would move beyond performance leaderboards to a more fundamental understanding of model behavior, guiding the development of more robust and generalizable IQA solutions by revealing their underlying decision-making processes and failure modes. This analytical use of XAI can help interpret the results of large-scale studies and steer future research more effectively.

2.4 Summary and Conclusions

This review has traversed the evolving landscapes of Image Quality Assessment (IQA) and Explainable Artificial Intelligence (XAI), culminating in their crucial integration. The state of the art in IQA has progressed substantially from traditional metrics to sophisticated deep learning models capable of handling diverse distortion types, including synthetic, authentic, and algorithm-generated artifacts, as highlighted in comprehensive studies such as [6, 21, 24]. Innovations in network architectures (CNNs, Transformers as seen in [12]), training strategies (unsupervised learning [9], weak supervision [10], meta-learning [5], adversarial learning [11], transfer learning [7]), and leveraging auxiliary information (intelligibility [25], semantics [27], vision-language correspondence [20, 22]) continue to enhance prediction accuracy and generalization. The development of large-scale, ecologically valid datasets [10, 28] remains critical for training robust deep models capable of performing well "in the wild."

Concurrently, the field of XAI has gained significant prominence, driven by the pressing need to demystify the "black box" nature of complex AI models, thereby increasing transparency, trust, and accountability [2, 29]. Research in XAI spans a wide spectrum, encompassing the development of inherently interpretable models and a diverse array of post-hoc techniques—such as saliency mapping [13, 14], rule extraction, and example-based explanations—designed to provide insights into model behavior after training. The importance of tailoring explanations to specific stakeholder needs [15] and considering the unique context of deployment, especially in sensitive domains like healthcare [16, 17, 19], is increasingly recognized as essential for creating explanations that are not only technically accurate but also meaningful, actionable, and trustworthy [1].

The integration of XAI methodologies into the IQA pipeline, particularly demonstrated in the context of medical image quality assessment [18, 23], and the broader potential for explainability in advanced IQA systems (e.g., via attention in [27] or multimodal reasoning in [20, 22]) represents a vital step towards building automated assessment systems that are both reliable and deployable in practice. By utilizing XAI methods, such as NormGrad-based saliency mapping, to visually pinpoint the sources of quality degradation or highlight the image features driving a particular quality score, these integrative approaches significantly enhance transparency. This capability allows users not only to receive a quality score but also to understand its basis, facilitating debugging, building user confidence, and enabling more informed decision-making.

Despite these advancements, significant challenges and compelling research opportunities remain at the intersection of IQA and XAI. A primary challenge is the lack of standardized metrics and robust protocols specifically designed to evaluate the faithfulness, utility, and robustness of explanations within the specific context of IQA. While localization metrics like the Pointing Game [18] provide a starting point, developing methods to assess how well an IQA explanation aids user understanding, task performance, or trust calibration requires substantial further investigation [15].

Another crucial future direction lies in the development of inherently interpretable IQA models, moving beyond the predominant reliance on post-hoc explanations for opaque black-box architectures. This could involve exploring architectures incorporating attention mechanisms designed for transparency [27], modular network designs, or hybrid approaches.

Furthermore, the application of diverse XAI techniques beyond saliency maps needs deeper exploration within IQA. Techniques such as counterfactual explanations, concept-based explanations, or example-based reasoning could offer different insights tailored to specific IQA tasks (e.g., assessing AI-generated content quality [21] and diverse user needs.

Finally, navigating trade-offs between explainability, accuracy, efficiency [27], and privacy [2] remains complex. Designing truly responsible and effective explainable IQA systems will necessitate careful consideration of these factors and continued interdisciplinary collaboration.

To facilitate a clear comparison and synthesis of the surveyed literature, Table 2.1 provides a comprehensive overview of the key papers discussed in this chapter. It details the methodologies, techniques, datasets, evaluation approaches, main findings, and limitations identified for each study, offering a structured summary of the current landscape at the intersection of IQA and XAI. Furthermore, Table 2.2 complements this by summarizing the specific empirical metrics and key quantitative results reported in these studies, allowing for a more direct comparison of the performance and characteristics of different approaches.

Table 2.1: State of the Art Summary

		Title (Year)	Methodology	Techniques	Datasets	Evaluation	Key Findings	Limitations
NR-IQA	IQA	Reference-Free Image Quality Metric for Degradation and Reconstruction Artifacts [8] (2024)	Deep Learning (CNN), Self-supervised Learning.	Lightweight FCN (7 layers) trained to predict JPEG Quality Factor (QF). Generalized to predict Gaussian Blur, Gaussian Noise severity, and MRI undersampling rate.	Self-supervised via JPEG compression. Tested on Flickr1024, fastMRI, ImageNet, LIVE.	Correlation curves (predicted QF vs. artifact severity/level). Average QF prediction for dataset quality estimation. Tested as perceptual loss.	CNNs can learn reference-free IQA via self-supervision on JPEG QF. Generalizes well to other common artifacts and MRI reconstruction artifacts. Can estimate overall dataset quality.	Application as a direct loss function is suboptimal without careful weighting against data consistency. MRI whole-image quality estimation less stable than patch-based.
IQA (MLLM)	IQA	A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment [20] (2024)	Comprehensive study comparing Multimodal Large Language Models (MLLMs) for IQA using various prompting strategies.	Compared 9 prompting systems (psychophysics methods + NLP strategies). Used difficult sample selection (via expert IQA models). Compared 3 open-source MLLMs vs. GPT-4V.	FR-KADID, Aug-KADID, TQD, SPCD (FR); NR-KADID, SPAQ, AGIQA-3K (NR).	SRCC comparisons across MLLMs, prompting systems, and datasets covering various visual attributes.	GPT-4V is reasonable but weak on fine-grained quality (color) and multi-image comparison. Open-source MLLMs perform poorly, especially in FR/multi-image scenarios. Chain-of-Thought prompting helps GPT-4V.	Only GPT-4V shows promise. Open-source models struggle significantly. Fine-tuning might cause catastrophic forgetting. Quantitative evaluation of text responses is hard.
AIGCIQA	IQA	SF-IQA: Quality and Similarity Integration for AI Generated Image Quality Assessment [21] (2024)	Score Fusion approach combining separate quality and similarity assessments for AI-Generated Images (AIGIs).	Quality Branch: Multi-layer feature extraction (Swin Transformer V2) fusing local/global features. Similarity Branch: Fine-tuned vision-language model (PickScore). Score Fusion Module: MLP fusing scores.	Pre-trained on various NR-IQA and AIGC-IQA datasets. Tested on AGIQA-3K and AIGIQA-20K (NTIRE 2024).	SRCC, PLCC on AGIQA-3K (overall and model subsets). NTIRE 2024 challenge ranking. Ablation studies on branches and fusion methods.	Integrating quality and similarity via score fusion improves AIGI QA. Achieved SOTA on AGIQA-3K and 4th in NTIRE 2024 challenge. Independent feature extraction and inner product similarity calculation work best.	Performance drops for 'bad' AIGI models (far from naturalistic). Relies on pre-trained models (Swin, PickScore).

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	Methodology	Techniques	Datasets	Evaluation	Key Findings	Limitations
IQA Review	IQA	Advancements in image quality assessment: a comparative study of image processing and deep learning techniques [6] (2024)	Literature review comparing traditional image processing (IP) techniques and Deep Learning (DL) methods for IQA.	Discusses IP (PSNR, SSIM, noise reduction, segmentation, feature extraction) and DL (CNNs, GANs, Transformers).	General application domains (media, medical, automotive, surveillance, remote sensing).	Comparative discussion of methodologies, strengths, limitations, and future directions (hybrid models, HVS-inspired, domain adaptation, lightweight models).	DL methods generally offer better accuracy, adaptability, and perceptual relevance than traditional IP techniques, especially for NR-IQA.	Review nature; no new method or quantitative comparison across specific datasets.
XAI Review (Healthcare)	XAI	A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges [16] (2024)	Comprehensive review of XAI in healthcare focusing on concepts, applications, and challenges.	Discusses various XAI techniques (categorization less technical than others). Emphasizes bias, ethics, fairness, trustworthiness.	General healthcare applications.	Discusses need to balance interpretability/accuracy, clinical workflow integration, regulatory standards.	XAI crucial for transparency, reliability, ethics in healthcare AI. Highlights need for practical application guidance.	High-level review, less technical detail on methods.
IQA (FR/NR)	IQA	TOPIQ: A Top-down Approach from Semantics to Distortions for Image Quality Assessment [27] (2023)	Deep Learning (CNN), Top-down approach using multi-scale features and attention.	Proposes Coarse-to-Fine Attention Network (CFANet). Uses Cross-Scale Attention (CSA) guided by high-level features. Introduces Gated Local Pooling (GLP) for efficiency. Uses ResNet50 backbone.	Tested on FR (LIVE, CSIQ, TID2013, PieAPP, PIPAL, BAPPS) and NR (CLIVE, KonIQ-10k, SPAQ, AVA, FLIVE) datasets. Uses ImageNet pre-training.	SRCC, PLCC comparisons (intra-/cross-dataset). Attention map visualization. Ablation studies on components (GLP, SA, CSA, Pos.). Backbone analysis (ResNet, Swin). Computational complexity (FLOPS).	Top-down semantic guidance improves multi-scale feature utilization for IQA. CFANet achieves SOTA/competitive performance on FR/NR benchmarks. More efficient (FLOPS) than SOTA Transformer models (e.g., AHIQ). GLP and CSA are effective components.	Still relies on pre-trained backbone (ResNet50). Performance gap between ResNet and Swin backbones suggests potential for further improvement. Evaluation primarily on standard benchmarks.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	Methodology	Techniques	Datasets	Evaluation	Key Findings	Limitations
BIQA (Multitask)	IQA	Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective [22] (2023)	Multitask learning for BIQA using vision-language correspondence (CLIP). Learns BIQA jointly with scene classification and distortion type identification.	Uses pre-trained CLIP. Describes tasks via textual templates. Computes joint probability via cosine similarity. Uses fidelity losses and dynamic loss weighting.	Trained/Tested on LIVE, CSIQ, KADID-10k, CLIVE, BID, KonIQ-10k. Cross-tested on TID2013, SPAQ, PIPAL.	SRCC, PLCC (intra-/cross-dataset). gMAD competition. MOS realignment analysis. Ablations on tasks, components.	BIQA benefits from auxiliary tasks. LIQE outperforms SOTA BIQA, generalizes better, realigns MOS scales effectively. Vision-language correspondence is effective for multitask BIQA.	Relies on pre-trained CLIP. Performance might be limited by CLIP's capabilities. Needs labeled data for scene/distortion tasks during training.
XAI (Medical IQA)	XAI + IQA	Explainable Image Quality Assessment for Medical Imaging [23] (2023)	Proposes explainable medical IQA pipeline using saliency methods (NormGrad) to localize quality issues.	Uses NormGrad for saliency maps. Compares with Grad-CAM, Guided Grad-CAM, Guided Backprop, Input x Grad. Evaluates faithfulness via Pointing Game, smoothing effects, randomization tests, repeatability, reproducibility (ResNet vs EfficientNet) using proposed DoM metric.	Object-CXR (foreign objects), LVOT dataset (cardiac MRI quality issue).	Pointing Game accuracy, Difference of Means (DoM), qualitative visualization.	NormGrad saliency maps are more accurate and consistent than baselines for localizing medical IQA issues. Combined-layer NormGrad improves precision. Trained models are essential for meaningful saliency maps.	Pointing Game metric less discriminative for multiple ROIs. NormGrad may not be ideal for segmenting large structures.
BIQA (Unsupervised)	IQA	Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild [9] (2023)	Unsupervised contrastive learning (Mixture of Experts) to learn separate content-aware and quality-aware image representations.	Content encoder trained via MoCo-v2 on ImageNet. Quality encoder trained via MoCo-v2 with novel augmentation bank, OLA-cropping, and Intra-pair Swapping on diverse images. Linear regressor trained on concatenated frozen features.	Training: ImageNet (content); KADIS, AVA, COCO, etc. (quality). Evaluation: KonIQ, CLIVE, FLIVE, SPAQ (authentic); LIVE, TID2013, CSIQ, KADID (synthetic).	SRCC, PLCC comparisons. t-SNE visualization. Ablations.	Learns complementary content/quality features unsupervisedly. Achieves SOTA/competitive performance. Unsupervised content features outperform supervised ones for IQA. Content dominates for authentic, quality for synthetic.	Computationally intensive training. Performance depends on contrastive framework effectiveness and augmentation design.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	Methodology	Techniques	Datasets	Evaluation	Key Findings	Limitations
XAI Review (Healthcare)	XAI	Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare [11] (2023)	Narrative review on XAI concepts and challenges in healthcare.	Discusses black/glass box, transparency/post-hoc, human-AI collaboration, sXAI, Granular Computing (GrC), Fuzzy Modeling (FM).	General healthcare applications.	Discusses challenges: legal compliance (GDPR, HIPAA, PIPL), privacy/security trade-offs, trust calibration, accuracy vs. explainability, measurement difficulties, future complexity.	XAI is crucial for trust in healthcare AI but faces significant challenges. Balance with accuracy is key. Explanations should be tailored. sXAI and Explainability-by-Design are promising.	Narrative review, lacks technical depth on methods and quantitative comparisons.
NR-IQA (Transformer)	IQA	No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency [12] (2022)	Hybrid CNN-Transformer NR-IQA model incorporating relative ranking and self-consistency losses.	Extracts multi-scale features (ResNet50). Uses Transformer encoder for non-local modeling. Relative ranking loss (triplet loss with adaptive margin). Self-consistency loss (horizontal flipping).	Tested on LIVE, CSIQ, TID2013, KADID-10K (synthetic); CLIVE, KonIQ-10k, LIVE-FB (authentic).	SRCC, PLCC (intra-/cross-dataset). Ablations on components and backbone size. Qualitative analysis (nearest neighbors, quality maps).	Combining CNN (local) and Transformer (non-local) is effective. Relative ranking and self-consistency improve performance and robustness. Achieves SOTA/competitive results.	Still shows some sensitivity to transformations despite consistency loss. Relies on patch sampling.
XAI Review (Medical)	XAI	Explainable artificial intelligence (XAI) in deep learning-based medical image analysis [19] (2022)	Systematic survey of XAI techniques in deep learning-based medical image analysis. Proposes taxonomy framework.	Reviews visual (backprop, perturbation, MIL), textual (captioning, TCAV), and example-based (triplets, influence functions, prototypes) XAI methods applied medically.	Surveys papers using various medical imaging datasets (MRI, CT, X-Ray etc.) across anatomical locations.	Discusses XAI evaluation criteria and robustness tests. Summarizes pros/cons of techniques.	Visual explanation (saliency) is dominant in medical XAI. Most methods are post-hoc, model-specific, local. Evaluation is lacking. Highlights critiques and future directions.	Review scope limited to papers up to Oct 2020.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	**Methodology**	**Techniques**	**Datasets**	**Evaluation**	**Key Findings**	**Limitations**
NR-IQA (Intelligibility)	IQA	IE-IQA: Intelligibility Enriched Generalizable No-Reference Image Quality Assessment [25] (2021)	Integrates image intelligibility (semantic understanding) with distortion features using a bilateral network for generalizable NR-IQA.	Uses two backbones: distortion (trainable EfficientNet-B0) and intelligibility (frozen, pre-trained on semantic tasks). Proposes contribution/sensitivity-based feature selection for intelligibility. Fuses features via Intelligibility Enhanced Module.	Tested on KonIQ-10k, SPAQ, LIVEW, CID2013, BID. Uses models pre-trained on ImageNet, Places-365, MS-COCO etc. for intelligibility.	SRCC, PLCC (intra-/cross-dataset). Ablations on tasks, feature selection, training from scratch. Grad-CAM visualization.	Intelligibility features significantly improve NR-IQA generalization. Feature selection is beneficial. Bilateral fusion works well. Different semantic tasks provide useful intelligibility priors.	Relies on availability of large pre-trained semantic models. Optimal semantic task/features might be application-dependent.
XAI Conceptual	XAI	What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research [15] (2021)	Proposes conceptual model linking XAI approaches to stakeholder desiderata satisfaction via understanding, moderated by context. Reviews stakeholders and desiderata.	Conceptual analysis. Discusses XAI approach families (ante-hoc/post-hoc, etc.), explanatory info types, understanding (mental models), desiderata facets (epistemic/substantial), context.	Not applicable (conceptual). Uses hypothetical scenarios.	Discusses need for empirical evaluation and interdisciplinary research.	XAI should target specific stakeholder desiderata. Understanding mediates the process. Context matters. Provides structured framework for evaluating/developing XAI. Highlights interdisciplinary needs.	Conceptual model needs empirical validation. Does not offer specific technical methods.
NR-IQA (Transfer Learning)	IQA	TTL-IQA: Transitive Transfer Learning based No-reference Image Quality Assessment [7] (2021)	Transitive Transfer Learning for NR-IQA via an auxiliary domain/task constructed by a Distortion Translation GAN (DT-GAN).	DT-GAN generates hallucinated distorted images with quality level labels. Semantic Features Transfer network (SFTnet) with SDA attention optimizes feature transfer (ImageNet -> Auxiliary -> IQA).	Pre-training uses Waterloo DB + target IQA DBs (LIVE, TID2013, CSIQ, LIVE MD, LIVEC). Evaluated on the same target DBs.	SRCC, PLCC (intra-/cross-dataset). Ablations on components, quality levels. Comparison with artificial simulation. Feature map visualizations.	Transitive transfer improves NR-IQA generalization. DT-GAN effectively simulates diverse distortions. SFTnet adaptively transfers useful features.	Requires complex DT-GAN training per target quality distribution. Relies on VGG backbone.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	**Methodology**	**Techniques**	**Datasets**	**Evaluation**	**Key Findings**	**Limitations**
XAI (Medical IQA)	XAI + IQA	Explainable Image Quality Analysis of Chest X-Rays [18] (2021)	Uses NormGrad saliency method for explainable quality assessment of Chest X-Rays (foreign object detection).	Compares NormGrad with Grad-CAM on Object-CXR dataset using Pointing Game accuracy. Qualitative analysis of saliency maps.	Object-CXR dataset.	Pointing Game accuracy, qualitative visualization.	NormGrad provides more accurate saliency maps than Grad-CAM for localizing quality issues (foreign objects) on Chest X-Rays.	Focuses on one specific dataset/task and compares only two main methods. Pointing Game has limitations.
XAI Saliency	XAI	There and Back Again: Revisiting Backpropagation Saliency Methods [13] (2020)	Analyzes backpropagation-based saliency methods. Proposes unifying framework, NormGrad, meta-saliency. Investigates layer combination & class sensitivity.	Extract & Aggregate framework. NormGrad (L2/Frobenius norm aggregation). Meta-saliency (inner SGD step). Layer combination strategies. Class-sensitivity metric.	PASCAL VOC 2007, ImageNet 2012. Uses VGG16, ResNet50.	Pointing Game accuracy, map correlations, weight sensitivity visualization.	Backprop methods fit framework. NormGrad viable. Layer combination improves localization. Class sensitivity increases with depth (explains Grad-CAM failure). Meta-saliency improves class sensitivity.	Focuses on non-invasive backprop methods. Pointing Game is main evaluation.
XAI Review	XAI	Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [2] (2020)	Comprehensive review defining XAI concepts, taxonomies (transparent vs post-hoc; DL specific), challenges, and Responsible AI.	Reviews wide range of XAI techniques (transparent models, post-hoc methods like text, visual, local, example, simplification, feature relevance).	Not applicable (review). Discusses various application domains.	Discusses interpretability vs performance trade-off, metrics, DL challenges, security, fairness, accountability, data fusion implications, guidelines.	Defines explainability focusing on audience. Provides extensive taxonomies. Highlights need for consensus, evaluation metrics, human factors. Introduces Responsible AI.	Review nature; breadth over technical depth.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	Methodology	Techniques	Datasets	Evaluation	Key Findings	Limitations
NR-IQA (Weak Supervision)	IQA	DeepFL-IQA: Weak Supervision for Deep IQA Feature Learning [10] (2020)	Two-stage NR-IQA via weakly supervised feature learning (predicting FR-IQA scores) and subsequent supervised fine-tuning/regression. Introduces KADID-10k dataset.	Stage 1: MTL (InceptionResNetV2) predicts 11 FR-IQA scores on KADIS-700k (unrated); uses HE-norm for stability. Stage 2: Extract MLSP features, train shallow regressor on rated datasets.	KADIS-700k (weak sup.), KADID-10k (prop., sup.). Tested on KADID-10k, LIVE, CSIQ, TID2013, LIVE-itW, KonIQ-10k.	MTL validation curves. SROCC/PLCC comparisons (intra-/cross-dataset). Ablations on learning schemes, regressors. Per-distortion analysis.	Weak supervision using FR scores is effective. HE-norm helps MTL. MLSP features work well. Outperforms conventional NR-IQA, competitive with deep NR-IQA on synthetic data.	Generalization to authentic distortions limited. Relies on FR metrics for weak labels. Requires large unrated dataset.
IQA DB	IQA	KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment [28] (2020)	Creation of large-scale, ecologically valid IQA database (KonIQ-10k) with authentic distortions and crowdsourced scores. Proposed deep BIQA model (KonCept512).	Database: Sampling from YFCC100m using quality/content indicators, diversity sampling, duplicate removal, manual filtering. Subjective Study: Crowdsourcing (ACR), expert-based filtering. BIQA Model: Transfer learning (CNNs), compared resolutions, losses, architectures. Feature training tested.	KonIQ-10k (proposed). Tested on KonIQ-10k, LIVE-itW.	DB: Diversity analysis. Scores: Reliability (ICC, inter-group SROCC). Model: SRCC/PLCC, training size effect, prediction power analysis (Nmax judges).	KonIQ-10k is diverse and reliably annotated. KonCept512 achieves SOTA on KonIQ-10k, generalizes well. Larger resolution helps. Performance comparable to 9 human judges.	Database focus on technical quality. Crowdsourcing limitations. Model training limited by resources for full resolution.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	Methodology	Techniques	Datasets	Evaluation	Key Findings	Limitations
NR-IQA (Meta-Learning)	IQA	MetaIQA: Deep Meta-learning for No-Reference Image Quality Assessment [5] (2020)	Deep meta-learning (optimization-based) to learn a shared quality prior from distortion-specific tasks for better NR-IQA generalization.	Defines distortion-specific tasks. Uses ResNet18 backbone. Bi-level gradient optimization (MAML-like with Adam). Fine-tunes meta-model on target task.	Meta-training: TID2013, KADID-10K (synthetic). Tested: Leave-one-out on synthetic DBs; generalization to authentic DBs (CID2013, LIVE challenge, KonIQ-10K).	SRCC, PLCC (leave-one-out, cross-dataset). Ablation vs direct pre-training. Parameter sensitivity. Gradient map viz.	Meta-learning improves generalization to unknown synthetic and authentic distortions. Outperforms SOTA NR-IQA in generalization. Prior model captures distortion locations.	Needs distortion-specific tasks for meta-training. Performance depends on meta-training diversity.
XAI Review	XAI	A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI [17] (2020)	Survey of XAI methods with categorization (Perceptive vs Mathematical Structure) and application to medical field.	Categories: Perceptive (Saliency, Signal, Verbal), Mathematical (Predefined Model, Feature Extraction, Sensitivity, Optimization). Discusses risks in medical XAI.	General ML and specific medical applications (stroke, Alzheimer's, EEG, skin lesions, etc.).	Categorizes methods, provides examples, discusses challenges/prospects for each, highlights medical risks (over-reliance, manipulation, noisy data).	Provides a structured overview of diverse XAI approaches. Highlights lack of human evaluation and potential risks in medical applications. Emphasizes need for domain-specific considerations.	Categorization might be subjective. Doesn't provide quantitative comparison of methods.
XAI Saliency	XAI	NormGrad: Finding the Pixels that Matter for Training [14] (2019)	Proposes NormGrad saliency method derived from 1x1 conv layer gradients. Introduces order 1 NormGrad (meta-learning).	NormGrad (order 0): Frobenius norm of gradient component outer product. 1x1 identity trick. NormGrad (order 1): Inner SGD step in loss.	ImageNet. Uses VGG16, ResNet50.	Qualitative visualization vs Grad-CAM at different depths. Comparison of map resolutions.	NormGrad (order 0) provides masks at any layer but isn't class-selective. Order 1 adds selectivity. Can produce higher-res maps than Grad-CAM.	Primarily qualitative. Order 1 is computationally more expensive.

Continued on next page

Table 2.1 – Continued from previous page

		Title (Year)	**Methodology**	**Techniques**	**Datasets**	**Evaluation**	**Key Findings**	**Limitations**
IQA Eval	IQA	A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms [24] (2019)	Large-scale performance evaluation of FR, fused FR, and NR IQA algorithms.	Evaluates 43 FR, 7 fused FR (22 versions), 14 NR methods using SRCC, PLCC, statistical significance testing, complexity analysis.	9 datasets (LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ, LIVE MD, MDID2013, MDID, MDIVL).	Detailed comparison tables (per dataset, overall weighted avg.), statistical significance matrices, execution times.	Identifies top FR (IWSSIM, FSIMc, VSI) and NR (CORNIA, HOSA, dipIQ) methods. Rank aggregation fusion (RAS) outperforms other fusion types and individual FR methods. NR lags FR significantly.	Evaluation limited to included datasets/methods. Doesn't cover authentic NR datasets fully.
XAI Review	XAI	Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) [29] (2018)	Literature survey defining XAI concepts, motivations, taxonomy (complexity, scope, model dependency), techniques, challenges.	Reviews intrinsic models, post-hoc (visualization, knowledge extraction, influence, example-based). Covers LIME, SHAP, Grad-CAM, LRP etc. briefly.	Mentions various application domains.	Discusses evaluation challenges, human factors, explain vs predict debate.	Provides broad overview and taxonomy. Highlights need for formalism, human factors, evaluation metrics. Notes technical challenges and interdisciplinary nature.	Breadth over depth. Taxonomy can be overlapping. Rapidly evolving field.
NR-IQA (GAN)	IQA	Hallucinated-IQA: No-Reference Image Quality Assessment via Adversarial Learning [11] (2018)	Generates a 'hallucinated' reference image using a GAN to compensate for missing reference in NR-IQA. Uses discrepancy map for regression.	Quality-Aware Generator (stacked hourglass) with pixel + quality-aware perceptual loss (VGG + NR-IQA features). IQA-Discriminator penalizes bad hallucinations. Regression network uses distorted image + discrepancy map, guided by high-level features from generator.	LIVE, CSIQ, TID2008, TID2013.	SRCC, LCC comparisons. Ablations on components. Qualitative visualizations.	Hallucinated reference improves NR-IQA. Quality-aware loss and IQA-discriminator enhance results. High-level fusion stabilizes regression. SOTA performance demonstrated.	Needs reference images for generator training. Performance relies on hallucination quality. Complex training.

Table 2.2: Summary of Key Empirical Metrics Reported

Paper (Year)	**Key Empirical Metric(s)**	**Reported Value(s) / Finding**
Reference-Free Image Quality Metric for Degradation and Reconstruction Artifacts [8] (2024)	Correlation (Predicted QF vs. Artifact Level); Avg. Dataset QF	Showed decreasing QF with increasing Gaussian Blur/Noise. Estimated Avg. QF for LIVE (0.780), Flickr1024 (0.952), ImageNet (0.971). Suboptimal results as perceptual loss.
A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment [20] (2024)	SRCC	GPT-4V achieved reasonable SRCC (e.g., >0.7 on FR-KADID/Aug-KADID/TQD, >0.8 on SPAQ) but struggled on color (SPCD SRCC 0.1). Open-source MLLMs showed much lower SRCC values across the board. Chain-of-Thought improved GPT-4V SRCC (e.g., FR-KADID 0.809 vs 0.745 std).
SF-IQA: Quality and Similarity Integration for AI Generated Image Quality Assessment [21] (2024)	SRCC, PLCC, Main Score (Avg.), NTIRE Rank	Achieved SOTA on AGIQA-3K (e.g., Good Models SRCC/PLCC 0.8239/0.8634 for quality branch). Placed 4th in NTIRE 2024 AIGIQA challenge (Main Score: 0.9138). Ablation showed score fusion improved over individual branches.
TOPIQ: A Top-down Approach from Semantics to Distortions for Image Quality Assessment [27] (2023)	SRCC, PLCC, BAPPS 2AFC Score, FLOPS	SOTA/competitive SRCC/PLCC on FR (e.g., PIPAL 0.813 SRCC) and NR (e.g., KonIQ-10k 0.926 SRCC) datasets. Best BAPPS 2AFC scores (e.g., 0.824 Synth, 0.714 Alg.). Significantly lower FLOPS (13% of AHIQ) for comparable performance.
Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective [22] (2023)	SRCC, PLCC, gMAD Win Rate, MOS Realignment SRCC	Outperformed SOTA on several datasets (Weighted Avg SRCC 0.922). Won gMAD competition against UNIQUE. Better MOS realignment (SRCC 0.879 vs. 0.851). Task ablation confirmed benefit of multitask learning.
Explainable Image Quality Assessment for Medical Imaging [23] (2023)	Pointing Game Accuracy, Difference of Means (DoM)	NormGrad achieved best Pointing Game: 0.853 (Object-CXR), 0.611 (LVOT). NormGrad (Combined) showed lower DoM (better reproducibility across architectures) than baselines like Grad-CAM.
Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild [9] (2023)	SRCC, PLCC	Achieved SOTA/competitive performance on 8 datasets (e.g., KonIQ SRCC 0.914, SPAQ SRCC 0.918, KADID SRCC 0.872). Unsupervised content features outperformed supervised ImageNet features.
No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency [12] (2022)	SRCC, PLCC	Achieved SOTA/competitive results on 7 datasets (Weighted Avg SRCC/PLCC: 0.685/0.732). Ablation confirmed benefits of Transformer, Relative Ranking, and Self-Consistency.
Explainable artificial intelligence (XAI) in deep learning-based medical image analysis [19] (2022)	N/A (Survey)	Summarized usage trends (visual dominant, post-hoc, local) and discussed evaluation criteria (application/human/functionally-grounded), robustness tests. No new empirical metrics from this paper.
IE-IQA: Intelligibility Enriched Generalizable No-Reference Image Quality Assessment [25] (2021)	SRCC, PLCC	Outperformed SOTA in cross-dataset tests (e.g., trained on KonIQ, tested on SPAQ: SRCC 0.859, LIVEW: 0.829). SOTA intra-dataset performance (e.g., KonIQ SRCC 0.900). Ablation confirmed benefit of intelligibility features.

Continued on next page

Table 2.2 – Continued from previous page

Paper (Year)	**Key Empirical Metric(s)**	**Reported Value(s) / Finding**
TTL-IQA: Transitive Transfer Learning based No-reference Image Quality Assessment [7] (2021)	SRCC, PLCC	Outperformed SOTA NR methods on LIVE (SRCC 0.979), LIVE MD (SRCC 0.952), LIVEC (SRCC 0.884). Better cross-database generalization than baselines. Ablation confirmed benefits of DT-GAN and SFTnet.
Explainable Image Quality Analysis of Chest X-Rays [18] (2021)	Pointing Game Accuracy	NormGrad achieved 0.862 Pointing Game accuracy on Object-CXR, outperforming Grad-CAM.
There and Back Again: Revisiting Backpropagation Saliency Methods [13] (2020)	Pointing Game Accuracy, Max-Min Class Correlation	Layer combination improved Pointing Game (e.g., Linear Approx. ResNet50 +1.84% on difficult VOC). Meta-saliency reduced max-min correlation (improved class sensitivity) across methods and layers (e.g., Linear Approx. conv5_3 correlation change -0.6).
DeepFL-IQA: Weak Supervision for Deep IQA Feature Learning [10] (2020)	SRCC, PLCC	Achieved high correlation on KADID-10k (SRCC 0.936). Competitive with deep NR-IQA on synthetic datasets (e.g., LIVE SRCC 0.972, TID2013 SRCC 0.858). Generalization to authentic datasets was lower (e.g., KonIQ-10k SRCC 0.877).
KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment [28] (2020)	Subject Score Reliability (ICC, Inter-group SRCC); BIQA Model Performance (SRCC, PLCC)	High score reliability (ICC 0.46, Inter-group SROCC 0.973). KonCept512 model: SRCC 0.921 (KonIQ-10k test), 0.825 (cross-test on LIVE-itW). Model performance comparable to Nmax $\approx$ 9 human judges.
MetaIQA: Deep Meta-learning for No-Reference Image Quality Assessment [5] (2020)	SRCC, PLCC	Outperformed SOTA in leave-one-distortion-out tests (Avg SRCC: 0.854 on TID2013, 0.767 on KADID-10K). Good generalization to authentic datasets (e.g., LIVE challenge SRCC 0.802, KonIQ-10K SRCC 0.850).
A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms [24] (2019)	Weighted Avg. SRCC/PLCC; Statistical Significance (F-test); Execution Time	Top FR: IWSSIM (0.875 SRCC), VSI (0.877 SRCC). Top NR: CORNIA (0.690 SRCC). RAS fusion outperformed best FR (e.g., RAS6 SRCC 0.893). NR performance significantly lags FR. Statistical tests showed significant differences between methods. Execution times varied widely.
Hallucinated-IQA: No-Reference Image Quality Assessment via Adversarial Learning [11] (2018)	SROCC, LCC (PLCC)	Outperformed SOTA NR methods on LIVE (SRCC 0.982), CSIQ (SRCC 0.949), TID2008 (SRCC 0.910), TID2013 (SRCC 0.879). Ablation showed benefits of quality loss, IQA-GAN, and fusion.

Chapter 3

Methodology

Addressing the problem statement from Chapter 1 (Section 1.7) and building on the review in Chapter 2, this chapter outlines the methodology for integrating gradient-based XAI into semantic segmentation for IQA-related tasks. The aim is to enhance model transparency by visually explaining predictions, focusing on identifying quality-relevant features.

An experimental pipeline (Figure 3.1) was developed, encompassing:

1. **Data Manipulation (Section 3.1):** Dataset preparation, preprocessing, and augmentation.

2. **Model Training & Selection (Section 3.2):** Training and validating a DeepLabV3+/ResNet50 segmentation model using K-Fold Cross-Validation.

3. **Explainability Methods (Section 3.3):** Applying Grad-CAM (Section 3.3.2) and NormGrad (Section 3.3.3) to generate saliency maps.

4. **Visualization (Section 3.4):** Synthesizing segmentation outputs and saliency maps for qualitative analysis.

Subsequent sections detail each stage of this framework.

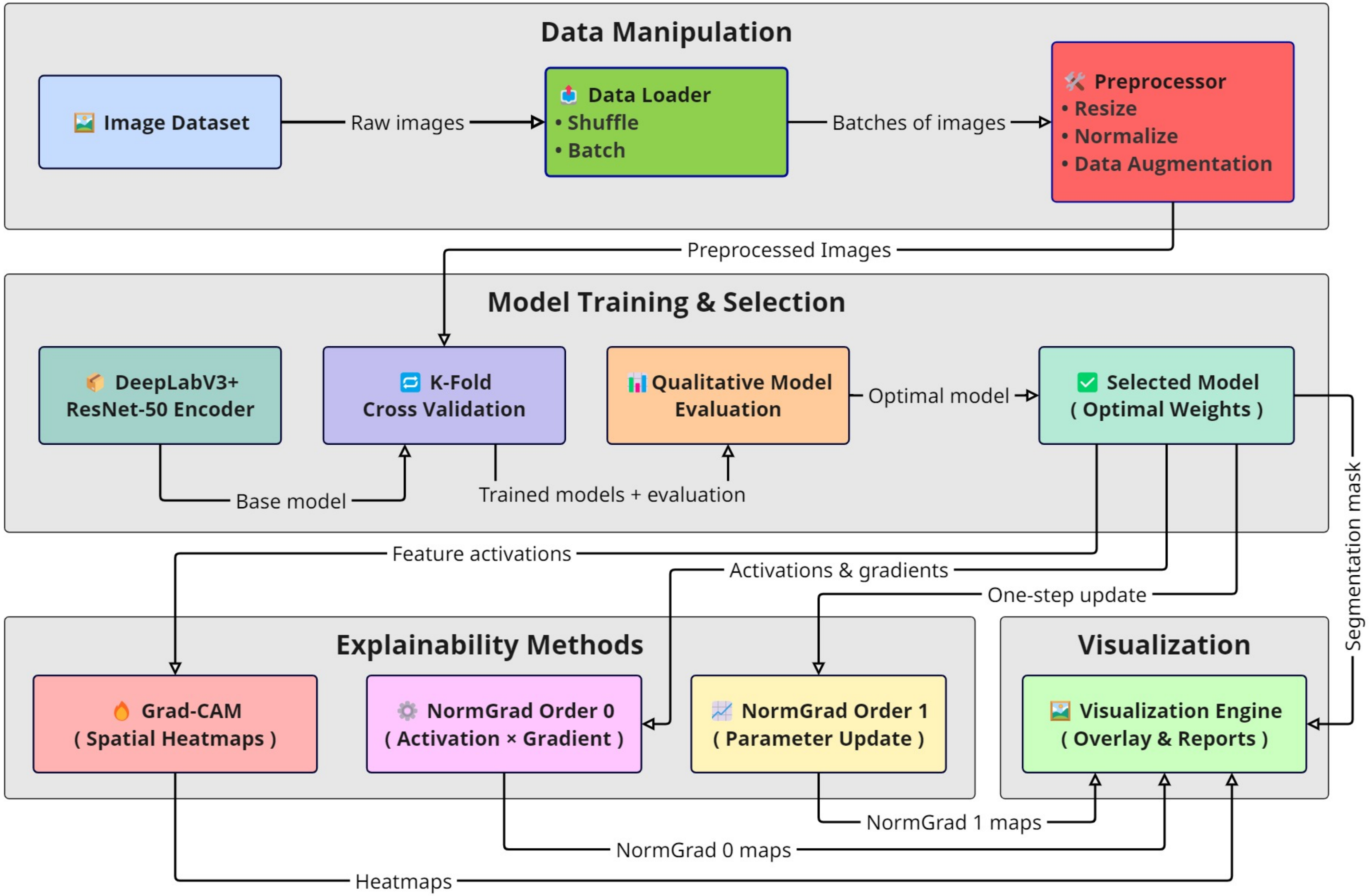


Figure 3.1: Pipeline overview: Data preparation, DeepLabV3+ (Φ_θ) training via K-Fold, explainability analysis (Grad-CAM, NormGrad), and visualization.

3.1 Data Manipulation

The Data Manipulation stage prepares raw image data and annotations for the deep learning model as illustrated in the overall pipeline (e.g., Figure 3.1).

This stage begins with the **Image Dataset** component, which accesses Chest X-ray images and associated metadata (e.g., CSV files). The metadata contains image identifiers and string-encoded annotations (rectangles, ellipses, polygons) for foreign objects. A key function is parsing these strings into binary segmentation masks, providing pixel-wise ground truth for each image.

Next, the **Data Loader** component efficiently manages data retrieval. It shuffles the training dataset each epoch to prevent order-based learning bias and improve generalization. It also groups image-mask pairs into batches, enabling parallel processing and stabilizing gradient estimates during training.

Finally, the **Preprocessor** receives data batches and applies transformations. All images and masks are resized to a uniform input dimension (e.g., 256×256). For training data, augmentations like random horizontal flipping and brightness/contrast adjustments are applied to enhance robustness. Test data preprocessing is typically limited to resizing for consistent evaluation.

The output of this stage comprises batches of preprocessed image and masks, formatted for input into the Model Training & Selection stage.

3.2 Model Training & Selection

Following data preparation, the Model Training & Selection stage, depicted in the overall pipeline (Figure 3.1), is responsible for developing and validating the core segmentation model for subsequent explainability methods.

3.2.1 DeepLabV3+ with ResNet50 Backbone

The core task addressed in this thesis involves not only assessing image quality but also understanding the spatial underpinnings of these assessments, often by identifying specific regions or objects within an image that influence perceived quality. To facilitate this spatial analysis, a robust semantic segmentation model is employed. The chosen architecture is DeepLabV3+ [30], a state-of-the-art model known for its effectiveness in dense prediction tasks. DeepLabV3+ utilizes an encoder-decoder structure, leveraging powerful pre-trained networks for feature extraction while incorporating mechanisms to capture multi-scale context and refine segmentation boundaries.

The specific configuration adopted in this work utilizes a ResNet50 [31] network as the encoder backbone, pre-trained on the large-scale ImageNet dataset [32]. The ResNet architecture, characterized by its deep structure and residual connections, serves as a powerful feature extractor, capable of learning rich hierarchical representations from input images. The residual connections are particularly crucial as they facilitate the training of very deep networks by mitigating the vanishing gradient problem, allowing for the effective learning of complex image features [31]. Using pre-trained weights allows the model to leverage knowledge learned from a vast dataset, significantly improving performance and reducing training time on the target task.

The selection of DeepLabV3+ with a ResNet50 backbone is motivated by several factors. This architecture represents a mature and well-validated approach within the semantic segmentation domain, offering stability and readily available implementations. Its effectiveness is supported by recent studies, such as [33], which found this combination to yield superior results in terms of segmentation accuracy (Dice and Jaccard scores) and computational efficiency compared to other variants in a medical imaging context. The use of a ResNet50 backbone, pre-trained on ImageNet, provides significant transfer learning advantages, reducing the need for extensive training data and leveraging robust feature extraction capabilities developed on a large-scale dataset, which is particularly beneficial for domain-specific tasks where data may be limited. Furthermore, the convolutional nature of this architecture aligns well with established gradient-based explainability techniques, such as NormGrad (discussed in Section 3.3.3), facilitating the generation of high-resolution saliency maps for interpretability. This contrasts with some alternative architectures like transformers, which may require larger datasets and present different challenges for applying certain XAI methods. The choice is further supported by applications in related fields; for instance, [34] employed a DeepLabV3-ResNet model for an explainable visual inspection system, highlighting its practical efficacy. Therefore, this configuration offers a pragmatic balance of performance, efficiency, data requirements, and compatibility with the explainability methods central to this research.

The DeepLabV3+ architecture builds upon this encoder by incorporating two key components: the Atrous Spatial Pyramid Pooling (ASPP) module and a dedicated decoder module, as illustrated in Figure 3.2.

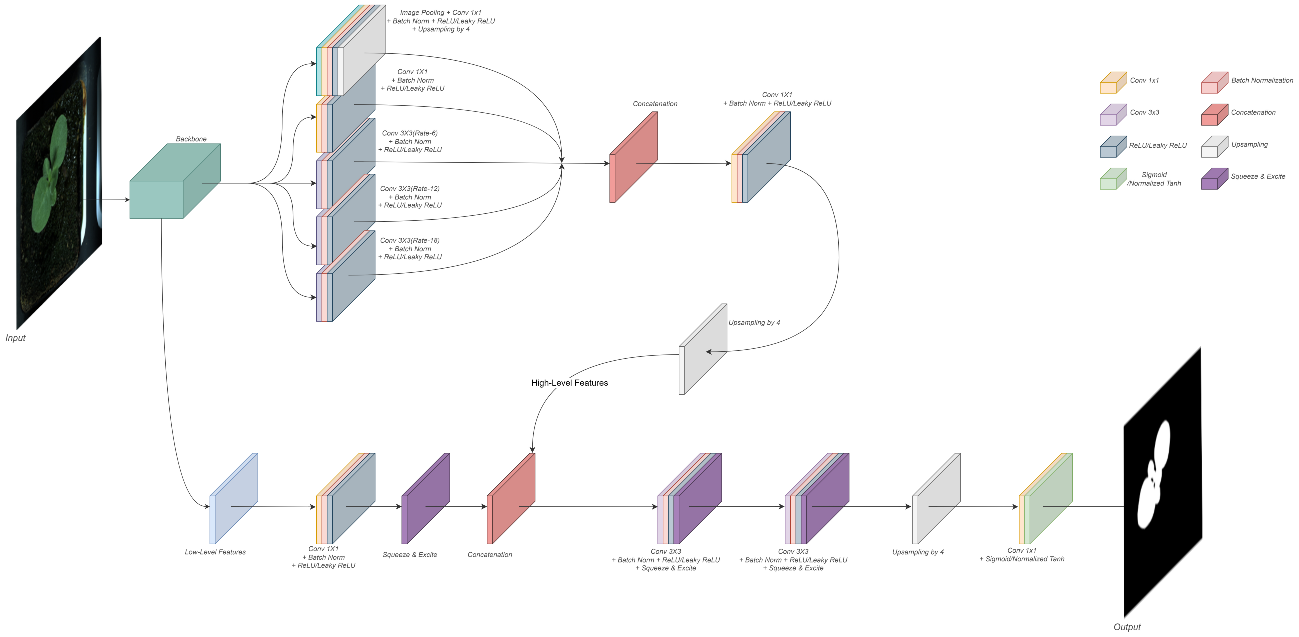


Figure 3.2: The architecture of the DeepLabV3+ model utilizing a ResNet50 backbone. Adapted from [30].

1. **Atrous Spatial Pyramid Pooling (ASPP):** The ASPP module employs atrous (dilated) convolutions with different dilation rates to capture multi-scale contextual information [30]. These convolutions expand the receptive field without increasing parameters. The module also incorporates global context through average pooling. By concatenating features from various scales, ASPP provides rich semantic representation for segmentation tasks.

2. **Decoder Module:** The decoder recovers spatial details lost during encoding by fusing high-level semantic features from ASPP with low-level features from earlier encoder stages [30]. After processing through 1x1 convolutions and bilinear upsampling, these features are concatenated and refined through additional convolutions to gradually restore the original image resolution. This architecture effectively combines semantic context with precise localization.

3. **Configuration for Binary Segmentation:** For binary segmentation tasks in this thesis, the final decoder layer uses a 1x1 convolution to reduce features to a

single channel, followed by a Sigmoid activation. This produces a probability map indicating the likelihood of each pixel belonging to the foreground class.

4. **Relevance to Explainable Image Quality Assessment:** DeepLabV3+ with ResNet50 is ideal for this research as it produces accurate, high-resolution segmentation maps crucial for localizing image quality features. Its multi-scale processing capabilities handle variations in quality-related features, while the encoder-decoder structure preserves both contextual information and fine details. This architecture provides an excellent foundation for applying XAI techniques (Section 3.3), enabling deeper understanding of features driving image quality assessment.

3.2.2 K-Fold Cross-Validation

To ensure the robustness and generalizability of the trained segmentation model, a K-Fold Cross-Validation strategy is employed, as depicted in the pipeline (Figure 3.1). This technique mitigates the risk of overfitting to a specific train-validation split and provides a more reliable estimate of the model's performance on unseen data.

The overall training dataset, consisting of preprocessed image-mask pairs, is partitioned into K mutually exclusive subsets, or "folds," of approximately equal size. The training process then iterates K times. In each iteration k (from 1 to K), the k-th fold is held out as the validation set, while the remaining K-1 folds are combined to form the training set for that iteration.

A fresh instance of the DeepLabV3+ model (or the model weights are reset) is trained on the combined K-1 training folds for a predefined number of epochs using the chosen optimizer and loss function (e.g., Dice Loss). After training within each fold iteration, the model's performance is evaluated on the held-out validation fold. This iterative process ensures that every sample in the original training dataset serves as validation data exactly once. The input to this component is the stream of preprocessed images from the Data Manipulation stage, and its intermediate outputs are the trained model instances and their corresponding validation performance metrics for each fold.

3.2.3 Qualitative Model Evaluation

While quantitative metrics provide objective scores, qualitative assessment through visual inspection is crucial for understanding the model's behavior in practice. A dedicated visualization procedure is implemented to facilitate this comparison, particularly when evaluating multiple models.

For a selected number of samples from the test set, this procedure generates a comparative visual report. For each sample, the report displays:

- The original input image.
- The ground-truth segmentation mask.
- For each model being compared:
  - The model's predicted binary segmentation mask.
  - An overlay of the predicted mask onto the original image for contextual understanding.
  - A heatmap generated using the an explainability method (detailed in Section 3.3), highlighting the input pixels deemed most influential for the segmentation prediction according to this method.
  - An overlay of the NormGrad heatmap onto the original image, allowing direct visual correlation between influential pixels and image structures.

This side-by-side visualization allows for direct comparison of segmentation quality (e.g., accuracy of boundaries, detection of small objects, false positives/negatives) and the corresponding saliency maps across different models for the same input image. It provides valuable insights into how different models achieve their results and where their attention mechanisms, as interpreted by NormGrad, differ.

Following the completion of both quantitative and qualitative evaluations, the **Model Selection** component identifies the optimal model configuration based on the previous quantitative and qualitative evaluations. The weights of this best-performing model are saved, representing the culmination of the training and validation process.

3.3 Explainability Methods

Having established the architecture of the segmentation model, the subsequent critical step involves employing methods to illuminate its internal decision-making process. The primary objective is to generate explanations for the segmentation predictions derived from the DeepLabV3+ model, thereby moving beyond a simple performance evaluation towards an interpretable understanding of *why* specific image regions are segmented. As discussed in Chapter 1 and conceptually illustrated in Figure 1.6, integrating explainability is fundamental for fostering trust, enabling model debugging, and deriving meaningful insights, particularly when applying complex models like those used in image quality assessment. This section details the specific post-hoc explainability techniques selected to achieve this goal, focusing on gradient-based methods capable of producing visual saliency maps that highlight the image features most influential to the segmentation output.

3.3.1 Gradient-based Saliency Methods

Gradient-based saliency methods constitute a significant category of post-hoc explainability techniques, particularly prevalent for interpreting the decisions of deep neural networks in computer vision tasks [2]. The fundamental principle underlying these methods is the utilization of gradient information, obtained through backpropagation, to infer the importance of input features (e.g., pixels) or intermediate neural activations with respect to a specific model output. The gradient of a chosen output score (such as the score for a predicted class or an aggregated segmentation metric) with respect to an input feature indicates the sensitivity of the output to changes in that feature. A higher gradient magnitude is typically interpreted as signifying greater importance or influence of that feature on the model's prediction for the given input [13].

These methods aim to generate visual explanations, commonly known as saliency maps or heatmaps, which overlay the input image and highlight regions deemed most influential for the model's decision. This visual feedback can provide valuable insights into the model's focus, helping to understand whether it attends to relevant structures or relies on spurious correlations. While basic gradient visualizations can sometimes be noisy or difficult to interpret directly, they form the foundation for numerous refinements and more sophisticated techniques designed to produce clearer and more robust explanations [13].

In the context of this thesis, which focuses on explaining the predictions of a segmentation model applied to image quality assessment tasks, gradient-based methods offer a powerful lens for understanding spatial reasoning. By analyzing how gradients flow back through the network, we can investigate which parts of an input image contribute most significantly to the segmentation output, potentially revealing how the model identifies quality defects or relevant objects. The subsequent sections will delve into two specific, advanced gradient-based techniques employed in this work: Gradient-weighted Class Activation Mapping (Grad-CAM) [35] and NormGrad [14], chosen for their distinct mechanisms and demonstrated utility in providing class-discriminative and fine-grained explanations, respectively.

3.3.2 GradCAM for Segmentation

Gradient-weighted Class Activation Mapping (Grad-CAM) [35] is a prominent gradient-based saliency technique designed to produce visual explanations for the predictions made by Convolutional Neural Network (CNN) models. Originally developed for image classification tasks, Grad-CAM identifies the regions within an input image that are most influential for a specific prediction (e.g., a particular class label) by analyzing the gradients flowing into the final convolutional layer. Its key advantage lies in its ability to generate class-discriminative localization maps without requiring architectural modifications or retraining of the model under investigation. Given the objective of this thesis to enhance the interpretability of models involved in image assessment, adapting Grad-CAM to the context of image segmentation provides a valuable tool for understanding *why* a model segments specific regions, which can be correlated with perceived quality defects or salient objects.

The core principle of Grad-CAM involves leveraging the spatial information preserved in the feature maps of convolutional layers. For a target class c and a chosen convolutional layer with K feature maps $A^k \in \mathbb{R}^{u \times v}$ (where $u \times v$ is the spatial dimension), Grad-CAM proceeds as follows:

1. **Gradient Computation:** The gradient of the score for the target class c, denoted y^c (typically the pre-softmax activation), is computed with respect to the feature map activations A^k of the chosen layer: $\frac{\partial y^c}{\partial A^k}$. This gradient signifies how changes in each feature map activation influence the score y^c.

2. **Neuron Importance Weighting:** The gradients are global average pooled across the spatial dimensions (i, j) to obtain the "neuron importance" weights α_k^c for each feature map k:
$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.1}$$
where $Z = u \times v$ is the number of pixels in the feature map. Each α_k^c represents the importance of feature map k for the target class c.

3. **Heatmap Generation:** A weighted combination of the forward activation maps A^k is computed, followed by a ReLU activation function:
$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{3.2}$$
The ReLU operation is crucial as it isolates features that have a positive influence on the class score y^c, ensuring that only regions contributing positively to the

prediction are highlighted. The resulting $L^c_{\text{Grad-CAM}}$ is a coarse heatmap of the same spatial resolution as the feature maps A^k, indicating the importance of each spatial location for the prediction c. This heatmap is typically upsampled to the original image resolution for visualization.Figure 3.3 illustrates typical Grad-CAM outputs, showcasing how the method generates class-discriminative heatmaps highlighting regions influential for a specific prediction, such as focusing on a dog's face when the target class is "dog" [13].

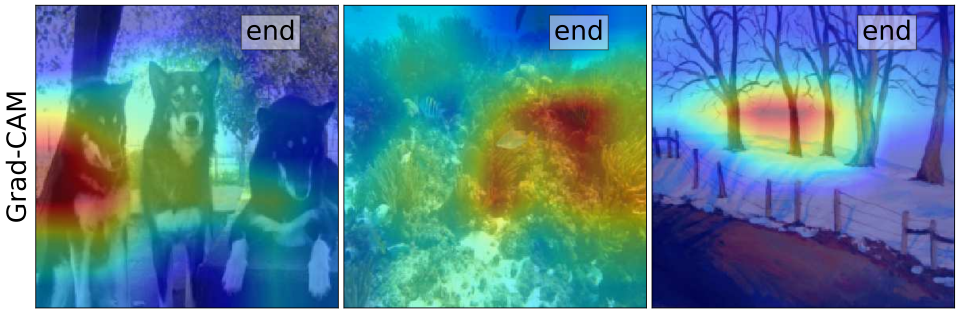


Figure 3.3: Example Grad-CAM saliency maps. [13].

3.3.2.1 Adapting Grad-CAM for Segmentation Tasks

Adapting Grad-CAM to explain the output of a segmentation model, such as the DeepLabV3+ architecture employed in this work (Section 3.2.1), requires defining an appropriate target score y^c. Unlike classification, segmentation models produce a pixel-wise output map predicting the class for each pixel. To apply Grad-CAM in this context, we need to aggregate this spatial output into a scalar value that represents the prediction of interest. A common approach, relevant for explaining the presence and localization of the segmented region, is to define y^c as the sum (or average) of the activations within the predicted segmentation mask for the class c. For instance, if we aim to explain why the model predicted the foreground object mask, y^c could be the sum of activations in the output layer corresponding to the foreground class for all pixels predicted as foreground. The gradients $\frac{\partial y^c}{\partial A^k}$ are then computed by backpropagating from this aggregated score.

The choice of the target convolutional layer is critical. Consistent with the original Grad-CAM paper [35] and common practice, the final convolutional layer of the encoder backbone (e.g., the last layer of the ResNet50 backbone before the ASPP module in DeepLabV3+) is typically selected. This layer represents the best compromise between capturing high-level semantic information and retaining sufficient spatial resolution. Identifying this layer programmatically often requires inspecting the model architecture, potentially using a helper function to locate the desired layer by name or type within the model's structure.

3.3.3 NormGrad

While Grad-CAM provides valuable localization maps by weighting feature map activations with gradient information related to the output score, alternative gradient-based saliency methods offer different perspectives on feature importance. NormGrad, introduced by Rebuffi et al. [14], presents such an alternative, originally designed to identify input pixels that are most influential during the model's *training* process, rather than just for inference. This focus on training dynamics can provide complementary insights, particularly when analyzing how specific image regions contribute to the learning of features relevant to quality assessment. As highlighted in Chapter 2 (Section 2.3), NormGrad has shown promise in generating fine-grained saliency maps capable of precisely localizing quality-degrading elements in medical imaging contexts [18, 23], motivating its inclusion in this methodology.

The core idea of NormGrad, as detailed in [14], stems from analyzing the gradient computation for the weights of a convolutional layer. Consider a specific convolutional layer $k_{\mathbf{w}}$ within the network Φ_θ. Let q represent the composition of layers preceding $k_{\mathbf{w}}$, and h represent the composition of layers following it, including the loss function ℓ. For an input $\mathbf{x}$, the activation entering the layer is $\mathbf{x}' = q(\mathbf{x})$, and the output is $\mathbf{x}'' = k_{\mathbf{w}}(\mathbf{x}') = \mathbf{w} * \mathbf{x}'$. The gradient of the loss with respect to the filter weights $\mathbf{w}$ can be expressed as a sum over spatial locations u in the output feature map:

$$\frac{d(h \circ k_{\mathbf{w}} \circ q)(\mathbf{x})}{d\mathbf{w}} = \sum_u \mathbf{g}_u {\mathbf{x}'}_u^\top \tag{3.3}$$

where $\mathbf{g}_u$ represents the gradient of the final loss backpropagated to the spatial location u of the layer's output ($\frac{dh}{d\mathbf{x}''_u}$), and $\mathbf{x}'_u$ represents the corresponding input activation patch at that location (or simply the activation vector at u for a 1×1 convolution). Each term $\mathbf{g}_u {\mathbf{x}'}_u^\top$ represents the contribution of spatial location u to the overall weight gradient update. NormGrad proposes using the norm of this contribution as a measure of importance for that spatial location:

$$\mathbf{m}_u = \|\mathbf{g}_u {\mathbf{x}'}_u^\top\|_F = \|\mathbf{g}_u\| \cdot \|{\mathbf{x}'}_u\| \tag{3.4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\cdot\|$ denotes the L2 norm for vectors. The resulting map $\mathbf{m}$, composed of values $\mathbf{m}_u$ for all spatial locations u, highlights regions where the interaction between the forward activations and the backpropagated gradients is strongest, indicating high influence on the weight gradients.

Within the scope of this thesis, two variants of NormGrad were implemented, offering different computational trade-offs and interpretations:

1. **Zero-Order NormGrad:** This variant directly implements the core NormGrad idea described above (Equation 3.4). It requires a single forward pass through the network to compute the activations $\mathbf{x}'$ at the target layer, followed by a single backward pass to compute the gradients $\mathbf{g}$ of the chosen output score (e.g., related to the segmentation mask) with respect to these activations. The importance map $\mathbf{m}$ is then calculated by taking the element-wise product of the L2 norms of the activation vectors and gradient vectors at each spatial location u. This method provides a computationally efficient way (comparable to standard gradient backpropagation) to generate a saliency map based on the magnitude of gradient contributions to the layer's weights, reflecting the influence of each spatial location on the model's learned parameters concerning the current input and task.

2. **First-Order NormGrad:** This variant adopts a meta-learning perspective, aiming to identify pixels that are most crucial for improving the model's performance on the given input via a hypothetical gradient descent step. It estimates how sensitive the loss is to changes in input pixels *after* a small update to the model weights θ, i.e., $\theta' \leftarrow \theta - \eta \nabla_\theta \ell(\Phi_\theta(\mathbf{x}), y)$. The computation involves approximating a Hessian-vector product using finite differences, which typically requires multiple forward and backward passes (often four, as noted in [14]). The parameter ϵ controls the step size used in this finite difference approximation. This method is computationally more intensive than the Zero-Order variant but aims to capture a different aspect of importance related to the dynamics of learning and optimization, potentially revealing regions critical for model adaptation or regions that might cause instability during training (if run in an adversarial mode).

These implemented NormGrad variants provide alternative saliency maps to those generated by Grad-CAM. The Zero-Order method offers an efficient way to visualize the spatial distribution of weight gradient magnitudes, while the First-Order method provides insights into the training dynamics and sensitivity. Both contribute to the goal of achieving a more comprehensive understanding of the segmentation model's behavior in the context of image quality assessment by highlighting influential input regions from different perspectives. Figure 3.4 provides visual examples comparing Zero-Order and First-Order NormGrad outputs. As shown in [13], Zero-Order NormGrad tends to produce class-agnostic maps highlighting all regions with strong gradient-activation interactions at a given layer, whereas First-Order NormGrad can yield more fine-grained, class-discriminative maps by incorporating a refinement step related to hypothetical weight updates.

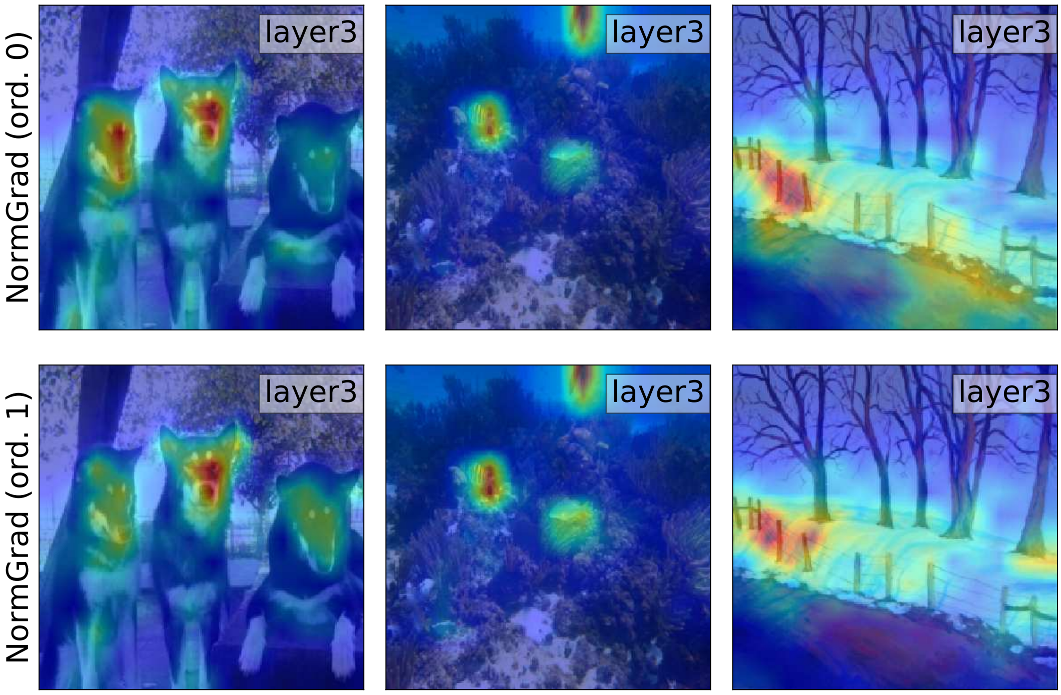


Figure 3.4: Example NormGrad saliency maps. [13].

3.4 Visualization and Scoring Engine

Building upon the explainability maps derived from methods like NormGrad, a dedicated Visualization and Scoring Engine was developed to facilitate both qualitative interpretation and quantitative assessment based on the segmentation model's outputs. As illustrated in the overall pipeline architecture (conceptually depicted in Figure 3.1), this engine synthesizes information from the trained model Φ_θ, generating insightful visual overlays and calculating object-based scores leading to overall image quality metrics. Its primary role is to translate raw model predictions into human-interpretable formats and actionable scores, aiding in the understanding of the model's spatial reasoning and the significance of detected features for image quality assessment.

The engine's workflow commences by processing the raw output map $O = \Phi_\theta(I)$ generated by the segmentation model (e.g., DeepLabV3+) for a given input image $I \in \mathbb{R}^{H\times W\times C}$. This output is converted into a binary mask $M \in \{0,1\}^{H\times W}$ by applying a sigmoid activation $\sigma(\cdot)$ followed by a threshold T (typically $T = 0.5$):

$$M(i,j) = \begin{cases} 1 & \text{if } \sigma(O(i,j)) > T \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

where (i,j) denotes the pixel coordinates. This mask M delineates the regions identified by the model as potential anomalies or foreign objects.

Following binarization, a connected components analysis (CCA) algorithm is applied to M. CCA identifies distinct, contiguous regions of foreground pixels, grouping them into a set of N distinct object instances $\mathcal{O} = \{o_1, o_2, ..., o_N\}$. The CCA process yields a labeled mask $L \in \{0, ..., N\}^{H\times W}$, where $L(i,j) = k$ if pixel (i,j) belongs to object o_k, and 0 for background. For each detected object o_k (where $k \in \{1, ..., N\}$), the following properties are extracted:

- **Pixel Area:** $A_k = |o_k| = \sum_{i,j}[L(i,j) = k]$, where $[\cdot]$ is the Iverson bracket, defined as:

  $$[P] = \begin{cases} 1 & \text{if proposition } P \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

  Thus, the area counts pixels belonging to object k.

- **Relative Size:** $S_k^{\text{rel}} = A_k/(H \times W)$. (Using S_k^{rel} to distinguish from S_k^{penalty} later)

- **Centroid:** $C_k = (\bar{i}_k, \bar{j}_k)$, calculated as:

  $$\bar{i}_k = \frac{1}{A_k} \sum_{(i,j):L(i,j)=k} i, \quad \bar{j}_k = \frac{1}{A_k} \sum_{(i,j):L(i,j)=k} j \tag{3.7}$$

- **Bounding Box:** The minimum enclosing rectangle $(x^k_{min}, y^k_{min}, w_k, h_k)$.

For informational purposes and visualization, the engine integrates spatial saliency information. This is typically achieved using NormGrad Order 1 heatmaps. While a heatmap H_{NG1} can be computed for a specific layer (as described in Section 3.3.3), the implemented system primarily generates a combined saliency map. This map is created by averaging the normalized NormGrad Order 1 heatmaps derived from a predefined list of multiple network layers. The resulting combined heatmap, initially potentially of different dimensions $H' \times W'$, is resized using bilinear interpolation to match the input image dimensions, yielding $H'_{NG1} \in \mathbb{R}^{H \times W}$, and subsequently normalized to the range $[0, 1]$. For each object o_k, the engine calculates the mean saliency value over its pixels, termed the 'Saliency Score' (or 'Confidence'):

$$C_k^{\text{saliency}} = \frac{1}{A_k} \sum_{(i,j):L(i,j)=k} H'_{NG1}(i, j) \tag{3.8}$$

This score C_k^{saliency} quantifies the model's focus on object o_k according to NormGrad. It is displayed for qualitative assessment and, while it does not directly factor into the primary image quality calculation (Equation 3.12), it is used to derive a secondary, 'filtered' image quality score by considering only objects exceeding a predefined saliency threshold $T_{saliency}$ (see Equation 3.13).

To contextualize the detected objects, a spatial 'Importance' score P_k^{imp} is calculated. A Region of Interest (ROI), $R_{ROI} = [y_{min}, y_{max}] \times [x_{min}, x_{max}]$, is defined (e.g., representing the expected chest area). The ROI center is C_{ROI}. For each object o_k with centroid C_k, the Euclidean distance $d_k = \|C_k - C_{ROI}\|_2$ is computed. Let d_{max} be the maximum distance from C_{ROI} to any image corner. The importance score $P_k^{\text{imp}} \in [0, 1]$ is:

$$P_k^{\text{imp}} = \begin{cases} 1.0 & \text{if } C_k \in R_{ROI} \\ 1.0 - \frac{d_k}{d_{max}} & \text{otherwise} \end{cases} \tag{3.9}$$

This prioritizes objects within or near the critical region R_{ROI}.

Furthermore, a 'Size Penalty' score P_k^{size} is introduced to penalize objects based on their relative size S_k^{rel}, using a sigmoid function parameterized by steepness k_{sig} and threshold τ_{sig}:

$$P_k^{\text{size}} = \frac{1}{1 + e^{-k_{sig}(S_k^{\text{rel}} - \tau_{sig})}} \tag{3.10}$$

This formulation enables flexible penalization of detected objects based on their size characteristics, allowing for tailored responses, such as assigning negligible penalties to very small detections (potential noise artifacts) while more significantly penalizing objects that occupy excessive portions of the image frame.

Figure 3.5 visually demonstrates how the size and location (importance) of detected objects influence their individual penalty. The figure contrasts scenarios leading to low versus high penalties. While saliency C_k^{saliency} is also computed, it serves as an independent informational metric about model attention and as a basis for filtering objects for a secondary quality score, rather than a direct component of the individual object penalty S_k^{penalty}.

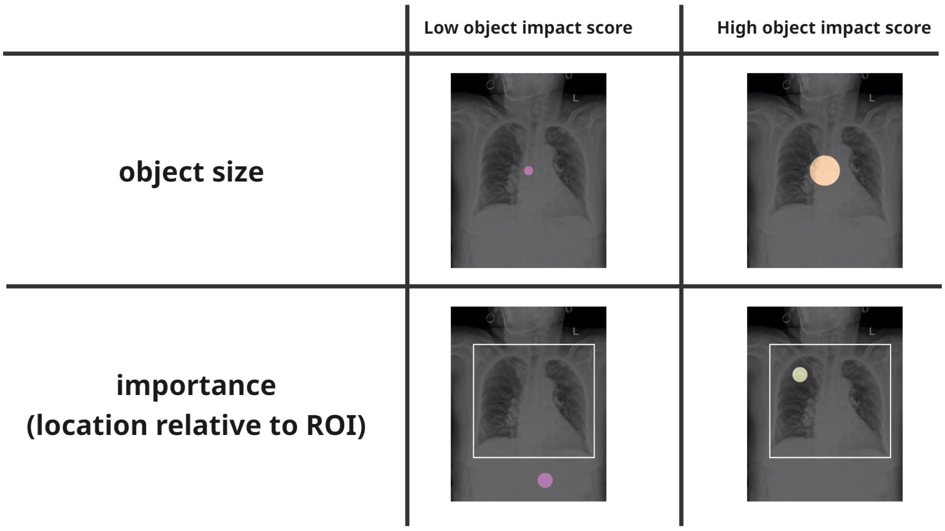


Figure 3.5: Illustration of how object size and importance (location relative to ROI) contribute to an object's penalty score.

Finally, the engine aggregates these metrics to compute overall image quality scores. First, an 'Individual Object Penalty' S_k^{penalty} is calculated for each object by combining its importance and size penalty using predefined weights w_{imp} and w_{size} (where $w_{imp}+w_{size} = 1$):

$$S_k^{\text{penalty}} = w_{imp}P_k^{\text{imp}} + w_{size}P_k^{\text{size}} \tag{3.11}$$

This S_k^{penalty} represents the degradation factor attributed to object o_k, bounded within $[0, 1]$. The primary 'Overall Image Quality' score Q_{img} is then derived using a multiplicative aggregation of "quality factors" for each object:

$$Q_{img} = \prod_{k=1}^{N}(1 - S_k^{\text{penalty}}) \tag{3.12}$$

If no objects are detected ($N = 0$), Q_{img} is defined as 1.0. A higher Q_{img} (closer to 1.0) indicates better image quality, implying fewer and/or less impactful (in terms of size and location) anomalies were detected. This product formulation ensures that the score remains within $[0, 1]$ and that each detected object contributes to a reduction in perceived quality.

In addition to this primary quality score, a secondary, 'filtered' image quality score, $Q_{img}^{\text{filtered}}$, is computed. This score aims to assess quality based on objects detected with higher confidence (saliency). It is calculated using the same product formulation as Equation 3.12, but the product is taken only over the subset of objects $\mathcal{O}' \subseteq \mathcal{O}$ for which the Saliency Score C_k^{saliency} (from Equation 3.8) exceeds a predefined threshold $T_{saliency}$:

$$Q_{img}^{\text{filtered}} = \prod_{o_k \in \mathcal{O}' : C_k^{\text{saliency}} > T_{saliency}} (1 - S_k^{\text{penalty}}) \tag{3.13}$$

If no objects meet the saliency threshold ($N' = |\mathcal{O}'| = 0$), $Q_{img}^{\text{filtered}}$ is also defined as 1.0. This dual-score approach allows for a nuanced understanding, distinguishing between the overall impact of all detected objects and the impact of those anomalies detected with higher model confidence.

Combining all the intermediate steps, the comprehensive equation for the primary final image quality score Q_{img} can be expressed as:

$$Q_{img} = \prod_{k=1}^{N} \left(1 - \left(w_{imp} P_k^{\text{imp}} + w_{size} \frac{1}{1 + e^{-k_{sig}(S_k^{\text{rel}} - \tau_{sig})}}\right)\right) \tag{3.14}$$

where:

- P_k^{imp} is the importance score defined in Equation 3.9.
- S_k^{rel} is the relative size of the object o_k.
- N is the total number of detected objects.

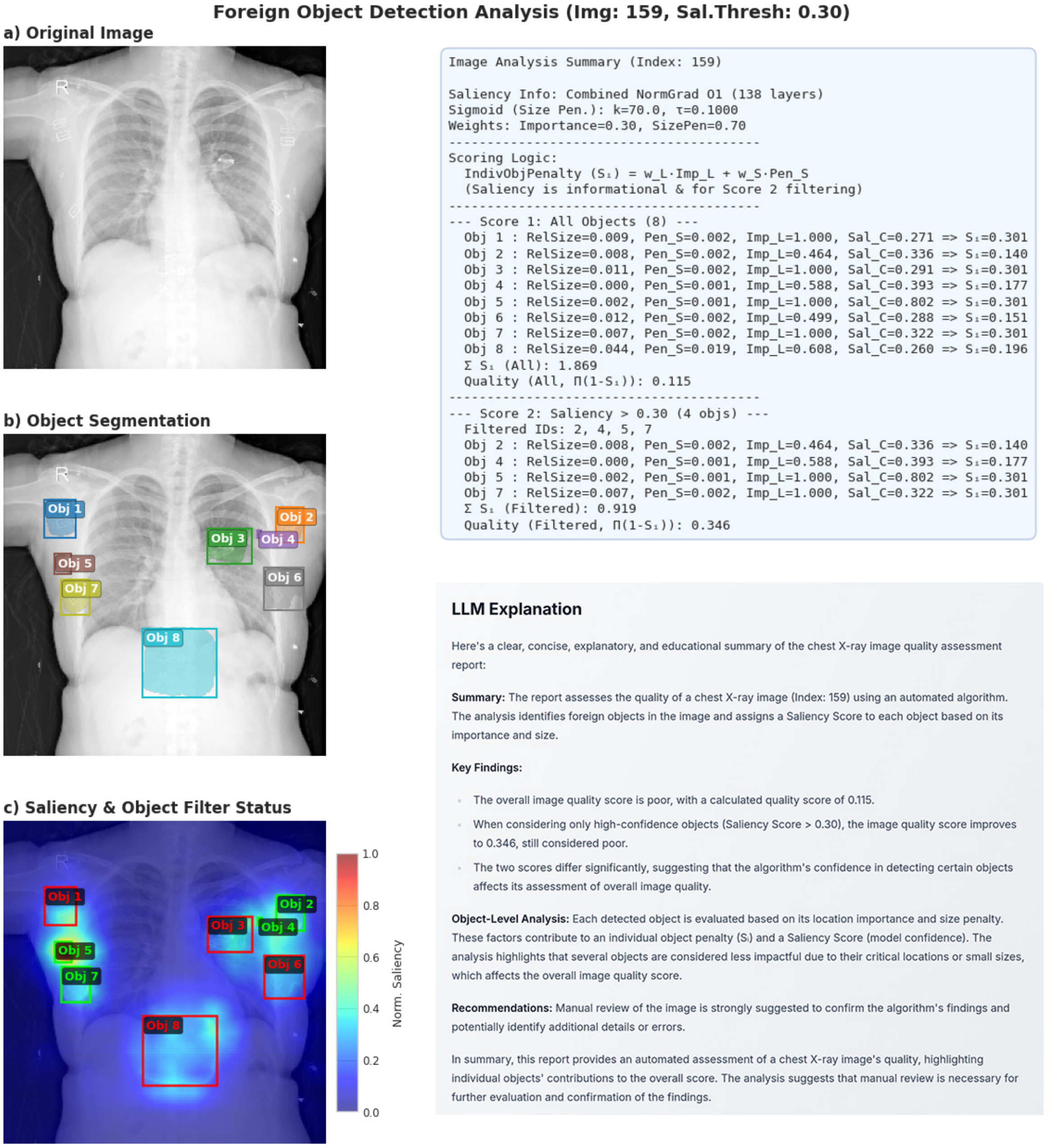
Finally, the engine generates composite visualizations and analysis outputs. These include:

- **Original image:** The unmodified input image I.
- **Object overlay:** The image with color-coded object masks (from L), bounding boxes, and identifiers for each o_k.
- **Saliency visualization (Informational):** The image overlaid with the resized and normalized NormGrad saliency heatmap H'_{NG1} using alpha blending, provided as supplementary information and to illustrate the basis for object filtering.
- **Analysis summary:** A text panel detailing:
  - Analysis parameters (details of the saliency map generation, such as the use of combined NormGrad Order 1 from specified layers; sigmoid parameters k_{sig}, τ_{sig} for size penalty; weights w_{imp}, w_{size} for individual object penalty; saliency filter threshold $T_{saliency}$).

- Per-object metrics for each object o_k: Relative size S_k^{rel}, size penalty P_k^{size}, importance P_k^{imp}, Saliency Score C_k^{saliency} (informational and for filtering), and individual object penalty S_k^{penalty}.
- Metrics for Score 1 (Q_{img}, based on all N objects):
  - Number of detected objects N.
  - The sum of individual object penalties $\sum_{k=1}^{N} S_k^{\text{penalty}}$ (for reference).
  - The primary image quality score Q_{img}.
- Metrics for Score 2 ($Q_{img}^{\text{filtered}}$, based on N' high-saliency objects):
  - Number of objects N' with $C_k^{\text{saliency}} > T_{saliency}$.
  - List of IDs for these N' filtered objects.
  - The sum of penalties for filtered objects $\sum_{o_k \in \mathcal{O}'} S_k^{\text{penalty}}$ (for reference).
  - The filtered image quality score $Q_{img}^{\text{filtered}}$.

Furthermore, to enhance the interpretability of these quantitative results, the system leverages a Large Language Model (LLM). Subsequent to the generation of visual outputs, the structured analysis data, including configuration parameters, per-object metrics, and the derived quality scores (Q_{img} and $Q_{img}^{\text{filtered}}$), are formatted into a detailed prompt. The LLM then generates a natural language report, providing a qualitative summary and explanation of the findings. This process involves structuring the numerical and categorical data from the analysis into a textual format designed to elicit a comprehensive and explanatory response from the language model. This report aims to make the complex analysis more accessible, offering insights into the detected objects' impact on overall image quality and highlighting any particularly noteworthy findings based on the automated assessment.

These visual outputs, calculated scores, and the LLM-generated natural language report collectively provide a comprehensive qualitative and quantitative assessment tool. This enables a deeper understanding of the model's behavior and offers derived metrics for image quality rooted in explainable AI principles.



```
Image Analysis Summary (Index: 159)

Saliency Info: Combined NormGrad O1 (138 layers)
Sigmoid (Size Pen.): k=70.0, τ=0.1000
Weights: Importance=0.30, SizePen=0.70
----------------------------------------
Scoring Logic:
  IndivObjPenalty (Sᵢ) = w_L·Imp_L + w_S·Pen_S
  (Saliency is informational & for Score 2 filtering)
----------------------------------------
--- Score 1: All Objects (8) ---
  Obj 1 : RelSize=0.009, Pen_S=0.002, Imp_L=1.000, Sal_C=0.271 => Sᵢ=0.301
  Obj 2 : RelSize=0.008, Pen_S=0.002, Imp_L=0.464, Sal_C=0.336 => Sᵢ=0.140
  Obj 3 : RelSize=0.011, Pen_S=0.002, Imp_L=1.000, Sal_C=0.291 => Sᵢ=0.301
  Obj 4 : RelSize=0.000, Pen_S=0.001, Imp_L=0.588, Sal_C=0.393 => Sᵢ=0.177
  Obj 5 : RelSize=0.002, Pen_S=0.001, Imp_L=1.000, Sal_C=0.802 => Sᵢ=0.301
  Obj 6 : RelSize=0.012, Pen_S=0.002, Imp_L=0.499, Sal_C=0.288 => Sᵢ=0.151
  Obj 7 : RelSize=0.007, Pen_S=0.002, Imp_L=1.000, Sal_C=0.322 => Sᵢ=0.301
  Obj 8 : RelSize=0.044, Pen_S=0.019, Imp_L=0.608, Sal_C=0.260 => Sᵢ=0.196
  Σ Sᵢ (All): 1.869
  Quality (All, Π(1-Sᵢ)): 0.115
----------------------------------------
--- Score 2: Saliency > 0.30 (4 objs) ---
  Filtered IDs: 2, 4, 5, 7
  Obj 2 : RelSize=0.008, Pen_S=0.002, Imp_L=0.464, Sal_C=0.336 => Sᵢ=0.140
  Obj 4 : RelSize=0.000, Pen_S=0.001, Imp_L=0.588, Sal_C=0.393 => Sᵢ=0.177
  Obj 5 : RelSize=0.002, Pen_S=0.001, Imp_L=1.000, Sal_C=0.802 => Sᵢ=0.301
  Obj 7 : RelSize=0.007, Pen_S=0.002, Imp_L=1.000, Sal_C=0.322 => Sᵢ=0.301
  Σ Sᵢ (Filtered): 0.919
  Quality (Filtered, Π(1-Sᵢ)): 0.346
```

LLM Explanation

Here's a clear, concise, explanatory, and educational summary of the chest X-ray image quality assessment report:

Summary: The report assesses the quality of a chest X-ray image (Index: 159) using an automated algorithm. The analysis identifies foreign objects in the image and assigns a Saliency Score to each object based on its importance and size.

Key Findings:

- The overall image quality score is poor, with a calculated quality score of 0.115.
- When considering only high-confidence objects (Saliency Score > 0.30), the image quality score improves to 0.346, still considered poor.
- The two scores differ significantly, suggesting that the algorithm's confidence in detecting certain objects affects its assessment of overall image quality.

Object-Level Analysis: Each detected object is evaluated based on its location importance and size penalty. These factors contribute to an individual object penalty (S_i) and a Saliency Score (model confidence). The analysis highlights that several objects are considered less impactful due to their critical locations or small sizes, which affects the overall image quality score.

Recommendations: Manual review of the image is strongly suggested to confirm the algorithm's findings and potentially identify additional details or errors.

In summary, this report provides an automated assessment of a chest X-ray image's quality, highlighting individual objects' contributions to the overall score. The analysis suggests that manual review is necessary for further evaluation and confirmation of the findings.

Figure 3.6: Example output of the Visualization and Scoring Engine, showcasing the original image, object overlay, informational saliency visualization, analysis summary including the primary and filtered overall image quality scores, and the LLM-generated natural language report.

Chapter 4

Experiments and Discussion

This chapter presents a comprehensive empirical investigation designed to rigorously evaluate the efficacy, characteristics, and practical utility of the Explainable Artificial Intelligence (XAI) framework for Image Quality Assessment (IQA) developed in this thesis. The primary focus is on the application of this framework to the detection of foreign objects in medical X-ray images, a critical task where both accuracy and interpretability are paramount. The code implementation underpinning the experiments described here is publicly available.[1]

The chapter systematically unfolds the experimental methodology and findings. It commences by delineating the **Experimental Setup** (Section 4.1), which details the Object-CXR dataset and its preprocessing, the training procedure for the DeepLabV3+ segmentation model employing Focal Tversky Loss, and the suite of quantitative metrics selected for assessing segmentation performance (e.g., Dice Coefficient, Boundary IoU) and the effectiveness of XAI techniques such as Grad-CAM and NormGrad (e.g., Pointing Game Accuracy, faithfulness metrics). Implementation specifics, including software libraries and hardware configurations, are also provided to ensure transparency and reproducibility.

Subsequently, the chapter transitions to an in-depth presentation and discussion of the **Experimental Results** (Section 4.2). This section encompasses a thorough quantitative analysis of the segmentation model's performance derived from K-Fold Cross-Validation, a detailed quantitative evaluation of the employed XAI methods, and a comparative benchmarking against established baseline classification models (ResNet34, EfficientNet-B0) to contextualize the proposed system's capabilities. Complementing

[1] https://github.com/MoncefDj/Explainable-Artificial-Intelligence-for-Image-Quality-Assessment

these numerical evaluations, a **Qualitative Analysis** (Section 4.2.2) explores the interpretability, relevance, and user-perceived quality of the generated saliency maps and Large Language Model (LLM)-based textual summaries, leveraging a custom-developed interactive Gradio interface for systematic user feedback.

Furthermore, a **Model Complexity Analysis** (Section 4.2.3) examines the computational demands of the system, including space (memory) and time (processing) complexities for both the core model and the XAI components. This analysis offers crucial insights into the practical deployment considerations and scalability of the proposed XAI-IQA solution.

Collectively, this chapter aims to provide robust empirical evidence supporting the capabilities of the XAI-IQA system, elucidating the nuances of its performance, the comparative advantages of its explainability features, and its potential for trustworthy application in medical imaging scenarios. The findings discussed herein form the cornerstone for validating the contributions of this research.

4.1 Experimental Setup

This section delineates the comprehensive experimental framework established to evaluate the proposed XAI-driven IQA system. It begins by detailing the dataset selected for identifying foreign objects in medical images and the specific preprocessing steps applied to prepare the data for model consumption. Subsequently, the training procedure for the core segmentation model is outlined, including the choice of loss functions, optimization strategies, and cross-validation techniques. The section further describes the array of metrics and methodologies employed for the quantitative evaluation of both the segmentation model's performance and the efficacy of the explainability techniques. Finally, crucial implementation details concerning the software, hardware, and key hyperparameters are provided to ensure reproducibility and contextualize the experimental outcomes.

4.1.1 Dataset and Preprocessing

For the empirical validation of our segmentation model and subsequent explainability analysis, we utilized the Object-CXR dataset (Foreign objects in chest X-rays by Darius Barušauskas, 2022, published in Kaggle[2]). This resource, previously employed in foundational studies on explainable quality assessment for Chest X-rays [18], originates

[2] https://www.kaggle.com/datasets/raddar/foreign-objects-in-chest-xrays

from JF Healthcare and is tailored for identifying foreign objects (e.g., clips, buttons, wires) within Chest X-ray images, a common source of diagnostic quality degradation. The dataset comprises 10,000 images, balanced between those with and without foreign objects. Images featuring foreign objects include bounding box annotations indicating their locations.

We adhered to the dataset's predefined structure, partitioning it into training and testing sets. The training set was used for model learning, while the testing set served as unseen data for evaluating generalization. Our data handling involved reading image files and converting bounding box annotations into binary segmentation masks matching image dimensions (1 for foreign objects, 0 for background), providing ground truth for segmentation model training.

To prepare data and enhance training diversity, several preprocessing and augmentation steps were applied. All images and masks were uniformly resized to 256×256 pixels. Training data underwent random augmentations, including horizontal flipping and brightness/contrast adjustments, applied with a set probability to improve robustness and mitigate overfitting. Test data were only resized to maintain consistent evaluation conditions. Finally, datasets were organized into batches of 16 samples for efficient processing during training and evaluation.

4.1.2 Training Procedure

The DeepLabV3+ segmentation model (detailed in Section 3.2.1) was trained using a structured procedure to ensure robust learning and reliable evaluation. This encompassed selecting an appropriate loss function, optimizer, cross-validation strategy, and model persistence mechanism.

The primary objective function minimized was the **Focal Tversky Loss** [36], well-suited for semantic segmentation with class imbalance. It builds upon the Tversky Index (TI), which generalizes Dice and IoU by differentially weighting False Positives (FP) and False Negatives (FN). For a prediction P and ground truth G, TI is:

$$\mathrm{TI}(P,G) = \frac{|P \cap G| + \epsilon}{|P \cap G| + \alpha|G \setminus P| + \beta|P \setminus G| + \epsilon} \tag{4.1}$$

where ϵ is a small constant for numerical stability. The Tversky Loss (TL) is $1 - \mathrm{TI}$. The Focal Tversky Loss (FTL) further refines this with a focusing parameter γ to down-weight loss from well-classified examples, concentrating on harder-to-segment pixels:

$$\mathrm{FTL}(P,G) = (1 - \mathrm{TI}(P,G))^{\gamma} \tag{4.2}$$

In our experiments, FTL parameters were $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.75$. This choice of $\alpha > \beta$ imposes a higher penalty on false negatives, often desirable in medical imaging to minimize missed detections of true foreign objects.

Model parameters were optimized using the **Adam optimizer** [37], chosen for its adaptive learning rate capabilities and general effectiveness. A learning rate of 1×10^{-4} was used.

A **K-Fold Cross-Validation** strategy ($K = 5$ folds, as specified by $N_{SPLITS} = 5$) was employed to ensure robust evaluation and mitigate overfitting (see Section 3.2.2). The training dataset was divided accordingly, and for each fold, the model was trained for 50 epochs. In each iteration, one fold served as the validation set, with the remaining $K - 1$ folds for training. The model was initialized from scratch (or weights reset) for each fold to ensure independent training runs. The training batch size was 16.

Upon completing training for each fold, model weights were saved, enabling subsequent evaluation on its validation set and use in explainability analysis. The framework allowed loading pre-existing weights, though for K-Fold experiments, training was performed for each fold.

4.1.3 Evaluation Methods for Segmentation Performance

During each K-Fold Cross-Validation iteration, the trained DeepLabV3+ model instance was quantitatively assessed on the held-out validation fold using established overlap-based and boundary-based metrics. These metrics, common in medical image segmentation, capture different aspects of segmentation quality. Average performance and standard deviations across all folds were computed to summarize overall model efficacy and stability.

- **Overlap-based Metrics:** Evaluate the overlap between predicted and ground truth masks.
  - **Dice Coefficient (DSC):** Measures volumetric overlap between prediction (X) and ground truth (Y), sensitive to segmented region size [38]. DSC=1 indicates perfect overlap, 0 no overlap.

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FP + FN} \tag{4.3}$$

    where TP, FP, FN are true positives, false positives, and false negatives.

  - **Recall (Sensitivity):** Quantifies the fraction of the true positive region correctly identified, crucial in medical applications to minimize false negatives [39].

$$\text{Recall} = \frac{|X \cap Y|}{|Y|} = \frac{TP}{TP + FN} \tag{4.4}$$

- **Boundary-based Metrics:** Focus on predicted contour accuracy, vital for precise localization [40].

  - **Boundary IoU (BIoU):** Calculates IoU for pixels on predicted (∂X) and ground-truth (∂Y) boundaries, sensitive to contour alignment discrepancies [40].

$$\text{BIoU} = \frac{|\partial X \cap \partial Y|}{|\partial X \cup \partial Y|} \tag{4.5}$$

  - **Average Symmetric Surface Distance (ASSD):** Computes the average distance between predicted and ground-truth mask surface points [41]. Lower ASSD indicates better surface agreement.

$$\text{ASSD}(X, Y) = \frac{1}{|\partial X| + |\partial Y|} \left(\sum_{x_s \in \partial X} d(x_s, \partial Y) + \sum_{y_s \in \partial Y} d(y_s, \partial X) \right) \tag{4.6}$$

    where $d(p, S) = \min_{s \in S} ||p - s||$ is the minimum Euclidean distance from point p to surface S.

  - **95th Percentile Hausdorff Distance (HD95):** Measures the maximum surface distance, robust to outliers by using the 95th percentile [42]. It reflects the largest surface discrepancy, excluding the 5% most extreme outlier distances.

$$\text{HD}(X, Y) = \max \left(\max_{x_s \in \partial X} \min_{y_s \in \partial Y} ||x_s - y_s||, \max_{y_s \in \partial Y} \min_{x_s \in \partial X} ||y_s - x_s|| \right) \tag{4.7}$$

    HD95 is the 95th percentile of all point-to-surface distances.

4.1.4 Evaluation of Explainability Methods

Beyond segmentation model assessment, a comprehensive evaluation of Grad-CAM and NormGrad (introduced in Section 3.3) was conducted to ascertain their utility and reliability for IQA via foreign object identification.

For objective measurement, a **quantitative evaluation** of saliency maps used faithfulness and localization metrics:

- **Faithfulness Metrics:** Assess how accurately saliency maps reflect the model's decision-making for identifying foreign objects.
  - **Deletion and Insertion Area Under Curve (DAUC & IAUC):** Measure how well saliency maps capture influential regions. DAUC iteratively removes pixels by decreasing importance, observing performance drop (S). IAUC iteratively adds pixels by importance to a blurred image, observing performance increase [43]. Higher IAUC and lower DAUC (for performance drops) indicate better faithfulness. Let $S(k)$ be model performance with top k (insertion) or bottom k (deletion, i.e., least important removed first, or most important removed first depending on definition) pixels according to saliency, N_p total pixels.

    $$\text{IAUC} = \frac{1}{N_p} \sum_{k=1}^{N_p} S_{\text{insert}}(k) \tag{4.8}$$

    $$\text{DAUC} = \frac{1}{N_p} \sum_{k=1}^{N_p} S_{\text{delete}}(k) \tag{4.9}$$

    (Note: Formulation can vary by normalization and whether score increase/decrease is measured. For DAUC, lower values are better if $S_{\text{delete}}(k)$ represents performance after removing important pixels.)
  - **Drop in Confidence:** Quantifies the decrease in model confidence (or segmentation score) when salient regions are perturbed/removed. A larger drop suggests critical regions were identified, indicating better faithfulness [44]. Let $f(I)$ be confidence for original image I, $f(I_M)$ for image I with saliency mask M regions perturbed.

    $$\text{Drop in Confidence} = f(I) - f(I_M) \tag{4.10}$$

    (A higher positive value is better for a "drop".)
- **Localization Metrics:** Evaluate saliency map's ability to pinpoint relevant regions (foreign objects).
  - **Pointing Game Accuracy (PGA):** Assesses if the maximum saliency point falls within the ground-truth foreign object mask. Accuracy is the proportion

of hits [45]. Let N_{total} be total test images, $P_i^{\max}$ max saliency coordinates for image i, M_i^{GT} ground-truth mask for image i.

$$\text{PGA} = \frac{\sum_{i=1}^{N_{\text{total}}} \mathbb{I}(P_i^{\max} \in M_i^{\text{GT}})}{N_{\text{total}}} \tag{4.11}$$

where $\mathbb{I}(\cdot)$ is the indicator function (1 if true, 0 otherwise).

These metrics provide a standardized, objective means to compare Grad-CAM and NormGrad. Illustrative examples and detailed discussion follow in subsequent results sections.

4.1.5 Implementation Details

The experimental pipeline was implemented primarily in Python. Key libraries included **PyTorch** [46] for deep learning, **segmentation-models-pytorch** [47] for DeepLabV3+, and **Albumentations** [48] for image augmentation. **Scikit-learn** [49] was used for K-Fold cross-validation and metrics. The dataset was accessed via the **Kaggle API** (‘kagglehub‘ library). Standard libraries (NumPy, Pandas, Matplotlib, OpenCV) and utilities (tqdm) were also employed. For LLM-generated summaries, `Meta-Llama-3-8B-Instruct.Q4_0.gguf` was used via `GPT4All` [50] (`max_tokens=2024`). The interactive qualitative evaluation interface (Section 4.2.2) used **Gradio** [51].

Experiments were conducted in a Jupyter Notebook environment. Model training, evaluation, and saliency map generation used a workstation with an NVIDIA GeForce RTX 3090 GPU (24GB VRAM), and the system possessed 125GB of RAM. GPU acceleration (‘device: cuda‘) was leveraged.

Core hyperparameters for K-Fold training are in Table 4.1, applied consistently across folds. NormGrad’s ϵ was 1×10^{-5} for numerical stability.

Table 4.1: Key Hyperparameters and Configuration Settings.

Parameter	Value
Image Size	256 × 256 pixels
Batch Size	16
Epochs per Fold	50
Number of Folds (K)	5
Learning Rate	1×10^{-4}
Optimizer	Adam
Loss Function α (FN weight)	0.7
Loss Function β (FP weight)	0.3
Loss Function γ (Focal)	0.75
Random Seed	42
Evaluation Threshold (segmentation)	0.5
NormGrad ϵ	1×10^{-5}
Device	CUDA GPU
Dataset Size (Training Split)	9000 images
Number of Workers (Dataloader)	2
GPT4All `max_tokens`	2024

These implementation choices and resulting metrics form the basis for subsequent discussions on model behavior and XAI utility.

4.2 Experimental Results

This section presents and discusses the results from the described experiments, covering quantitative segmentation performance, qualitative and quantitative XAI assessments, and model complexity.

4.2.1 Quantitative Analysis

This subsection details the quantitative performance of the DeepLabV3+ segmentation model and the XAI methods, organized into segmentation model evaluation, XAI technique assessment, and baseline model comparison.

4.2.1.1 Segmentation Model Performance

The DeepLabV3+ model's ability to segment foreign objects was evaluated using 5-Fold Cross-Validation. Average performance metrics (Table 4.2) provide an overall assessment, while per-fold metrics (Table 4.3) show variability across data partitions.

Table 4.2: Average K-Fold Cross-Validation Performance (Mean ± Std. Dev.).

Metric	Mean Value ± Std. Dev.
Dice Coefficient (DSC)	0.6482 ± 0.0118
Recall (Sensitivity)	0.7008 ± 0.0118
Boundary IoU (BIoU)	0.4943 ± 0.0061
Average Symmetric Surface Distance (ASSD)	15.9032 ± 1.5337 (px)
Hausdorff Distance 95% (HD95)	57.1983 ± 5.3253 (px)
Training Time per Fold (s)	4622.3565 ± 58.7702
Evaluation Time per Fold (s)	24.3479 ± 0.4443

Table 4.3: Per-Fold Performance Metrics on Validation Set.

Fold	DSC	Recall	BIoU	ASSD (px)	HD95 (px)	Train Time (s)	Eval Time (s)
1	0.6431	0.6955	0.4895	14.4877	52.0988	4724.57	24.53
2	0.6608	0.6971	0.5037	14.1236	51.0948	4591.51	24.92
3	0.6307	0.6858	0.4952	16.8116	59.5075	4581.03	24.50
4	0.6559	0.7146	0.4882	16.4143	60.0178	4596.01	23.94
5	0.6506	0.7110	0.4948	17.6786	(missing)	4638.66	23.85

The DeepLabV3+ model achieved a mean DSC of 0.6482 ± 0.0118 (Table 4.2), suggesting good, though not perfect, overlap with ground truth foreign objects. The Recall of 0.7008 ± 0.0118 indicates identification of 70% of actual foreign object pixels, aligning with the Focal Tversky Loss's emphasis on penalizing false negatives. Boundary IoU (0.4943 ± 0.0061) indicates moderate contour delineation accuracy. ASSD (15.9032 ± 1.5337 pixels on 256×256 images) and HD95 (57.1983 ± 5.3253 pixels) reveal that while average boundary agreement is fair, instances of larger local deviations exist.

Critically, these segmentation metrics may be conservative due to inherent limitations in the Object-CXR dataset's ground truth annotations. Exploratory data analysis revealed inconsistencies: bounding boxes without discernible objects, oversized/inaccurate bounding boxes, and unannotated visible foreign objects. Such inaccuracies can penalize correct predictions or fail to reward true detections, potentially underestimating the model's true proficiency on accurately annotated data. The per-fold results (Table 4.3) show some variability, but the small standard deviations for most metrics (e.g., 0.0118 for DSC, 0.0061 for BIoU) in Table 4.2 indicate largely consistent performance across data splits. This suggests good stability, generalization, and reproducibility of the training process, crucial for assessing potential reliability.

4.2.1.2 Quantitative Evaluation of Explainability Methods

The quantitative performance of Grad-CAM, NormGrad Order 0, and NormGrad Order 1 was assessed using metrics from Section 4.1.4. Table 4.4 shows average performance across 5 folds (500 'affected' samples/fold, combined saliency from 138 layers). Table 4.5 provides per-fold details.

Table 4.4: Average Quantitative XAI Metrics (Mean ± Std. Dev.) Across 5 Folds. DAUC: Lower is better. IAUC, Drop in Performance, Pointing Game Accuracy: Higher is better.

XAI Method	DAUC	IAUC	Drop in Performance	Pointing Game Acc.	Eval Time (s) per Fold
Grad-CAM	0.0267 ± 0.0063	0.1658 ± 0.0180	0.3401 ± 0.0260	0.5119 ± 0.0125	1113.74 ± 9.99
NormGrad Order 0	0.0399 ± 0.0063	0.2046 ± 0.0254	0.3172 ± 0.0298	0.2603 ± 0.0510	1127.11 ± 4.77
NormGrad Order 1	0.0398 ± 0.0064	0.2054 ± 0.0255	0.3168 ± 0.0292	0.2683 ± 0.0559	5662.65 ± 17.22

Table 4.5: Per-Fold Quantitative XAI Metrics on Validation Set. (GC: Grad-CAM, NG0: NormGrad Order 0, NG1: NormGrad Order 1).

Fold	XAI Method	DAUC	IAUC	Drop in Perf.	PGA	Eval Time (s)
1	GC	0.0191	0.1821	0.3582	0.5258	1118.16
	NG0	0.0324	0.2233	0.3414	0.2083	1131.48
	NG1	0.0318	0.2259	0.3417	0.2202	5662.12
2	GC	0.0278	0.1662	0.3580	0.5218	1100.84
	NG0	0.0450	0.2074	0.3341	0.3234	1129.05
	NG1	0.0452	0.2066	0.3329	0.3294	5654.77
3	GC	0.0288	0.1364	0.3026	0.4940	1105.35
	NG0	0.0404	0.1692	0.2735	0.2163	1122.52
	NG1	0.0412	0.1705	0.2724	0.2083	5688.23
4	GC	0.0355	0.1785	0.3587	0.5079	1122.70
	NG0	0.0471	0.2326	0.3379	0.3016	1131.04
	NG1	0.0463	0.2327	0.3352	0.3214	5666.76
5	GC	0.0224	0.1657	0.3229	0.5099	1121.65
	NG0	0.0348	0.1906	0.2989	0.2520	1121.47
	NG1	0.0345	0.1912	0.3017	0.2619	5641.38

Grad-CAM achieved the lowest average DAUC (0.0267 ± 0.0063), suggesting better identification of influential pixels for removal, and the highest PGA (0.5119 ± 0.0125), indicating superior saliency peak alignment with ground-truth objects. NormGrad methods (Orders 0 and 1) showed higher IAUC values (approx. 0.205 ± 0.025) than Grad-CAM

(0.1658 ± 0.0180). Drop in Performance was comparable, with Grad-CAM slightly higher. NormGrad Order 1 was significantly more computationally expensive (approx. 5663s/fold) than Grad-CAM (approx. 1114s) and NormGrad Order 0 (approx. 1127s). Per-fold metrics (Table 4.5) generally reflected these trends, with Grad-CAM's PGA consistently higher.

4.2.1.3 Baseline Model Comparison for Classification

To contextualize our model's ability to distinguish images with foreign objects from those without (binary classification), we benchmarked its performance against reimplementations of ResNet34 and EfficientNet-B0, previously used by Ozer et al. [23]. Our DeepLabV3+ model was adapted for classification: output logits underwent sigmoid activation, thresholding (0.01) for a binary mask, then connected components analysis. If any component area >1 pixel, the image was classified as "Object" (positive); otherwise, "No Object" (negative).

Results are in Table 4.6 and Figure 4.1.

Table 4.6: Comparative Classification Performance Metrics. "Affected"/"Normal" for ResNet34/EfficientNet-B0; "Object"/"No Object" for adapted DeepLabV3+.

Model	**Class Label**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Specificity**
ResNet34 (reimplemented)	Class 1 (Affected)	0.8460	0.8089	0.9060	0.8547	0.7860
	Class 0 (Normal)	0.8460	0.8932	0.7860	0.8362	0.9060
EfficientNet-B0 (reimplemented)	Class 1 (Affected)	0.8600	0.8435	0.8840	0.8633	0.8360
	Class 0 (Normal)	0.8600	0.8782	0.8360	0.8566	0.8840
DeepLabV3+ (adapted)	Class 1 (Object)	0.8030	0.9662	0.6280	0.7612	0.9780
	Class 0 (No Object)	0.8030	0.7244	0.9780	0.8323	0.6280

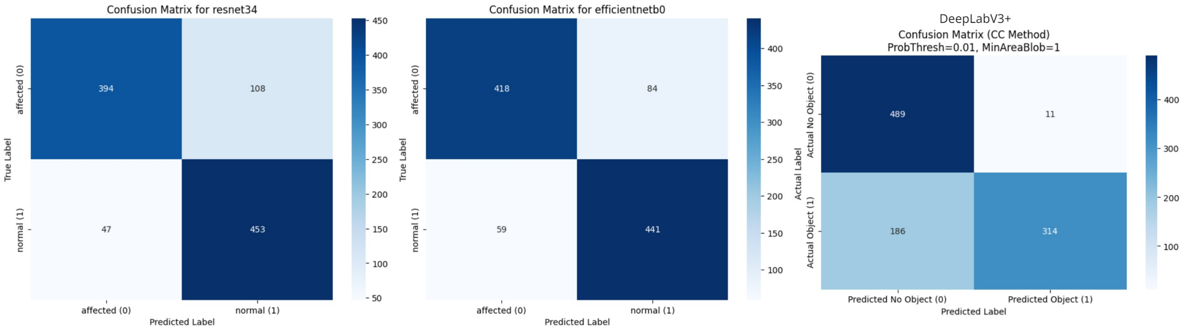


Figure 4.1: Confusion matrices for the binary classification task (Object vs. No Object / Affected vs. Normal). From left to right: ResNet34 (reimplemented), EfficientNet-B0 (reimplemented), and DeepLabV3+ (adapted).

Performance Discussion: EfficientNet-B0 had the highest accuracy (0.8600), followed by ResNet34 (0.8460), then adapted DeepLabV3+ (0.8030). However, DeepLabV3+ showed exceptional precision (0.9662) for the "Object" class, substantially higher than ResNet34 (0.8089) and EfficientNet-B0 (0.8435), indicating high reliability of its positive predictions. Its recall for "Object" (0.6280) was lower than ResNet34 (0.9060) and EfficientNet-B0 (0.8840). DeepLabV3+ also achieved the highest specificity for "Object" (0.9780). The F1-score for "Object" was lower (0.7612) for DeepLabV3+, primarily due to its lower reported recall. These metrics suggest DeepLabV3+ prioritizes precision and specificity, a valuable trade-off in medical contexts where false positives are costly.

Impact of Annotation Quality: As noted in Section 4.2.1.1, Object-CXR dataset annotation inconsistencies (e.g., oversized/misplaced/missing annotations) disproportionately affect segmentation-based approaches like ours, which are more sensitive to precise localization than classification models. To investigate the low recall, we analyzed misclassified cases (Figures 4.2, 4.3).

Manual review of all misclassifications revealed that 74/186 (approx. 40%) false negatives were ground truth errors (GT indicated an object, but none visible). Similarly, 6/11 (approx. 55%) false positives were correct model detections of objects missed in GT. Accounting for these GT errors significantly improves the true recall of our adapted DeepLabV3+ model and further validates its high precision. For the "No Object" class, DeepLabV3+ showed very high recall (0.9780). This analysis underscores that while standard metrics might suggest lower performance, DeepLabV3+ excels in precision and specificity, and its recall is artificially suppressed by GT issues. Its ability to provide pixel-level localization is an inherent advantage over pure classifiers.

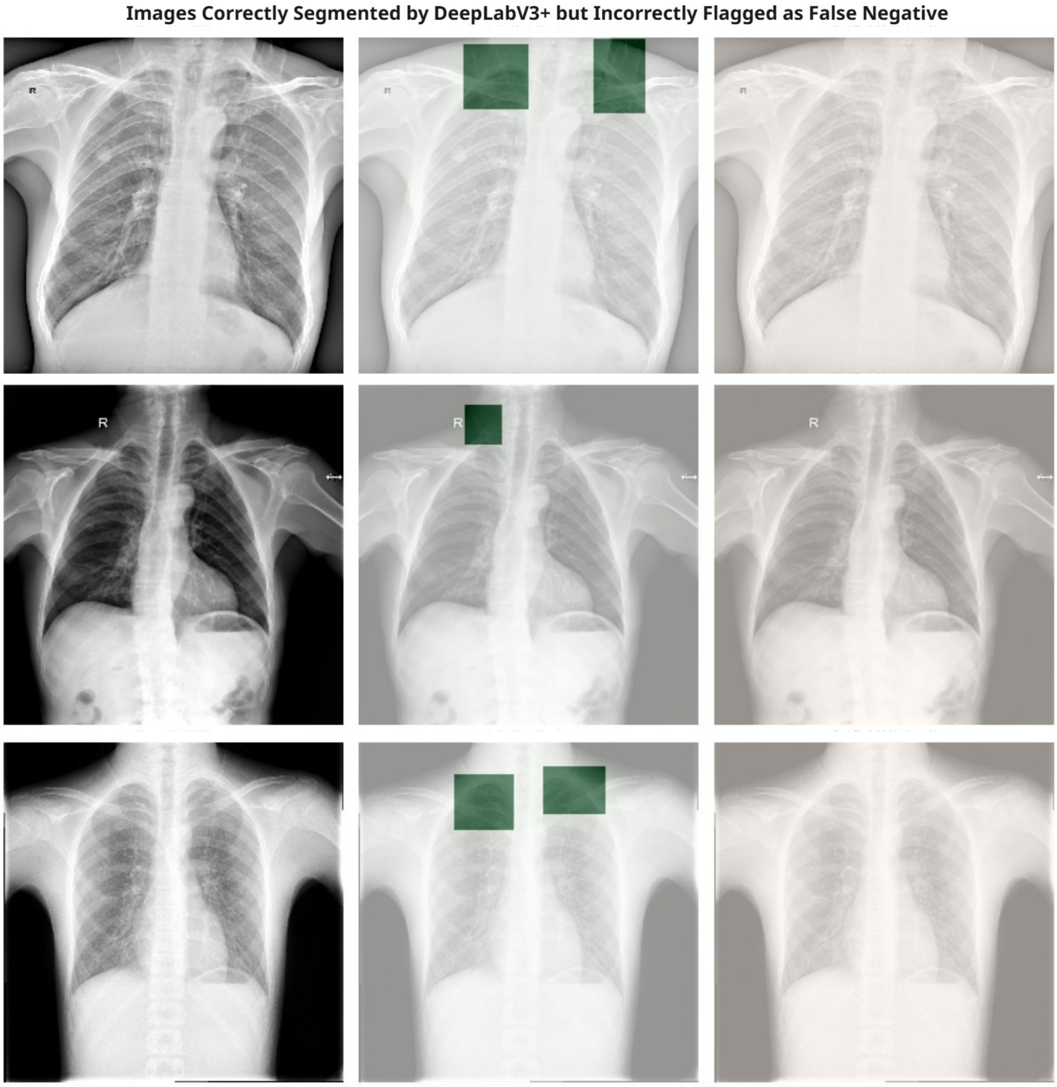

Images Correctly Segmented by DeepLabV3+ but Incorrectly Flagged as False Negative
R
R
R
R
R
R

Figure 4.2: Examples of false negative predictions. Left: Original X-ray. Middle: Ground truth annotation. Right: Model prediction (empty). Manual verification revealed GT errors where no object was present.

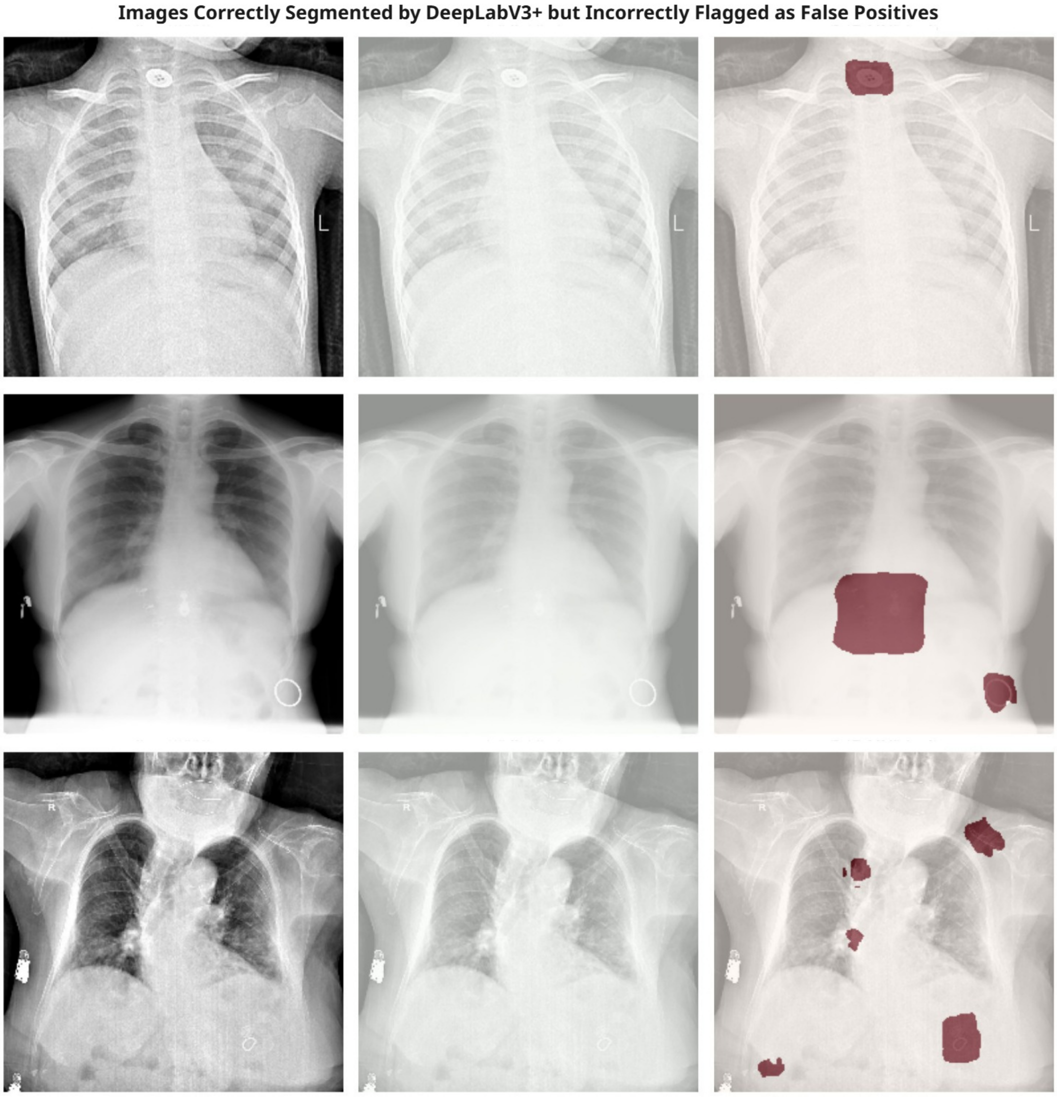

Images Correctly Segmented by DeepLabV3+ but Incorrectly Flagged as False Positives
L
L
L
R
R
R

Figure 4.3: Examples of false positive predictions. Left: Original X-ray. Middle: Ground truth (no annotation). Right: Model prediction. Manual verification confirmed correct model identification of unannotated objects.

Pointing Game Accuracy of Baseline Classifiers' Explanations: To compare explainability, PGA was evaluated for reimplemented ResNet34 and EfficientNet-B0 using Grad-CAM and NormGrad variants, considering all major convolutional/feature blocks. Table 4.7 compares these with DeepLabV3+'s XAI PGA.

Table 4.7: Comparative Pointing Game Accuracy (PGA) for Baseline Classification Models and DeepLabV3+ Segmentation Model. Baseline PGA from dedicated XAI runs; DeepLabV3+ values from Table 4.4.

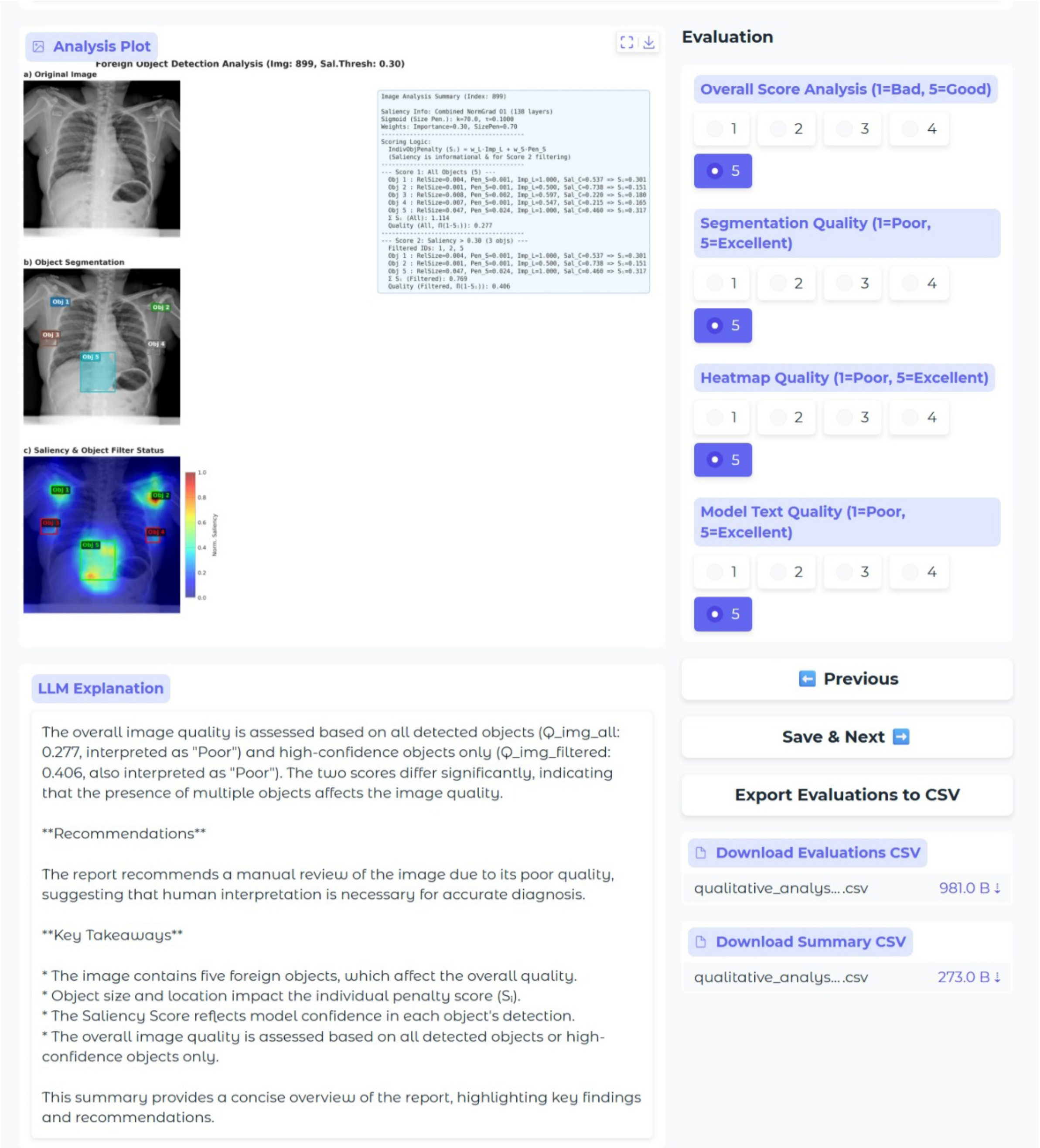
Model	**XAI Method**	**Mean PG Accuracy**
ResNet34 (reimpl.)	NormGrad (conv1x1, all layers)	0.206
	GradCAM	0.186
EfficientNet-B0 (reimpl.)	NormGrad (conv3x3, all layers)	0.178
	GradCAM	0.102
DeepLabV3+ (adapted seg.)	Grad-CAM	0.512
	NormGrad Order 0	0.260
	NormGrad Order 1	0.268

DeepLabV3+ with Grad-CAM achieved a PGA of 0.512, substantially higher than baseline classifiers (best PGA $\approx$ 0.206). This indicates that in over half the cases, DeepLabV3+'s explanation correctly localized the foreign object, a stark contrast to classifiers where explanations mislocalized objects in 80% of cases. This 2.5× improvement highlights the advantage of segmentation architectures, trained for precise spatial delineation, in generating more trustworthy and accurately localized explanations. This difference is clinically significant, enhancing trust and reducing risks associated with misinterpretation.

4.2.2 Qualitative Analysis

Complementing quantitative metrics, this subsection qualitatively assesses generated explanations (Grad-CAM, NormGrad, LLM summaries) for practical utility, interpretability, and trustworthiness in IQA for foreign object detection. A custom Gradio interface [51] (Figure 4.4) facilitated systematic evaluation. Reviewers scored 50 test images (1-5 Likert scale: 1=Bad/Poor, 5=Good/Excellent) across four dimensions: Overall Score Analysis Quality, Segmentation Quality, Heatmap Quality, and Model Text Quality. Each dimension was assessed independently.

Grad-CAM heatmaps, consistent with higher quantitative PGA (Table 4.4), generally provided well-localized, focused attributions, perceived as more directly interpretable.



```
Image Analysis Summary (Index: 899)

Saliency Info: Combined NormGrad O1 (138 layers)
Sigmoid (Size Pen.): k=70.0, τ=0.1000
Weights: Importance=0.30, SizePen=0.70
----------------------------------------
Scoring Logic:
  IndivObjPenalty (Sᵢ) = w_L·Imp_L + w_S·Pen_S
  (Saliency is informational & for Score 2 filtering)
----------------------------------------
--- Score 1: All Objects (5) ---
  Obj 1 : RelSize=0.004, Pen_S=0.001, Imp_L=1.000, Sal_C=0.537 => Sᵢ=0.301
  Obj 2 : RelSize=0.001, Pen_S=0.001, Imp_L=0.500, Sal_C=0.738 => Sᵢ=0.151
  Obj 3 : RelSize=0.008, Pen_S=0.002, Imp_L=0.597, Sal_C=0.220 => Sᵢ=0.180
  Obj 4 : RelSize=0.007, Pen_S=0.001, Imp_L=0.547, Sal_C=0.215 => Sᵢ=0.165
  Obj 5 : RelSize=0.047, Pen_S=0.024, Imp_L=1.000, Sal_C=0.460 => Sᵢ=0.317
  Σ Sᵢ (All): 1.114
  Quality (All, Π(1-Sᵢ)): 0.277
----------------------------------------
--- Score 2: Saliency > 0.30 (3 objs) ---
  Filtered IDs: 1, 2, 5
  Obj 1 : RelSize=0.004, Pen_S=0.001, Imp_L=1.000, Sal_C=0.537 => Sᵢ=0.301
  Obj 2 : RelSize=0.001, Pen_S=0.001, Imp_L=0.500, Sal_C=0.738 => Sᵢ=0.151
  Obj 5 : RelSize=0.047, Pen_S=0.024, Imp_L=1.000, Sal_C=0.460 => Sᵢ=0.317
  Σ Sᵢ (Filtered): 0.769
  Quality (Filtered, Π(1-Sᵢ)): 0.406
```

Evaluation

Overall Score Analysis (1=Bad, 5=Good)

○ 1 ○ 2 ○ 3 ○ 4 ● 5

Segmentation Quality (1=Poor, 5=Excellent)

○ 1 ○ 2 ○ 3 ○ 4 ● 5

Heatmap Quality (1=Poor, 5=Excellent)

○ 1 ○ 2 ○ 3 ○ 4 ● 5

Model Text Quality (1=Poor, 5=Excellent)

○ 1 ○ 2 ○ 3 ○ 4 ● 5

qualitative_analys....csv 981.0 B ↓

qualitative_analys....csv 273.0 B ↓

LLM Explanation

The overall image quality is assessed based on all detected objects (Q_img_all: 0.277, interpreted as "Poor") and high-confidence objects only (Q_img_filtered: 0.406, also interpreted as "Poor"). The two scores differ significantly, indicating that the presence of multiple objects affects the image quality.

Recommendations

The report recommends a manual review of the image due to its poor quality, suggesting that human interpretation is necessary for accurate diagnosis.

Key Takeaways

* The image contains five foreign objects, which affect the overall quality.
* Object size and location impact the individual penalty score (S_i).
* The Saliency Score reflects model confidence in each object's detection.
* The overall image quality is assessed based on all detected objects or high-confidence objects only.

This summary provides a concise overview of the report, highlighting key findings and recommendations.

Figure 4.4: Gradio-based interface for systematic qualitative evaluation of segmentation outputs, saliency maps, and LLM-generated explanations, allowing synchronized viewing and scoring.

Table 4.8: Summary of Mean Qualitative Evaluation Scores (N=50 images). Scores: 1 (Poor/Bad) to 5 (Excellent/Good).

Evaluation Dimension	**Mean Score**	**Std. Deviation**
Overall Score Analysis Quality	4.28	1.23
Segmentation Quality (Visual)	4.48	0.95
Heatmap Quality (Saliency)	4.62	0.83
Model Text Quality (LLM)	4.80	0.57

However, visual inspection of the heatmaps revealed that NormGrad often produced more reliable and meaningful attributions, despite Grad-CAM's superior quantitative scores. NormGrad maps (combined multi-layer) were often more diffuse, sometimes capturing broader context or multiple objects but occasionally obscuring precise contributions of specific parts. The choice depends on user needs: precise localization (Grad-CAM) versus broader, less focused attribution (NormGrad). Overall Heatmap Quality (4.62 ± 0.83) was high.

LLM-generated summaries received exceptionally high scores for Model Text Quality (4.80 ± 0.57), indicating consistent success in producing clear, relevant, and accurate narratives synthesizing quantitative findings and saliency information. Visual Segmentation Quality (4.48 ± 0.95) and Overall Score Analysis Quality (4.28 ± 1.23) also scored favorably.

It is important to contextualize these favorable qualitative results. A significant portion of the 50 images in the qualitative analysis set contained no foreign objects. The model's high accuracy in correctly identifying these "empty" cases (high specificity of 0.9780, Table 4.6) likely contributed to the high average scores, as evaluating correct negative predictions is straightforward. A qualitative set with more challenging cases (subtle or multiple objects) might yield different ratings, especially for segmentation and heatmap quality.

In summary, qualitative analysis highlighted Grad-CAM's localization precision and NormGrad's broader contextual view. LLM summaries effectively translated complex data into user-friendly language, enhancing overall system interpretability.

4.2.3 Model Complexity Analysis

This subsection details the computational complexity (space and time) of the DeepLabV3+ model and its XAI methods, informing feasibility and aiding comparison.

4.2.3.1 Space Complexity

Space complexity includes static (model parameters) and dynamic (activations, gradients) memory. In this context, $K_{\text{params\_total}}$ denotes the total number of model parameters. For data and activations, B represents the batch size, C the number of channels (e.g., 3 for RGB images, 1 for masks, or feature channels for layers), H and W the height and width of the input or feature maps respectively. Num_Features refers to the number of feature maps (output channels) of a layer, while H_{layer} and W_{layer} are its spatial dimensions, and C_{layer} refers to its number of channels.

Static and Data Memory: DeepLabV3+ (ResNet50 backbone, 26.68M parameters) requires 101.77MB (float32) for static parameter storage ($O(K_{\text{params\_total}})$). Input data batch memory ($B = 16$, $256 \times 256 \times 3$ float32 images) is 12MB for images and 4MB for masks ($O(B \cdot C \cdot H \cdot W)$).

Dynamic Memory (Training): Activations ($B \cdot \text{Num\_Features} \cdot H_{\text{layer}} \cdot W_{\text{layer}}$), parameter gradients ($O(K_{\text{params\_total}})$), and optimizer states (e.g., Adam, $O(K_{\text{params\_total}})$) significantly contribute.

XAI Memory: Grad-CAM is modest. NormGrad Order 0 adds an input tensor clone and target layer data ($O(BCHW + BC_{\text{layer}}H_{\text{layer}}W_{\text{layer}})$). NormGrad Order 1 is most demanding, temporarily needing two model copies and all original model gradients ($\approx 2 \cdot O(K_{\text{params\_total}})$ plus gradient storage). Key components are summarized in Table 4.9.

Table 4.9: Summary of Key Space Complexity Components.

Component	**Theoretical Space Complexity**
Model Parameters	$O(K_{\text{params\_total}})$
Input Data Batch (Images/Masks)	$O(B \cdot C \cdot H \cdot W)$
Activations (per layer, training)	$O(B \cdot \text{Num\_Features} \cdot H_{\text{layer}} \cdot W_{\text{layer}})$
Parameter Gradients (training)	$O(K_{\text{params\_total}})$
Optimizer States (e.g., Adam)	$O(K_{\text{params\_total}})$
NormGrad O0 (additional)	$O(BCHW + BC_{\text{layer}}H_{\text{layer}}W_{\text{layer}})$
NormGrad O1 (peak, model copy)	$\approx 2 \cdot O(K_{\text{params\_total}}) + \text{gradients}$

4.2.3.2 Time Complexity

Time complexity refers to computational effort. Here, L is the number of layers in the CNN. H, W are the spatial dimensions of feature maps, C_{in} and C_{out} are the input and output channels for a layer respectively, and KernelSize is the spatial dimension of a convolutional kernel (e.g., 3 for a 3x3 kernel). B is the batch size, C is the number of image channels (for pixel-wise loss), N_{train} is the total number of training samples, and $N_{\text{layers\_xai}}$ is the count of layers included in combined XAI computations. T_{fwd}, T_{bwd}, $T_{\text{partial\_bwd}}$, $T_{\text{bwd\_to\_target}}$, and $T_{\text{bwd\_full}}$ represent time for a forward pass, a full backward pass, a partial backward pass, a backward pass to a target layer, and a full backward pass respectively. $T_{\text{XAI\_single}}$ is the time for a single layer XAI computation.

Core Model & Training: CNN forward pass (L layers) is roughly $O(L \cdot HWC_{\text{in}}C_{\text{out}} \cdot \text{KernelSize}^2)$. Pixel-wise loss is $O(BCHW)$/batch. Training epoch involves N_{train}/B iterations of (forward + loss + backward + optimizer step). Evaluation involves forward passes and metric calculations.

XAI Cost: Grad-CAM needs one forward and one partial backward pass ($\approx T_{\text{fwd}} + T_{\text{partial\_bwd}}$). NormGrad Order 0 (forward, backward to target, plus operations, $\approx T_{\text{fwd}} + T_{\text{bwd\_to\_target}}$) is empirically similar in time to Grad-CAM (Table 4.4). NormGrad Order 1 is far more intensive ($\approx 2 \cdot (T_{\text{fwd}} + T_{\text{bwd\_full}})$). Combined multi-layer XAI scales with $N_{\text{layers\_xai}}$ ($O(B \cdot N_{\text{layers\_xai}} \cdot T_{\text{XAI\_single}})$). Key components are summarized in Table 4.10.

Table 4.10: Summary of Key Time Complexity Components.

Operation/Component	Theoretical Time Complexity
Model Forward Pass (CNN)	$O(L \cdot HWC_{\text{in}}C_{\text{out}}\text{Kernel}^2)$
Pixel-wise Loss (per batch)	$O(BCHW)$
Training Epoch (simplified)	$O(N_{\text{train}}/B \cdot (T_{\text{fwd}} + T_{\text{bwd}})) + T_{\text{eval}}$
Grad-CAM (single layer)	$\approx T_{\text{fwd}} + T_{\text{partial\_bwd}}$
NormGrad O0 (single layer)	$\approx T_{\text{fwd}} + T_{\text{bwd\_to\_target}}$
NormGrad O1 (single layer)	$\approx 2 \cdot (T_{\text{fwd}} + T_{\text{bwd\_full}})$
Combined XAI (multi-layer)	$O(B \cdot N_{\text{layers\_xai}} \cdot T_{\text{XAI\_single}})$

4.2.3.3 Computational Resources and Practical Implications

Experiments used an NVIDIA RTX 3090 (24GB VRAM). Batch size (16) and image dimensions (256 × 256) were chosen considering VRAM. While the 100MB model size is manageable, dynamic training memory and NormGrad Order 1's demands are significant. Training deep segmentation models is time-consuming. Advanced XAI like NormGrad Order 1, despite potentially richer explanations, is far costlier than Grad-CAM or NormGrad Order 0. This power vs. expense trade-off is key for deployment. Empirical runtimes (Tables 4.2, 4.4, and consolidated in Table 4.11) confirm these points.

Table 4.11: Empirical Runtimes for Key System Components.

Component/Operation	**Time (s)**	**Notes**
Model Training and Evaluation (per fold)		
Model Training	4622.36	Mean (Table 4.2)
Model Evaluation (segmentation)	24.35	Mean (Table 4.2)
XAI Saliency Map Generation (per fold, 500 images)		
Grad-CAM	1113.74	Table 4.4
NormGrad Order 0	1127.11	Table 4.4
NormGrad Order 1	5662.65	Table 4.4
Visualization Pipeline (per single image, mean over test runs)		
Image Objects Analysis	4.49	50 image test (Std Dev: 5.48)
Plotting Results	0.68	50 image test (Std Dev: 0.15)
LLM Model Load	1.32	50 image test (Std Dev: 0.06)
LLM Generation	3.33	50 image test (Std Dev: 0.68)
Combined Heatmap Generation	11.17	20 image test (Std Dev: 0.36)
Total Pipeline	10.75	50 image test (Std Dev: 6.20)

As detailed in Table 4.11, model training is a lengthy process, averaging approximately 4622s per fold. XAI methods also vary in cost: NormGrad Order 1 (approx. 5663s per 500 images) is notably more time-consuming than Grad-CAM (approx. 1114s) or NormGrad Order 0 (approx. 1127s). For the per-image visualization pipeline, which has a reported mean total processing time of approximately 10.75s, individual components such as combined heatmap generation (approx. 11.17s) and LLM-based summary generation (approx. 3.33s for generation, plus 1.32s for model loading) represent significant contributions to latency. These figures are crucial for resource budgeting and assessing the practical deployment feasibility, particularly for applications requiring rapid processing or operating under constrained computational resources.

Summary

This chapter detailed the comprehensive empirical evaluation of the proposed XAI framework for IQA, focusing on foreign object detection in medical X-rays. It described the experimental setup (Object-CXR dataset, DeepLabV3+ with Focal Tversky Loss, 5-Fold CV, evaluation metrics for segmentation and XAI), implementation details, and presented an in-depth analysis of quantitative and qualitative results, baseline comparisons, and model complexity.

Quantitative analysis showed the DeepLabV3+ model achieved a mean DSC of 0.6482 ± 0.0118 and Recall of 0.7008 ± 0.0118. These scores were noted as potentially conservative due to ground truth annotation inconsistencies in the Object-CXR dataset. For XAI methods, Grad-CAM demonstrated superior localization (PGA 0.5119 ± 0.0125) and faithfulness (DAUC 0.0267 ± 0.0063), while NormGrad Order 1 was computationally much more expensive.

Comparative analysis against baseline classifiers (ResNet34, EfficientNet-B0) revealed the adapted DeepLabV3+ model's exceptional precision (0.9662) and specificity (0.9780). Its lower reported recall (0.6280) was significantly impacted by ground truth errors (approx. 40% of FNs). Crucially, XAI applied to DeepLabV3+ yielded vastly superior PGA (e.g., Grad-CAM PGA 0.512) compared to baselines (PGA $\approx 0.1-0.2$), underscoring the enhanced reliability of explanations from the segmentation architecture.

Qualitative analysis via a Gradio interface rated segmentation mask quality highly (4.48/5). Grad-CAM maps were perceived as more focused, while LLM-generated summaries effectively synthesized findings (4.80/5). The favorable scores were contextualized by the presence of many "empty" (no object) images in the qualitative set.

Model complexity analysis detailed space (DeepLabV3+ parameters 101.77MB) and time demands, with empirical runtimes highlighting the cost of NormGrad Order 1 and LLM generation, crucial for deployment considerations.

In conclusion, this chapter systematically validated the XAI-IQA system, demonstrating robust object segmentation and, importantly, more accurate and localized explanations than standard classification-based XAI. Findings highlighted the impact of dataset quality on evaluation and the utility of combining visual saliency with LLM summaries for enhanced interpretability, providing a comprehensive understanding of the system's strengths, limitations, and computational aspects for medical IQA.

Chapter 5

Conclusion

This thesis embarked on an investigation into enhancing the interpretability of Image Quality Assessment (IQA) systems, particularly within high-stakes domains such as medical imaging where understanding the rationale behind automated assessments is paramount. Addressing the prevalent "black-box" nature of many contemporary Artificial Intelligence (AI) models, this research focused on developing and evaluating a novel framework for eXplainable Artificial Intelligence (XAI) tailored to IQA tasks. The primary objective was to move beyond mere predictive accuracy and provide transparent, human-understandable insights into how image quality determinations are made, specifically demonstrated through the task of foreign object detection in medical X-ray images.

5.1 Summary of the Research

The research commenced with a critical review of the state-of-the-art in both IQA and XAI, identifying a significant gap in integrative approaches that combine the predictive power of deep learning for IQA with the explanatory capabilities of XAI. To bridge this gap, a comprehensive XAI-IQA framework was proposed and implemented. This framework is centered around a DeepLabV3+ semantic segmentation model, utilizing a ResNet50 backbone, trained to identify foreign objects in Chest X-ray images—a task directly related to assessing diagnostic image quality.

The core of the explainability component involved the integration and adaptation of gradient-based saliency methods, specifically Grad-CAM and NormGrad (Orders 0 and 1), to visually highlight image regions most influential to the segmentation model's predictions. Beyond visual explanations, a sophisticated *Visualization and Scoring Engine* was developed. This engine processes the segmentation outputs and saliency maps to:

1. Extract per-object characteristics (size, location relative to a Region of Interest, saliency score).
2. Calculate an *Individual Object Penalty* based on these characteristics.
3. Derive two overall image quality scores: Q_{img} (based on all detected objects) and $Q_{img}^{filtered}$ (based on objects exceeding a saliency threshold), providing a nuanced quality assessment.
4. Generate a composite visual report including the original image, object overlays, and saliency heatmaps.
5. Leverage a Large Language Model (LLM) to produce natural language summaries explaining the automated assessment, thereby enhancing the accessibility of the findings.

The efficacy of this framework was rigorously evaluated through a 5-Fold Cross-Validation scheme on the Object-CXR dataset. The evaluation encompassed quantitative metrics for both segmentation performance (e.g., Dice Coefficient, Boundary IoU) and XAI method effectiveness (e.g., Pointing Game Accuracy, Deletion/Insertion AUC, Drop in Confidence). Qualitative assessments were conducted using a custom-built Gradio interface, and a comparative analysis was performed against baseline classification models (ResNet34, EfficientNet-B0) to contextualize the system's capabilities. Model complexity in terms of space and time was also analyzed.

5.2 Key Contributions and Findings

This thesis makes several key contributions to the field of explainable IQA:

- **Novel XAI-IQA Framework:** The primary contribution is the design and implementation of an end-to-end framework that integrates a deep learning segmentation model with advanced XAI techniques and an LLM-powered reporting mechanism for explainable IQA. This system not only predicts the presence of quality-degrading elements but also explains *why* and *where* these elements are identified.
- **Superiority of Segmentation-based XAI for Localization:** A crucial finding was that applying XAI methods to a segmentation architecture (DeepLabV3+) yielded vastly superior localization accuracy (e.g., Grad-CAM PGA ≈ 0.512) compared to applying similar XAI techniques to baseline classification models (PGA

$\approx 0.1 - 0.2$). This underscores the inherent advantage of task-specific architectures (segmentation for localization) in generating more reliable and trustworthy explanations for spatially-grounded IQA tasks.

- **Comprehensive Evaluation of XAI Methods for IQA:** The research provided a detailed quantitative and qualitative comparison of Grad-CAM and NormGrad variants within the IQA context. Grad-CAM demonstrated better faithfulness (DAUC 0.0267 ± 0.0063) and localization (PGA 0.5119 ± 0.0125) and was computationally more efficient. NormGrad methods, while offering potentially richer insights into training dynamics, were more computationally intensive, particularly NormGrad Order 1.

- **Development of an Advanced Visualization and Scoring Engine:** The engine, with its dual-score system and integration of object-specific penalties based on size, location, and model confidence (saliency), offers a nuanced and interpretable approach to quantifying image quality based on detected anomalies. The incorporation of LLM-generated summaries (rated highly at 4.80/5 for quality) significantly enhances the interpretability and accessibility of the automated analysis for end-users.

- **Highlighting the Impact of Dataset Quality:** The study revealed significant ground truth annotation inconsistencies within the Object-CXR dataset. While the DeepLabV3+ model achieved a mean DSC of 0.6482 ± 0.0118, this was deemed conservative. Manual analysis showed that approximately 40% of False Negatives were due to GT errors (no visible object where one was annotated). Accounting for these improved the effective recall and validated the model's high precision (0.9662) and specificity (0.9780) in its adapted classification role. This finding emphasizes the critical need for high-quality, precisely annotated datasets in developing and benchmarking IQA and XAI systems.

- **Robust Segmentation Performance with Explainability:** The DeepLabV3+ model itself demonstrated robust performance in identifying foreign objects, forming a solid foundation for the subsequent XAI analysis. The system's ability to provide accurate segmentations is a prerequisite for meaningful explanations.

5.3 Limitations of the Study

Despite the promising results, this research has several limitations that should be acknowledged:

- **Dataset-Specific Focus:** The framework was primarily developed and validated on the Object-CXR dataset for foreign object detection in medical X-rays. While this serves as a strong proof-of-concept for explainable IQA in a critical domain, the generalizability of the specific findings and the scoring engine parameters to other IQA tasks, image modalities, or distortion types requires further investigation.

- **Annotation Quality of the Dataset:** As discussed, the inherent inconsistencies in the Object-CXR dataset's ground truth annotations likely impacted the reported segmentation metrics, potentially underestimating the model's true performance. This also complicates the precise evaluation of XAI methods that rely on accurate ground truth for localization.

- **Computational Cost of Advanced XAI and LLM Integration:** While Grad-CAM was relatively efficient, NormGrad Order 1 exhibited significant computational overhead. Furthermore, the LLM-based summary generation, while valuable, adds to the per-image processing latency, which could be a concern for real-time applications.

- **Parameter Tuning for Scoring Engine:** The Visualization and Scoring Engine employs several parameters (e.g., weights for importance and size penalties, saliency thresholds). While default values were used based on domain understanding, optimal settings might vary across different applications or datasets, potentially requiring domain-specific tuning.

- **Subjectivity in Qualitative Evaluation:** Although a structured interface was used, qualitative assessments of heatmap quality and LLM summaries inherently contain a degree of subjectivity. The favorable scores might also be influenced by the dataset composition (many "empty" images) in the qualitative test set.

- **Limited Scope of XAI Techniques Explored:** The study focused on gradient-based saliency methods. Other XAI paradigms, such as example-based explanations or counterfactuals, were not investigated but could offer complementary insights.

5.4 Future Research Directions

The findings and limitations of this thesis open up several avenues for future research:

- **Broader Applicability and Generalization:** Extend and adapt the XAI-IQA framework to a wider range of IQA tasks, including the assessment of common image distortions (blur, noise, compression artifacts) across diverse image types (natural images, other medical modalities).
- **Development of Computationally Efficient XAI Methods:** Investigate or develop XAI techniques that offer a comparable level of explanatory detail to methods like NormGrad Order 1 but with significantly reduced computational demands, making them more suitable for practical deployment.
- **User Studies for Practical Utility Assessment:** Conduct formal user studies with domain experts (e.g., radiologists for medical IQA) to evaluate the practical utility, trustworthiness, and impact of the generated explanations and integrated quality scores on their diagnostic confidence and workflow.
- **Creation of High-Quality Benchmarking Datasets:** Address the challenge of dataset limitations by contributing to the development of large-scale, meticulously annotated datasets specifically designed for explainable IQA research, covering various quality attributes and image types.
- **Exploration of Inherently Interpretable Models:** Investigate the potential of designing IQA models that are inherently interpretable (e.g., using attention mechanisms designed for transparency, or hybrid symbolic-AI approaches), reducing the reliance on post-hoc XAI techniques.
- **Advanced LLM Integration and Prompt Engineering:** Further explore the capabilities of LLMs in the XAI-IQA loop, including more sophisticated prompt engineering, interactive querying of explanations, and generating multi-modal explanations.

5.5 Concluding Remarks

The increasing complexity and ubiquity of AI-driven image assessment tools necessitate a paradigm shift towards more transparent and interpretable systems. This thesis has contributed to this endeavor by proposing and validating a comprehensive framework for Explainable Artificial Intelligence for Image Quality Assessment. By integrating deep learning-based semantic segmentation with advanced saliency methods, a novel scoring engine, and LLM-generated textual summaries, this work has demonstrated a viable pathway to not only achieve accurate IQA but also to provide meaningful insights into the decision-making processes of these AI systems.

The empirical results, particularly the superior localization accuracy of explanations derived from segmentation models and the positive reception of LLM-synthesized reports, highlight the potential of such integrated systems to foster trust, facilitate debugging, and enhance user understanding in critical IQA applications. While challenges related to dataset quality, computational efficiency, and evaluation standardization remain, the research presented here lays a strong foundation for future advancements. It is hoped that this work will inspire further research into developing more robust, efficient, and genuinely explainable IQA systems, ultimately leading to their safer and more effective deployment across various domains, especially where image quality is intrinsically linked to critical outcomes. The journey towards truly trustworthy AI in image quality assessment is ongoing, and explainability is an indispensable compass on this path.

Bibliography

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera.
Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence.
Information Fusion, 99:101805, 2023.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera.
Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.
Information Fusion, 58:82–115, 2020.

[3] Volker Strobel.
Pold87/academic-keyword-occurrence: First release, April 2018.

[4] Ashish Sharma.
Consumer perception and attitude towards the visual elements in social campaign advertisement.
IOSR Journal of Business and Management, 3:6–17, 2012.

[5] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi.
Metaiqa: Deep meta-learning for no-reference image quality assessment, 2020.

[6] V. Karthikeyan and C. Jayanthi.
Advancements in image quality assessment: a comparative study of image processing and deep learning techniques.
The Scientific Temper, 2024.

[7] Xiaohan Yang, Fan Li, and Hantao Liu.
Ttl-iqa: Transitive transfer learning based no-reference image quality assessment.
IEEE Transactions on Multimedia, 23:4326–4340, 2021.

[8] Han Cui, Alfredo De Goyeneche, Efrat Shimron, Boyuan Ma, and Michael Lustig.
Reference-free image quality metric for degradation and reconstruction artifacts, 2024.

[9] Avinab Saha, Sandeep Mishra, and Alan C. Bovik.
Re-iqa: Unsupervised learning for image quality assessment in the wild, 2023.

[10] Hanhe Lin, Vlad Hosu, and Dietmar Saupe.
Deepfl-iqa: Weak supervision for deep iqa feature learning.
arXiv preprint arXiv:2001.08113, 2020.

[11] Kwan-Yee Lin and Guanxiang Wang.
Hallucinated-iqa: No-reference image quality assessment via adversarial learning, 2018.

[12] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani.
No-reference image quality assessment via transformers, relative ranking, and self-consistency, 2022.

[13] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi.
There and back again: Revisiting backpropagation saliency methods, 2020.

[14] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, Hakan Bilen, and Andrea Vedaldi.
Normgrad: Finding the pixels that matter for training, 2019.

[15] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum.
What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research.
Artificial Intelligence, 296:103473, 2021.

[16] Ibomoiye Domor Mienye, George Obaido, Nobert Jere, Ebikella Mienye, Kehinde Aruleba, Ikiomoye Douglas Emmanuel, and Blessing Ogbuokiri.
A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges.
Informatics in Medicine Unlocked, 51:101587, 2024.

[17] Erico Tjoa and Cuntai Guan.
A survey on explainable artificial intelligence (xai): Toward medical xai.
IEEE Transactions on Neural Networks and Learning Systems, 32(11):4793–4813, Nov 2021.

[18] Caner Ozer and Ilkay Oksuz.
Explainable image quality analysis of chest x-rays.
In Mattias Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schläfer, and Floris Ernst, editors, *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 567–580. PMLR, 07–09 Jul 2021.

[19] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever.
Explainable artificial intelligence (xai) in deep learning-based medical image analysis.
Medical Image Analysis, 79:102470, 2022.

[20] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang.
A comprehensive study of multimodal large language models for image quality assessment, 2024.

[21] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen.
Sf-iqa: Quality and similarity integration for ai generated image quality assessment.
In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6692–6701, June 2024.

[22] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma.
Blind image quality assessment via vision-language correspondence: A multitask learning perspective, 2023.

[23] Caner Ozer, Arda Guler, Aysel Turkvatan Cansever, and Ilkay Oksuz.
Explainable image quality assessment for medical imaging, 2023.

[24] Shahrukh Athar and Zhou Wang.
A comprehensive performance evaluation of image quality assessment algorithms.
IEEE Access, 7:140030–140070, 2019.

[25] T. Song, L. Li, H. Zhu, and J. Qian.
ie-iqa: intelligibility enriched generalizable no-reference image quality assessment.
Frontiers in Neuroscience, 15, 2021.

[26] Deepti Ghadiyaram and Alan C. Bovik.
Massive online crowdsourced study of subjective and objective picture quality.
IEEE Transactions on Image Processing, 25(1):372–387, Jan 2016.

[27] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin.
Topiq: A top-down approach from semantics to distortions for image quality assessment, 2023.

[28] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe.
Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment.
IEEE Transactions on Image Processing, 29:4041–4056, 2020.

[29] Amina Adadi and Mohammed Berrada.
Peeking inside the black-box: A survey on explainable artificial intelligence (xai).
IEEE Access, 6:52138–52160, 2018.

[30] M. K. Surehli, N. Aggarwal, and G. Joshi.
DeepLabV3Plus-PyTorch: A DeepLab V3+ Model with ResNet 50 Encoder to perform Binary Segmentation Tasks. Implemented with PyTorch.
https://github.com/mukund-ks/DeepLabV3Plus-PyTorch, aug 2023.
Repository content as of 2023-08-06. Accessed [Insert Date Accessed].

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition, 2015.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei.
Imagenet large scale visual recognition challenge, 2015.

[33] N. Sharma, S. Gupta, D. H. Elkamchouchi, and S. Bharany.
Encoder-decoder variant analysis for semantic segmentation of gastrointestinal tract using uw-madison dataset.
Bioengineering (Basel), 12(3):309, Mar 2025.

[34] Tobias Clement, Truong Thanh Hung Nguyen, Mohamed Abdelaal, and Hung Cao.
Xai-enhanced semantic segmentation models for visual quality inspection, 2024.

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.
Grad-cam: Visual explanations from deep networks via gradient-based localization.
International Journal of Computer Vision, 128(2):336–359, October 2019.

[36] Nabila Abraham and Naimul Mefraz Khan.
A novel focal tversky loss function with improved attention u-net for lesion segmentation, 2018.

[37] Diederik P. Kingma and Jimmy Ba.
Adam: A method for stochastic optimization, 2017.

[38] Reuben R Shamir, Yuval Duchin, Jinyoung Kim, Guillermo Sapiro, and Noam Harel.
Continuous dice coefficient: a method for evaluating probabilistic segmentations, 2019.

[39] Junjiao Tian, Niluthpol Mithun, Zach Seymour, Han-Pang Chiu, and Zsolt Kira.
Striking the right balance: Recall loss for semantic segmentation, 2022.

[40] Fan Sun, Zhiming Luo, and Shaozi Li.
Boundary difference over union loss for medical image segmentation, 2023.

[41] Martina Finocchiaro, Ronja Stern, Abraham George Smith, Jens Petersen, Kenny Erleben, and Melanie Ganz.
Hqcolon: A hybrid interactive machine learning pipeline for high quality colon labeling and segmentation, 2025.

[42] Thien B. Nguyen-Tat, Hoang-An Vo, and Phuoc-Sang Dang.
Qmaxvit-unet+: A query-based maxvit-unet with edge enhancement for scribble-supervised segmentation of medical images.
Computers in Biology and Medicine, 187:109762, March 2025.

[43] Tristan Gomez, Thomas Fréour, and Harold Mouchère.
Metrics for saliency map evaluation of deep learning explanation methods, 2022.

[44] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian.
Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks.

In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

[45] Ruozhen He, Ziyan Yang, Paola Cascante-Bonilla, Alexander C. Berg, and Vicente Ordonez.
Learning from synthetic data for visual grounding, 2024.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala.
Pytorch: An imperative style, high-performance deep learning library, 2019.

[47] Pavel Iakubovskii.
Segmentation models pytorch.
https://github.com/qubvel/segmentation_models.pytorch, 2019.

[48] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin.
Albumentations: Fast and flexible image augmentations.
Information, 11(2), 2020.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

[50] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar.
Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo.
https://github.com/nomic-ai/gpt4all, 2023.

[51] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou.
Gradio: Hassle-free sharing and testing of ml models in the wild, 2019.