



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



UGR

Universidad
de Granada

Department of Mathematics
Computer Science

Image Quality Assessment in Forensic Facial Comparison

Master Candidate: Mahshad Golafshan

Supervisor

Lamberto Ballan
University of Padova

Co-Supervisors

Pablo Mesejo Santiago
University of Granada

Enrique Bermejo Nieves
Panacea Cooperative Research

Granada/Padova, September 2025

Image Quality Assessment in Forensic Facial Comparison

Keywords: Forensic facial comparison, image quality assessment, face recognition, surveillance imagery, forensic identification, morphological analysis, biometrics, computer vision, deep learning.

Abstract

In forensic science, the reliability of facial comparison heavily depends on the quality of the images analyzed. This thesis examines the role of Image Quality Assessment (IQA) in facial comparison, focusing specifically on how different image characteristics impact the accuracy. Existing general-purpose IQA methods are evaluated, revealing significant limitations when applied to realistic forensic scenarios such as surveillance footage or uncontrolled imaging environments. To address these issues, a specialized IQA framework tailored explicitly for facial comparison tasks is proposed, integrating objective computational metrics and subjective assessments from trained examiners. The experimental evaluation conducted and demonstrates that the proposed framework is strongly correlated (up to 83%) with examiner decisions, outperforming standard IQA metrics by approximately 25%. These findings underline the importance of developing standardized IQA protocols tailored for forensic workflows, thus enhancing the reliability of facial comparisons and supporting higher evidentiary standards in identity verification contexts.

“That which can be seen is always only a perspective, not the truth itself.”

— *Friedrich Nietzsche, *On the Genealogy of Morals**

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisors, Pablo Mesejo Santiago, Enrique Bermejo, and Lamberto Ballan, for giving me the opportunity to develop this project under their guidance. I am deeply thankful for their unwavering patience, insightful feedback, and continuous support whenever I encountered challenges throughout this journey.

I would also like to thank the University of Granada and the University of Padova for offering me the invaluable opportunity to pursue my studies across both institutions, which has greatly enriched my academic and personal experience.

Finally, my heartfelt thanks go to my parents, whose love and support have been my greatest strength—especially through the long distance from home.

Contents

| | |
|--|-----------|
| Acknowledgments | 6 |
| 1 Introduction | 13 |
| 1.1 Problem Statement | 13 |
| 1.2 Motivation | 15 |
| 1.3 Objectives | 17 |
| 1.4 Project Planning | 18 |
| 2 Theoretical Background | 20 |
| 2.1 Machine Learning and Deep Learning | 20 |
| 2.1.1 Machine Learning | 20 |
| 2.1.2 Deep Learning | 21 |
| 2.2 Face Detection | 23 |
| 2.2.1 YOLO | 25 |
| 2.2.2 Single Shot MultiBox Detector | 28 |
| 2.2.3 SSD vs YOLO | 30 |
| 2.3 Image Quality Assessment | 30 |
| 2.4 Face Image Quality Assessment | 31 |
| 3 State of the art | 33 |
| 3.1 State of the art of FIQA | 33 |
| 3.2 FIQA in Forensic Applications | 34 |
| 3.3 Integration of YOLO in FIQA Pipelines | 35 |
| 3.4 Taxonomy of FIQA Methods | 35 |
| 3.4.1 Single-Score Prediction Methods | 36 |
| 3.4.2 Component-Based Estimation: OFIQ | 37 |
| 3.4.3 Machine Learning and Deep Learning Approaches for FIQA | 38 |
| 4 Materials and Methods | 40 |
| 4.1 Dataset | 40 |
| 4.2 Dataset Usage and Partitioning | 42 |
| 4.3 Model: OFIQ | 43 |
| 4.4 Evaluation | 47 |
| 4.4.1 Comparison of OFIQ and YOLO Bounding Boxes | 47 |
| 4.4.2 Landmark Accuracy Across Different Facial Yaw Angles | 49 |

| | | |
|----------|---|-----------|
| 5 | Implementation and Experiments | 51 |
| 5.1 | Implementation Overview | 51 |
| 5.2 | Datasets and Protocol | 52 |
| 5.3 | Experiments and Results | 53 |
| 5.3.1 | E1. OFIQ (SSD) acceptance by camera | 53 |
| 5.3.2 | E2. YOLO detector configuration (small-face recovery) | 54 |
| 5.3.3 | E3. OFIQ on YOLO crops | 54 |
| 5.3.4 | E4. Failure gap: YOLO vs. OFIQ (whole dataset) | 55 |
| 5.3.5 | E5. Inter-eye distance: code verification | 55 |
| 5.3.6 | E6. Yaw threshold and landmark visibility | 55 |
| 5.3.7 | E7. FIPP metric (face scale) | 57 |
| 5.4 | Discussion | 57 |
| 6 | Conclusion and Future Work | 58 |
| 6.1 | Conclusion | 58 |
| 6.1.1 | Summary of Contributions | 58 |
| 6.2 | Future Work | 59 |
| | Bibliography | 60 |

List of Figures

| | | |
|------|--|----|
| 1.1 | Example of image quality used for biometric. | 14 |
| 1.2 | Examples of facial images from a single subject, with associated Quality Scores (QS). | 16 |
| 1.3 | General workflow underlying the development of the proposed OFIQ framework. | 17 |
| 2.1 | Supervised learning and unsupervised learning. | 21 |
| 2.2 | Graphical example of a neural network. (a) and (b) show a biological and an artificial neuron, respectively. (c) illustrates a synapse. (d) shows a shallow neural network (left) and a deep neural network (right). | 22 |
| 2.3 | CNN applied to a biomedical image classification problem. | 23 |
| 2.4 | Overview of object detection methods categorized into two-stage and one-stage approaches. | 24 |
| 2.5 | Comparison between one-stage and two-stage object detectors. | 24 |
| 2.6 | Schematic diagram of the YOLO object detection pipeline. | 25 |
| 2.7 | Comparing YOLO precision with other detection models. | 26 |
| 2.8 | SSD framework | 29 |
| 2.9 | Summary of IQA subproblems | 31 |
| 2.10 | Summary of facial image validation under different CCTV conditions | 32 |
| 3.1 | Annual number of Scopus-indexed publications from 2009 to 2024 related to FIQA | 34 |
| 3.2 | Taxonomy of Face Image Quality Assessment (FIQA) methods | 36 |
| 4.1 | Illustration of the viewpoints and environments associated with each surveillance camera type used in this study | 41 |
| 4.2 | Outdoor and indoor camera setup during data acquisition. | 42 |
| 4.3 | The usage of checkerboard in Normalize scale and perspective of frames | 42 |
| 4.4 | Overview of the OFIQ framework | 44 |
| 4.5 | ADNet diagram. how ADNet draw the points. | 45 |
| 4.6 | Semantic and location of the 98 landmarks output by ADNet | 45 |
| 4.7 | Bounding box comparison between YOLO and SSD | 47 |
| 4.8 | Illustration of the three primary facial pose variations | 49 |
| 4.9 | Example sequence showing face rotation across yaw angles from 0° to 180°. | 49 |
| 4.10 | Superimposition of facial landmarks and segmentation mask on a profile face image. | 50 |

| | | |
|-----|---|----|
| 5.1 | Side-by-side comparison of the Baseline OFIQ pipeline (SSD-based, left) and the Enhanced OFIQ pipeline (YOLO-based, right). | 52 |
| 5.2 | YOLO cropping experiment | 55 |
| 5.3 | Example of facial landmark annotation illustrating various anthropometric measurements. | 56 |

List of Tables

| | | |
|-----|--|----|
| 1.1 | Initial planning of the project. | 18 |
| 1.2 | The final project planning. | 19 |
| 1.3 | Total hours and days worked. | 19 |
| 1.4 | Final project cost estimate. | 19 |
| 2.1 | Comparison of YOLO and SSD for face detection. | 30 |
| 3.1 | Comparison of representative single-score FIQA models with respect to supervision strategy, network architecture, score type, and interpretability. | 37 |
| 4.1 | Capture-related and subject-related quality components assessed by the OFIQ framework. | 44 |
| 4.2 | Face detection outcome for YOLO (red) and OFIQ (green) on the first frame shown in Figure 4.7 | 48 |
| 4.3 | Face detection outcome for YOLO (red) and OFIQ (green) on the second frame shown in Figure 4.7. | 48 |
| 5.1 | OFIQ quality metrics and their corresponding image dependencies. | 53 |
| 5.2 | OFIQ (SSD) positives by camera. Camera totals are from the report; Camera #1 positive rate is given as 22.79%. Camera #3 had 0 positives; the Camera #2 count is implied by the reported total of 609 positives. | 54 |
| 5.3 | YOLO cropping statistics on 3,632 images. Lowering <code>min_box_size</code> from 50 to 25 drastically reduces missed small faces. | 54 |
| 5.4 | OFIQ outputs over YOLO crops. A non-trivial portion of YOLO crops still yields no OFIQ face box $(-1, -1, -1, -1)$ | 54 |
| 5.5 | summarizes three anthropometric measures: lateral head size, vertical head size, and inter-eye distance. | 56 |
| 5.6 | Yaw vs. two-eye visibility from tilt frames and SCD samples. | 56 |
| 5.7 | FIPP examples from our experiments. Values reflect relative size of face to image area. | 57 |

Chapter 1

Introduction

1.1 Problem Statement

Forensic identification is a specialized field within forensic science [36] that seeks to determine or confirm the identity of individuals—whether living or deceased—based on physical, biological, or digital evidence. This process plays a vital role in a variety of contexts: criminal investigations (e.g., identifying suspects or victims), legal proceedings (e.g., verifying identity claims), and humanitarian efforts (e.g., locating missing persons or disaster victim identification) [36].

A key technique within forensic identification is the analysis of facial features [42, 77]. However, the terminology used in this context often requires clarification. Facial identification refers broadly to the process of determining an individual’s identity using their facial features [20]. Facial recognition, in contrast, typically describes an automated, algorithm-driven approach that compares a given face against a database of known individuals [62]. Finally, forensic facial comparison is a more specialized process performed manually or semi-automatically by trained experts [6, 12]. This method involves a detailed, often case-specific comparison between two facial images—such as a surveillance frame and a mugshot—with the aim of confirming or excluding identity under legal scrutiny [6, 73].

The effectiveness of facial comparison—automated or expert-led—heavily depends on the quality of the images being analyzed. Ideal conditions, such as high resolution, good lighting, and frontal poses, support accurate and reliable comparisons. However, real-world conditions are far from perfect. Surveillance footage is often grainy, poorly lit, or affected by motion blur; passport photos may show outdated appearances; suspects may wear disguises or masks. These factors contribute to a wide range of **image quality limitations** that challenge both human and algorithmic comparison methods [7, 5].

Despite significant advances in artificial intelligence and image processing, several persistent challenges remain in the forensic application of facial comparison [6].

- **Image Quality Issues:** Degradations such as poor lighting, low resolution, motion blur, and compression artifacts hinder accurate feature extraction and comparison [5].
- **Variability in Facial Appearance:** Aging, facial hair, expressions, makeup,

and pose variation introduce inconsistencies in facial features that complicate recognition [71].

- **Subjectivity in Manual Comparisons:** Expert analysis, while valuable, may introduce inconsistencies and subjective judgments into the comparison process [7].
- **Bias in Automated Systems:** AI-driven facial recognition technologies have been shown to exhibit demographic performance disparities, raising ethical concerns around fairness and bias [59].

Some of these metrics which can lead to substantial errors in forensic face matching are shown in the Figure 1.1.

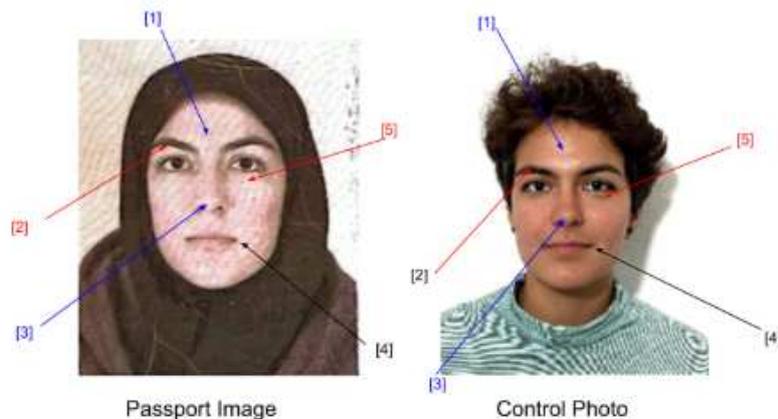


Figure 1.1: Example of image quality used for biometric identification in passport control. The numbered points correspond to cephalometric landmarks that should match in both images for them to be considered as belonging to the same person. Factors such as illumination, resolution, occlusions (e.g., a scarf), and inter-eye distance play a crucial role in accurate identification.

These challenges underscore the importance of assessing image quality as part of the forensic workflow. **Image Quality Assessments (IQA)** is the field dedicated to quantifying the perceptual quality of images. Traditionally, this has been done using subjective metrics like Mean Opinion Score, where human raters judge image quality, or objective models that attempt to replicate human judgments through computational means [53]. However, many deep learning-based IQA methods used in commercial or research contexts operate as “black boxes,” providing a single scalar score without explaining how that score was derived. In forensic scenarios—where transparency, explainability, and auditability are paramount—this lack of interpretability poses a significant problem [59]. To address this issue, we turn to **Open Source Facial Image Quality (OFIQ)**, a framework developed specifically for facial image quality assessment in biometric and forensic contexts [39]. Unlike traditional IQA models, OFIQ does not generate a single opaque score. Instead,

it performs a component-wise analysis of image quality, offering interpretable insights into key factors such as facial alignment, sharpness, lighting, occlusion, and resolution. As an open-source and modular tool, OFIQ allows forensic experts and developers to tailor quality assessment criteria to their specific domain needs. This makes it particularly well-suited for applications where evidence must be interpreted in court or where transparency in decision-making is legally mandated.

In this thesis, we aim to explore and enhance the OFIQ framework, with a focus on improving its performance in challenging forensic scenarios—especially images featuring occlusions or non-frontal (profile) views. These are among the most common and problematic cases encountered in real-world investigations. By improving OFIQ’s ability to robustly evaluate facial image quality under such conditions, we contribute toward more interpretable, reliable, and forensically valid tools for facial comparison and identification.

1.2 Motivation

The OFIQ framework was originally developed to evaluate facial image quality in biometric applications, such as identity verification, passport control, and access authentication [2, 24, 39]. In these contexts, images are typically captured under controlled conditions characterized by consistent lighting, frontal pose, and high resolution [29, 72]. OFIQ provides interpretable and component-wise quality scores (e.g., alignment, sharpness, occlusion) to help systems determine whether an image meets recognition standards [11, 39].

With recent advances in technology, particularly the development of deep learning algorithms, have dramatically reduced error rates in automated face recognition systems. Despite these improvements, recognition performance can still be significantly affected by external factors. These include the imaging process—such as lighting conditions, camera angle, resolution, and motion blur—as well as the level of cooperation from the subject being photographed. For example, individuals may not always face the camera directly, may move during capture, or their faces may be partially occluded by objects like hats, glasses, or masks. These variations introduce challenges that reduce the reliability of face recognition, especially in uncontrolled or forensic scenarios.



Figure 1.2: Examples of facial images from a single subject, with associated Quality Scores (QS). The QS indicates the suitability of images for face recognition systems, with higher scores denoting better quality. (a) has the highest score (95), clearly showing facial details against a uniform background. (b) receives a slightly lower score (85) due to insufficient separation between the face and background. (c) has a moderate score (62), reflecting the impact of facial shadows and a cluttered background. (d) scores lowest (42), significantly affected by strong shadows and poor visibility of facial details [39].

However, forensic facial comparison operates in a vastly different setting. Images are often extracted from CCTV footage, social media, or personal devices, where conditions are uncontrolled and quality is frequently compromised [6, 21, 46, 79]. These images may suffer from low resolution, poor lighting, occlusions (e.g., masks or sunglasses), and non-frontal views, significantly reducing the reliability of typical biometric frameworks [10, 21]. As a result, the use of IQA tools in forensics demands frameworks that remain robust, interpretable, and reliable under these real-world degradations [43, 46].

Although OFIQ presents a modular and transparent approach to IQA, its underlying assumptions—optimized for biometric-quality images—limit its direct applicability to forensic cases. Key challenges include:

- Poor generalization to severely degraded or occluded images [10, 21];
- Lack of validation in forensic-like datasets [6, 46];
- Absence of forensic-relevant metrics such as inter-eye distance, face-to-image pixel proportion (FIPP), or head size [22, 43].

These limitations emphasize the need for adapting and rigorously evaluating OFIQ in forensic contexts, ensuring outputs remain interpretable and actionable by forensic experts [12, 73].

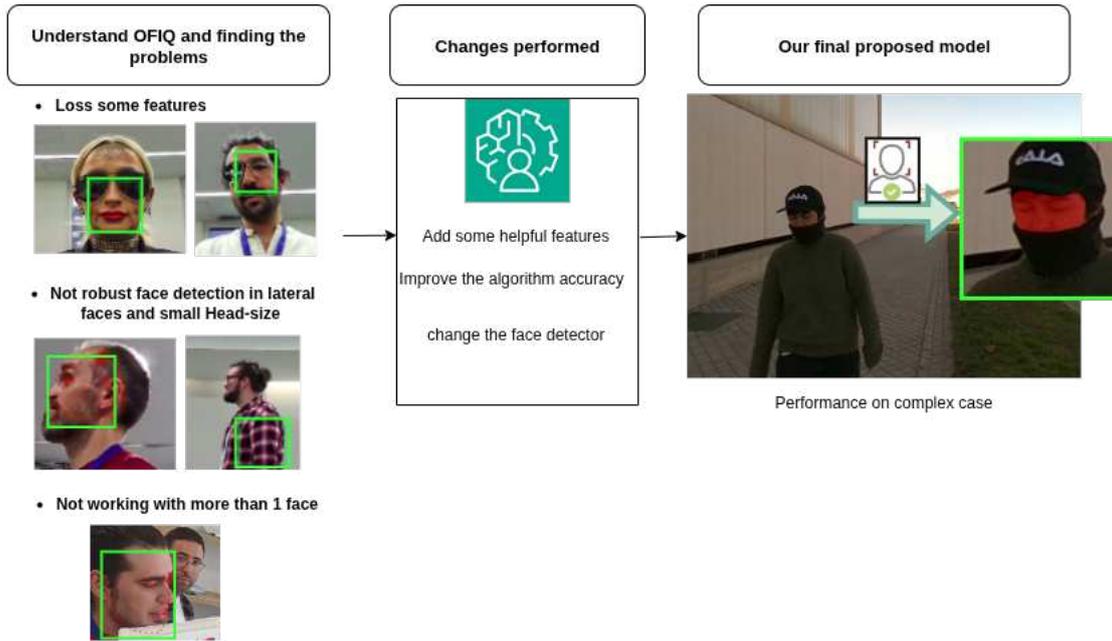


Figure 1.3: General workflow underlying the development of the proposed OFIQ framework. The first stage (left block) consists of a diagnostic analysis of prior methods, highlighting recurrent deficiencies such as limited robustness of face detection under heterogeneous imaging conditions, partial loss of salient facial features, and inadequate handling of multi-face inputs. The intermediate stage (central block) outlines the methodological refinements introduced to mitigate these shortcomings, including the integration of additional discriminative features, algorithmic optimization to enhance accuracy, and substitution of the baseline face detector with a more resilient alternative. The final stage (right block) demonstrates the resulting system’s capability to generalize effectively to complex forensic scenarios, with improved reliability of image quality assessment even in challenging operational environments.

1.3 Objectives

The main objective of this thesis is to develop and validate an automatic IQA method specifically tailored for forensic facial comparison scenarios, characterized by uncontrolled, degraded, and highly variable conditions.

To achieve this main objective, the following sub-objectives are defined:

- Evaluate the performance of the original OFIQ framework on forensic-like images to identify its strengths and limitations.
- Adapt and extend the OFIQ framework to effectively handle forensic-specific challenges, including occlusions, lateral views, and low-resolution images.
- Integrate additional forensic-specific quality indicators, such as inter-eye distance, FIPP and head size to enhance robustness and interpretability.
- Validate the adapted IQA framework using real-world or synthetic forensic datasets, assessing its utility, accuracy, and reliability within investigative

workflows.

1.4 Project Planning

The planning of this MSc Thesis has been structured according to the official academic load of 33 ECTS credits, which translates into an estimated workload of approximately 1080 hours. With the second semester spanning 6 months, this corresponds to an average dedication of 30 hours per week, or 6 hours per day over a standard five-day work week.

The nature of the project does not present a significant complexity in terms of scope and requirements, and the team does not include a large group of people whose comust be synchronized, which allows addressing its development through a waterfall model approach [64]. However, this approach avoids backtracking in any of the phases of the cycle, and although the system design and requirements are expected to be stable, there is the possibility of minor adjustments as more the problem and formation methods is obtained. This is why we use a small variant, the feedback version. The phases of the waterfall model are:

- **Requirements Analysis:** Initial meetings with the project supervisors were conducted to define the objectives and scope.
- **Design:** Based on the findings from the analysis phase, suitable methods and tools were selected. Preliminary experiments were also designed to evaluate feasibility and define the structure of the experimental framework
- **Implementation:** This phase involved the adaptation and integration of selected techniques, the development of necessary functionalities, and the creation of a robust model for the project’s objectives
- **Testing and Evaluation:** A series of experiments were conducted to simulate relevant distortions, validate the methodology, and assess model performance in terms of quality evaluation and reliability

The initial planning is shown in the Gantt diagram displayed in Table 1.1.

| Task | Weeks – Hours | October | November | December | January | February | March |
|-----------------------|---------------|---------|----------|----------|---------|----------|-------|
| Requirements Analysis | 4 – 60 | | | | | | |
| Design | 6 – 90 | | | | | | |
| Implementation | 9 – 135 | | | | | | |
| Tests | 6 – 90 | | | | | | |

Table 1.1: Initial planning of the project.

In addition, some delays were expected to occur, especially in the implementation and writing the thesis in a academic form as can be seen in Table 1.2, given the novelty of the proposal and the difficulty of the problem. In particular, for example, finetuning the code was an iterative and manual process that took longer than expected.

| Task | Weeks - Hours | October | November | December | January | February | March | April | May | June |
|-----------------------|---------------|---------|----------|----------|---------|----------|-------|-------|-----|------|
| Requirements Analysis | 3 - 51 | | | | | | | | | |
| Design | 5 - 75 | | | | | | | | | |
| Implementation | 8 - 120 | | | | | | | | | |
| Tests | 7 - 85 | | | | | | | | | |
| Writing | 12 - 155 | | | | | | | | | |

Table 1.2: The final project planning.

To carry out this project, the following materials were taken into account:

| | |
|-------------------|----------------------------|
| Start date | 01/10/2024 |
| End date | 10/06/2025 |
| Duration | 240 days, 217 working days |

Table 1.3: Total hours and days worked.

| Item | Cost |
|--------------------|----------------|
| Salary | 1 250.00€ |
| Chat GPT | 160.00€ |
| Google Colab Pro | 55.50€ |
| GPU Server | 2 109.80€ |
| Google Drive 100GB | 10.00€ |
| Cloud | 300.00€ |
| Total | 3 965 € |

Table 1.4: Final project cost estimate.

Chapter 2

Theoretical Background

2.1 Machine Learning and Deep Learning

2.1.1 Machine Learning

Machine Learning (ML) [56] is one of the core branches of Artificial Intelligence (AI). It enables computers to learn from data without being explicitly programmed. Through algorithms and models, computers can recognize patterns, make predictions, and make decisions based on the information provided.

ML is particularly valuable for solving complex problems for which no analytical solution exists, or where such solutions are very expensive to obtain. In these cases, the computer is responsible for identifying patterns in the data and making predictions about them [70].

Formally, a program is said to learn from experience E with respect to some class of tasks T and a performance measure P , if its performance on tasks T , measured by P , improves with experience E [56].

Depending on the problem requirements, the nature of the data, and the objectives to be achieved, different types of learning algorithms can be applied. In this thesis, we focus on two main categories:

- **Supervised learning:** the model is trained on annotated data, i.e., examples with known outputs.
- **Unsupervised learning:** the model is given unlabelled data and must discover the underlying patterns.

Figure 2.1 indicates the difference between supervised learning and unsupervised learning. Unsupervised learning can deal with some tasks more easily than supervised learning because it does not need the guidance of annotation in data. However, the research of unsupervised learning is in its infancy, and it cannot replace supervised learning algorithms.

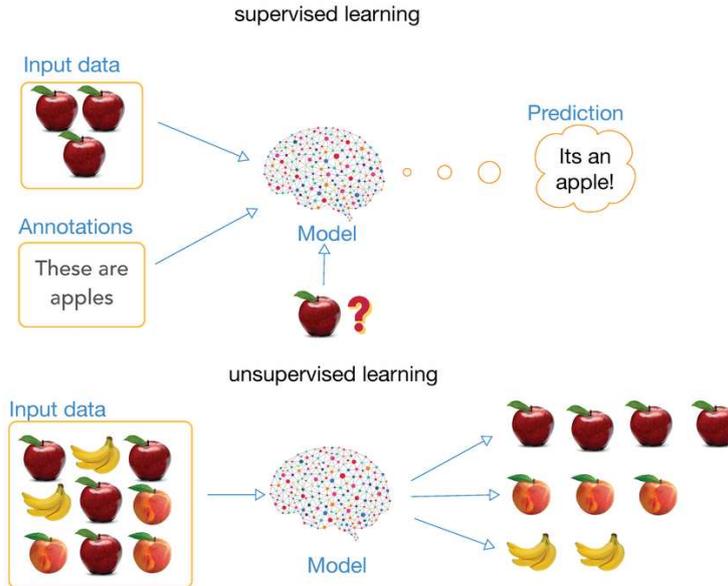


Figure 2.1: Supervised learning and unsupervised learning. Supervised learning uses annotation guidance to draw learning-task-related conclusions about the data. Unsupervised learning uses the latent factors in data to conclude the relationship between data and the corresponding learning task, and there is no need to mark the data [78]

In general, ML techniques are applied to large datasets, from which we aim to extract the hidden structures and patterns [28].

Given these descriptions, the problem addressed in this work can be approached using ML techniques. We have input data (additional features of distorted point clouds) and outputs (quality values). Furthermore, public datasets exist with labels for different types of distortions. Thus, this is a supervised learning problem.

2.1.2 Deep Learning

Deep Learning (DL) is a subfield of ML in which feature extraction is performed automatically by the model itself, rather than being manually engineered by a human expert [28, 48]. In practice, automatic feature extraction often outperforms handcrafted features.

Most DL models are based on hierarchical data processing through multiple layers. The best-known models are *artificial neural networks (ANNs)*, bio-inspired architectures that abstractly simulate the functioning of neurons in the human brain [69].

At a high level, a neural network consists of three main stages:

- **Input stage:** the model receives a set of data or features to analyze. The data are propagated forward through the network (feed-forward).
- **Processing stage:** each neuron computes a weighted sum of its inputs, applies an activation function, and passes the result to the next layer. The

weights encode the relative importance of the inputs, while the activation functions (e.g., sigmoid, ReLU) determine the neuron's output. As data propagate, hidden layers extract increasingly abstract features.

- **Output stage:** the network produces a prediction based on the extracted features. A loss function then computes the error with respect to the expected output, and the weights are adjusted accordingly.

The learning process, known as *training*, consists of iteratively repeating this procedure over many examples. The quality of the dataset is crucial: it must be representative, extensive, and clean, since the model will learn features directly from it.

In short, a neural network is a set of parameters (weights, biases, activation functions, loss functions, and optimizers) tuned to achieve the desired task. A common challenge is *overfitting*, which occurs when the model learns the training data too well, losing its ability to generalize to unseen data. To mitigate overfitting, techniques such as *regularization* are applied during training.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [47, 48] are a neural network architecture specifically designed for processing structured data, such as images. CNNs have also been successfully applied to text, audio, and more recently, 3D data. Their key idea is the use of *convolutional layers*, which apply learnable filters over local regions of the input to extract features.

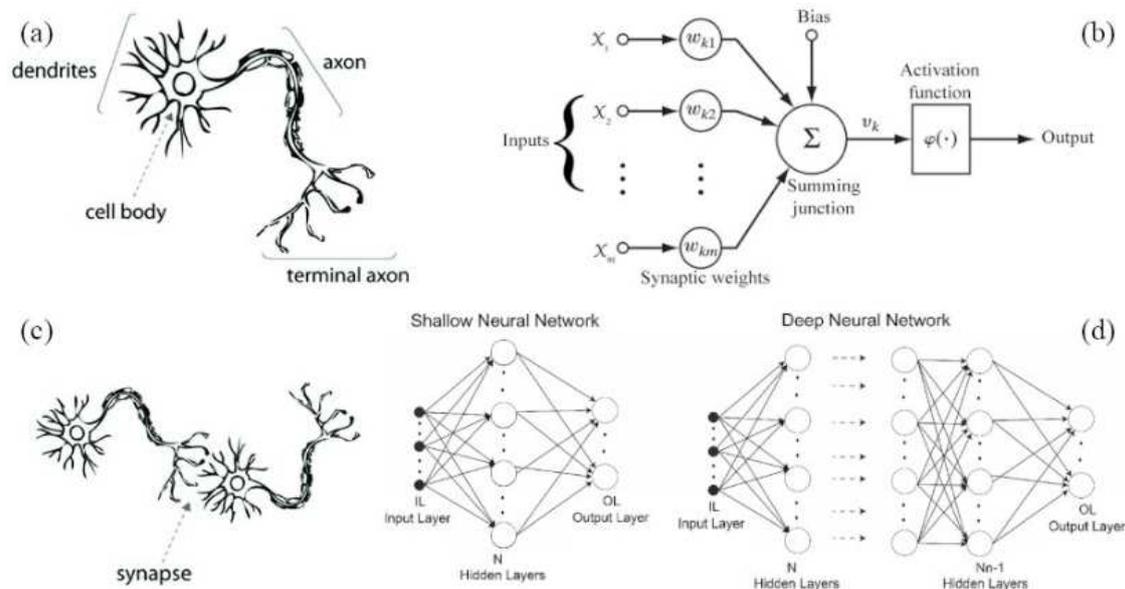


Figure 2.2: Graphical example of a neural network. (a) and (b) show a biological and an artificial neuron, respectively. (c) illustrates a synapse. (d) shows a shallow neural network (left) and a deep neural network (right).

To explain CNNs, consider the 2D image case. A convolution operation involves sliding a filter (kernel) across the image, computing element-wise products with the local patch, and summing the results to form an entry in the *feature map*. This operation is repeated across the image, capturing spatial patterns. The stride determines how the filter moves across the input, and padding may be applied to preserve dimensionality.

Pooling layers are often included to reduce the spatial size of feature maps, thereby controlling the number of parameters and computation. Combined with activation functions and fully connected layers, CNNs provide a powerful architecture for learning hierarchical feature representations.

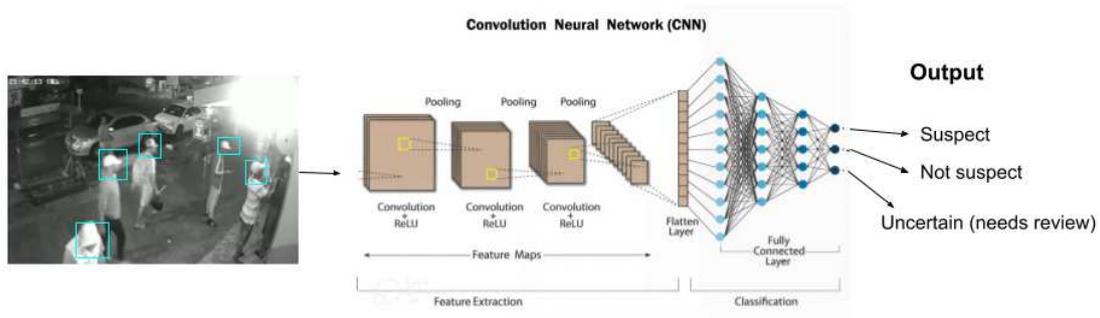


Figure 2.3: CNN applied to a biomedical image classification problem [47]. The main building blocks of CNNs are convolutional layers, pooling layers, activation functions, and fully connected layers.

2.2 Face Detection

Face detection is a crucial step in computer vision and biometric applications, acting as a foundation for tasks such as facial recognition, emotion detection, and IQA. Broadly speaking, face detection falls under the category of object detection, which aims to identify and localize objects within an image. The road map of traditional and deep learning methods shown in the Figure 2.4.

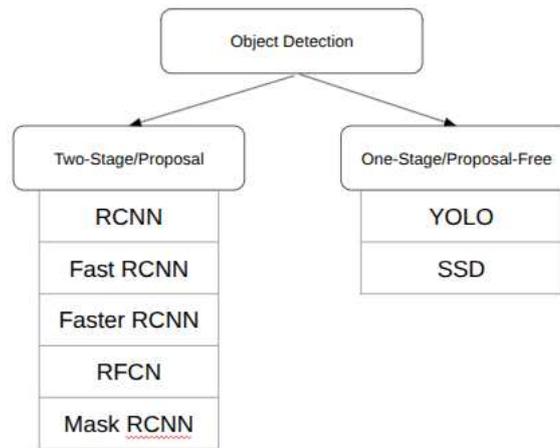


Figure 2.4: Overview of object detection methods categorized into two-stage and one-stage approaches. Two-stage proposal-based detectors, such as RCNN [27], Fast RCNN [26], Faster RCNN [68], R-FCN [16], and Mask RCNN [32], first generate region proposals and subsequently classify them. In contrast, one-stage proposal-free detectors, such as YOLO [67] and SSD [52], directly predict bounding boxes and object categories in a single step, enabling faster inference.

Object detection algorithms can be categorized into two main paradigms:

- **Two-stage detectors:** These methods, such as R-CNN, Fast R-CNN, and Faster R-CNN, first generate region proposals and then classify those regions. They generally offer high accuracy but are computationally intensive.
- **One-stage detectors:** Methods like YOLO and SSD predict object bounding boxes and class probabilities directly from full images in a single forward pass. These models are faster and suitable for real-time applications.

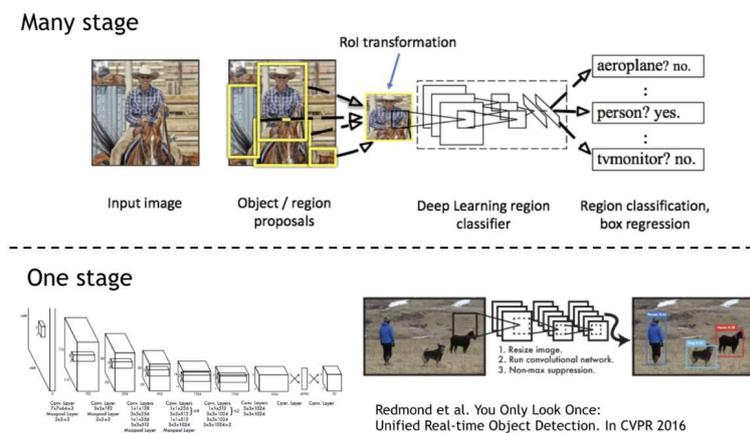


Figure 2.5: Comparison between one-stage and two-stage object detectors. One-stage models predict directly from the image, while two-stage models involve a region proposal step [63].

Among the one-stage object detection methods, two prominent architectures—YOLO and SSD—stand out for their balance of speed and accuracy. The following sections provide a detailed overview of each, with emphasis on their relevance to face detection and then image quality assessment tasks.

2.2.1 YOLO

YOLO (You Only Look Once)[redmon2016you] is a family of real-time object detection algorithms that are widely used due to their high speed and competitive accuracy. Traditional object detectors, such as R-CNN or Faster R-CNN, rely on multi-stage pipelines: first proposing regions of interest (ROIs), and then classifying them. In contrast, YOLO performs both tasks simultaneously, allowing it to achieve high speed and competitive accuracy. YOLO divides the input image into an $S \times S$ grid. Each grid cell predicts a fixed number of bounding boxes, along with confidence scores and class probabilities. These predictions are output from a single forward pass of a CNN, enabling real-time object detection even on resource-constrained hardware.

Figure 2.6 shows a simplified schematic of the YOLO detection pipeline.

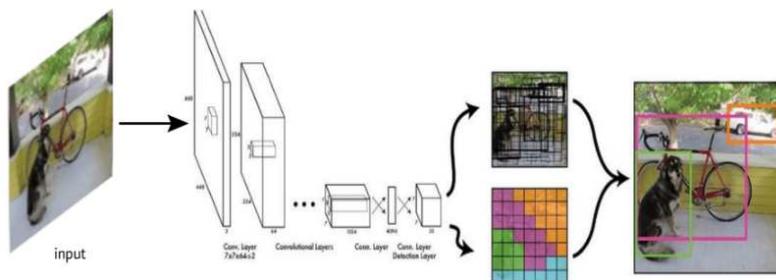
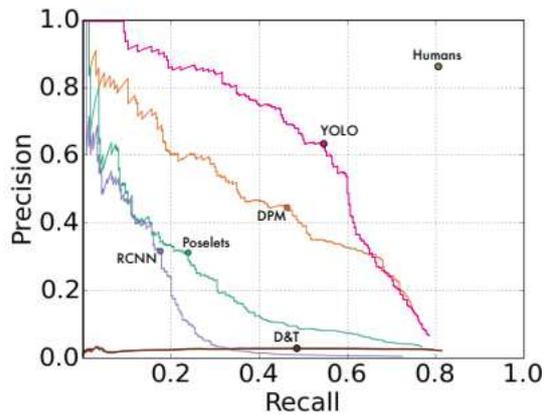


Figure 2.6: Schematic diagram of the YOLO object detection pipeline. The input image is processed through a CNN to produce bounding boxes and class confidence scores in a single pass.

In the official paper of YOLO [redmon2016you], they compare YOLO to other detection systems on the Picasso Dataset [25] and the People-Art Dataset [14], two datasets for testing person detection on artwork. Figure 2.7 shows comparative performance between YOLO and other detection methods.



(a) Picasso Dataset precision-recall curves.

| | VOC 2007 | Picasso | | People-Art |
|--------------|-------------|-------------|--------------|------------|
| | AP | AP | Best F_1 | AP |
| YOLO | 59.2 | 53.3 | 0.590 | 45 |
| R-CNN | 54.2 | 10.4 | 0.226 | 26 |
| DPM | 43.2 | 37.8 | 0.458 | 32 |
| Poselets [2] | 36.5 | 17.8 | 0.271 | |
| D&T [4] | - | 1.9 | 0.051 | |

(b) Quantitative results on the VOC 2007, Picasso, and People-Art Datasets. The Picasso Dataset evaluates on both AP and best F1 score.

Figure 2.7: Figure adapted from [redmon2016you], Comparison of precision–recall performance for different object detection models, including YOLO, RCNN, Fast RCNN, DPM, and others, on benchmark datasets. YOLO demonstrates a favorable trade-off between precision and recall compared to other detectors, while humans still outperform all models in this metric. This comparison highlights YOLO’s efficiency in real-time detection scenarios.

In the context of *FIQA*, precise and consistent face detection is a crucial preprocessing step. YOLO-based detectors such as YOLOv5 or YOLOv7 are widely used due to their ability to quickly and accurately localize faces, even under difficult conditions—such as low resolution, motion blur, partial occlusion, or poor lighting—all of which are common in forensic and surveillance imagery.

Machine Learning and Deep Learning Techniques Used in YOLO

YOLO leverages a combination of advanced ML and DL techniques to achieve real-time object detection with high accuracy. At its core, YOLO employs deep CNNs, an architecture particularly suited for extracting hierarchical feature representations directly from raw image data.

Deep Learning Backbone: CNN Architecture

YOLO’s primary strength originates from its deep CNN architecture. CNNs are neural network architectures specifically designed for processing grid-like data such as images. They consist of multiple layers, each specialized in capturing distinct types of visual features. In YOLO:

- **Convolutional Layers:** These layers apply convolution operations to input images or feature maps to detect local visual features such as edges, textures, and patterns. Multiple convolutional layers, stacked sequentially, allow the model to extract increasingly complex and abstract visual representations.
- **Pooling Layers:** Typically, max-pooling layers are interleaved with convolutional layers. Pooling operations reduce spatial dimensions of feature maps,

enhancing computational efficiency and introducing translational invariance into the network.

- **Activation Functions (ReLU):** Non-linear activation functions like the Rectified Linear Unit (ReLU) enable CNNs to model complex non-linear relationships within data, significantly enhancing representation capability.

This deep CNN structure provides YOLO with an effective hierarchical representation of image data, allowing it to robustly recognize objects across various scales and contexts.

Prediction as a Regression Task

Unlike traditional region-based object detection methods (e.g., R-CNN, Fast R-CNN) that require separate region proposals and classification steps, YOLO frames detection as a unified regression problem. This means YOLO directly predicts:

1. **Bounding Box Coordinates (x, y, width, height):** Using regression outputs from the final CNN layers, YOLO estimates exact bounding box positions and dimensions directly from learned image features.
2. **Confidence Scores:** The model simultaneously predicts a confidence score that reflects both the presence probability of an object and the accuracy of the predicted bounding box.
3. **Class Probabilities:** YOLO also outputs a probability distribution across object classes for each predicted bounding box, indicating the likelihood that a detected object belongs to a specific class.

This single-stage approach significantly reduces computational complexity, enabling YOLO to perform at real-time speeds while maintaining competitive accuracy levels.

Loss Functions and Optimization

YOLO uses a carefully designed loss function to train its network. The total YOLO loss function combines multiple components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{class}} \quad (2.1)$$

where:

- $\mathcal{L}_{\text{coord}}$ penalizes inaccuracies in predicted bounding box coordinates using mean squared error.
- $\mathcal{L}_{\text{conf}}$ handles incorrect confidence scores, distinguishing between cells containing and not containing objects.
- $\mathcal{L}_{\text{class}}$ manages misclassification errors, computed using categorical cross-entropy loss.

Training YOLO involves optimizing this loss function via backpropagation, typically using gradient descent optimization algorithms like stochastic gradient descent or adaptive variants such as Adam.

Non-Maximum Suppression

During inference, YOLO employs Non-Maximum Suppression (NMS), an essential ML-based post-processing step to remove redundant bounding boxes. NMS sorts the predicted bounding boxes based on their confidence scores, then iteratively selects the highest-confidence box while discarding overlapping boxes that exceed a defined Intersection over Union (IoU) threshold. This significantly refines the output detections and enhances final accuracy.

Transfer Learning and Fine-tuning

Typically, YOLO models are pre-trained on extensive datasets such as ImageNet or COCO to learn generic visual features. Subsequently, these models can be fine-tuned on smaller, domain-specific datasets to enhance detection performance for specific tasks—such as facial detection in biometric and forensic contexts—leveraging transfer learning to improve generalization and robustness.

By integrating deep convolutional architectures, regression-based prediction, sophisticated loss functions, and careful post-processing techniques, YOLO effectively combines ML and DL methodologies. This integration allows it to robustly handle real-world complexities, such as varying scales, occlusions, and diverse image quality conditions—making YOLO an ideal choice for integration into forensic facial image quality assessment frameworks like OFIQ..

2.2.2 Single Shot MultiBox Detector

The Single Shot MultiBox Detector (SSD) is a widely-used one-stage object detection algorithm introduced by Liu et al. [52]. Unlike two-stage detectors that rely on region proposals, SSD performs object localization and classification in a single pass through the network, making it highly efficient for real-time applications. In Figure 2.8, the SSD framework is shown.

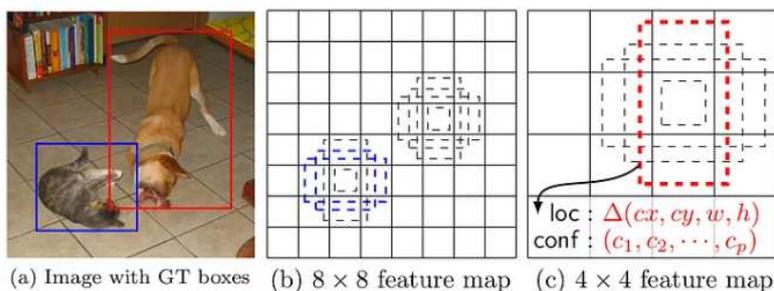


Figure 2.8: SSD framework. (a) SSD only needs an input image and ground truth boxes for each object during training. In a convolutional fashion, we evaluate a small set (e.g. 4) of default boxes of different aspect ratios at each location in several feature maps with different scales (e.g. 8×8 and 4×4 in (b) and (c)). For each default box, we predict both the shape offsets and the confidences for all object categories ((c_1, c_2, \dots, c_p)). At training time, we first match these default boxes to the ground truth boxes. For example, we have matched two default boxes with the cat and one with the dog, which are treated as positives and the rest as negatives. The model loss is a weighted sum between localization loss (e.g. Smooth L1 [26]) and confidence loss (e.g. Softmax) [51]

SSD’s distinguishing feature is its use of multiple feature maps from different layers in the convolutional backbone. These maps, representing progressively lower spatial resolutions, allow SSD to detect objects of various scales more effectively. Lower layers capture finer details for small objects, while deeper layers are better suited for larger-scale detections [23, 37].

SSD in Face Detection Quality Assessments

In the context of facial image quality assessment (FIQA), SSD has been adopted in frameworks like the original OFIQ [39], where it serves as the primary face detector. Its advantages in this setting include:

- **Efficient detection of small and medium-sized faces:** SSD’s multi-scale design enables it to detect faces of varying sizes across different image resolutions [80].
- **Real-time performance:** SSD is optimized for speed, achieving frame rates suitable for real-time biometric and surveillance systems [37].
- **High accuracy in controlled environments:** SSD performs reliably in scenarios with consistent lighting, frontal poses, and minimal occlusion, which are common in regulated biometric acquisitions [75].

However, SSD is less robust in unconstrained environments. It exhibits performance degradation when applied to forensic-like images, which often involve low resolution, motion blur, severe occlusion, or lateral face views [50]. These limitations reduce its effectiveness in forensic workflows, where facial detection must operate under suboptimal conditions.

2.2.3 SSD vs YOLO

| Metric | YOLO | SSD |
|---------------------------|------------------|------------------|
| Speed (FPS) | High (up to 140) | Moderate (30–60) |
| Detection Accuracy | High | Moderate |
| Occlusion Handling | Good | Fair |
| Multi-scale Detection | Good | Very Good |
| Suitability for Forensics | Excellent | Moderate |

Table 2.1: Comparison of YOLO and SSD for face detection. The values are derived from published benchmark studies and comparative analyses. YOLO demonstrates higher speed (up to 67 FPS) and stronger detection accuracy, particularly in real-time scenarios [66, 37], while SSD achieves competitive performance with robust multi-scale detection based on feature maps [51]. Occlusion handling has been explored in extensions of YOLO for partially visible objects [31], whereas SSD has been evaluated under occlusion and illumination variations in urban detection tasks [58]. The assessment of forensic suitability is inferred from the literature, given YOLO’s emphasis on speed and applicability to surveillance contexts.

YOLO provides better generalization for unconstrained environments and outperforms SSD in speed and occlusion robustness. Consequently, it has become the preferred choice for preprocessing in forensic face quality analysis. Integrating YOLO into the OFIQ pipeline ensures that quality metrics are computed only on the relevant facial region, free from background clutter or misalignment. This alignment is vital for assessing local quality components like sharpness, illumination, and facial symmetry. Furthermore, using a real-time detector like YOLO allows the system to scale to large datasets or operate under time constraints, which is especially valuable in biometric authentication or forensic triage workflows.

2.3 Image Quality Assessment

IQA encompasses a broad set of algorithms and methodologies aimed at quantifying perceptual quality, which is crucial for tasks like diagnostic imaging, surveillance, photography, and biometric recognition [13, 57, 76]. IQA seeks to measure how distortions such as blur, noise, compression, and illumination affect an image’s utility for human or machine interpretation. IQA methods are generally divided into three categories, illustrated in Figure 2.9):

- **Full-Reference IQA (FR-IQA):** Requires an undistorted reference image to compare with the test image. Common examples include SSIM [76] and PSNR.
- **Reduced-Reference IQA (RR-IQA):** Uses a partial set of reference features or metadata to estimate image quality.

- **No-Reference IQA (NR-IQA):** Assesses quality without any reference, relying on intrinsic image characteristics, often learned from data [13, 57].

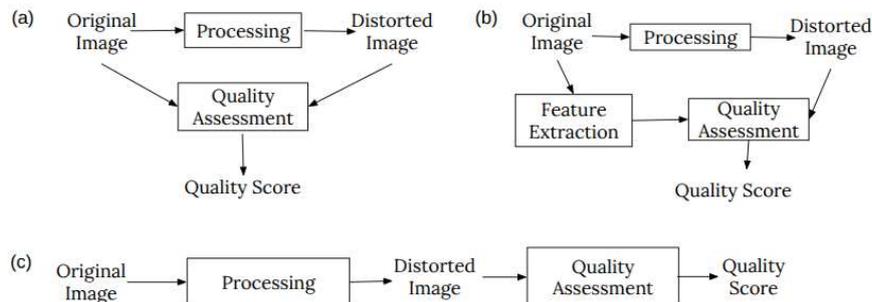


Figure 2.9: Summary of IQA subproblems. Image (a) shows the FR-IQA setup; (b) shows RR-IQA; and (c) represents NR-IQA. The proposed work in this thesis falls under category (c), as it evaluates facial image quality in forensic contexts where no reference images are available.

2.4 Face Image Quality Assessment

FIQA specifically refers to estimating the quality of face images in the context of automated facial recognition. This is especially important in biometric systems such as border control, surveillance and eKYC¹ system where low-quality images can lead to high false match or false non-match rates [8, 30, 61].

Most FIQA methods produce a scalar quality score representing how suitable an image is for recognition tasks. Recent methods also produce interpretable sub-scores [34, 55, 74]. Standards such as ISO/IEC 19794-5:(2005, 2011) and 39794-5:2019 [41] and ICAO guidelines [40] define technical criteria for acceptable facial images, covering:

- **Illumination:** Balanced lighting without shadows or hotspots.
- **Sharpness:** The face must be in focus, with minimal blur.
- **Pose:** Near-frontal views are preferred for accurate recognition.
- **Occlusion:** Key facial regions (e.g., eyes, mouth) should not be blocked.

For example, European regulation states that the EES (Entry Exit System) facial quality algorithm shall be comprehensible in terms of the ISO/IEC 19794-5:2011 criteria.

In Figure 2.10 some of these criteria are shown.

¹eKYC (electronic Know Your Customer) refers to digital identity verification processes used by organizations such as banks and telecom providers to remotely confirm a person's identity.

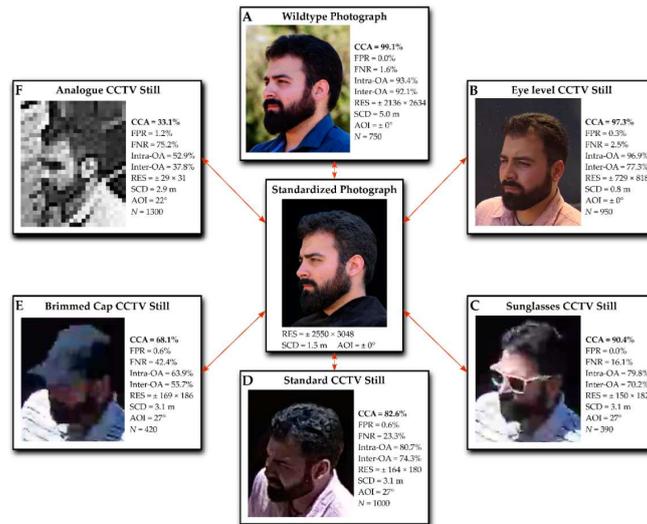


Figure 2.10: Summary of facial image validation using morphological analysis across six different capture conditions from the Wits Face Database [3]. The figure illustrates how various image sources—ranging from standardized photographs to analogue CCTV stills—impact recognition performance based on metrics such as Corrected Classification Accuracy, False Positive/Negative Rates, and resolution. Each capture condition is compared to the central standardized photograph, demonstrating how image quality factors like pose, illumination, and occlusion influence forensic face comparison reliability.

Early FIQA methods were rule-based, using handcrafted features like histogram distributions for lighting or Laplacian operators for focus estimation. These methods offered limited generalizability and interpretability [30]. Modern FIQA frameworks leverage deep learning models trained to correlate image content with downstream recognition scores, improving robustness and generalization [34, 55, 74].

In summary, FIQA is a crucial pre-processing step in biometric systems, helping filter low-quality images before they enter recognition pipelines.

Chapter 3

State of the art

The assessment of face image quality has become increasingly critical in the realm of biometric and forensic applications. As facial recognition systems are deployed in diverse and often uncontrolled environments, ensuring the reliability and accuracy of these systems hinges on the quality of the input images. Factors such as pose variation, illumination changes, occlusions, and image resolution significantly impact the performance of facial recognition algorithms. Consequently, the field has witnessed a surge in research aimed at developing robust FIQA techniques that can predict and enhance the utility of facial images in recognition tasks.

3.1 State of the art of FIQA

Over the past decade, FIQA has become a critical subfield in biometrics and computer vision. Its goal is to predict whether a given face image is suitable for use in automated face recognition systems. FIQA methods have progressed from simple heuristic models to highly complex deep learning-based systems that evaluate quality in both supervised and unsupervised manners.

To conclude this section, it is pertinent to illustrate the growing academic interest in the field of Facial Image Quality Assessment (FIQA) within the broader context of facial analysis. As shown in 3.1, the number of publications indexed in Scopus related to FIQA has steadily increased over the past decade. This trend is particularly evident in reputable journals such as IEEE Access, PLOS ONE, and Multimedia Tools and Applications, highlighting a sustained and rising focus on this research area. The notable surge in recent years underscores the relevance and timeliness of the topic addressed in this thesis.

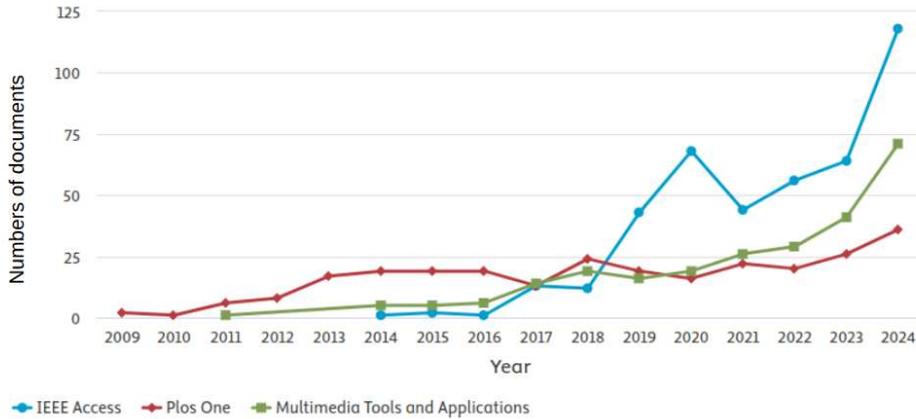


Figure 3.1: Annual number of Scopus-indexed publications from 2009 to 2024 related to Facial Image Quality Assessment (FIQA), retrieved on June 14, 2025, using the query: ("facial image quality assessment" OR "face image quality"). The graph compares results across three major journals: IEEE Access, Plos One, and Multimedia Tools and Applications. A clear upward trend is observed, reflecting a growing research interest in FIQA for applications in biometric recognition and forensic image analysis.

Early approaches to FIQA predominantly relied on handcrafted features and statistical models. These methods utilized metrics such as signal-to-noise ratio, contrast, and sharpness to estimate image quality. However, with the advent of deep learning, there has been a paradigm shift towards data-driven models that can learn complex representations of image quality. Recent surveys highlight this transition, noting the emergence of deep learning-based FIQA methods that outperform traditional techniques, especially in unconstrained scenarios [18].

One notable advancement is the development of the Detectability (DET) score, a novel metric that quantifies the suitability of an image for face detection tasks. The DET score evaluates the performance of face detectors across varying IoU thresholds, providing a comprehensive measure of image quality in the context of detection performance [49]. Building upon this, supervised and unsupervised quality estimators have been proposed, leveraging the DET score to train models that can predict image quality without explicit ground truth labels.

3.2 FIQA in Forensic Applications

In forensic settings, the quality of facial images is paramount, as it directly influences the accuracy of facial comparisons and identifications. Studies have demonstrated that higher image quality correlates with increased correctness in facial comparisons, emphasizing the need for reliable FIQA methods in forensic workflows [54]. Moreover, the lack of standardized image quality assurance systems in forensic facial comparison underscores the importance of developing robust FIQA frameworks tailored to forensic requirements.

Recent research has explored semi-quantitative scoring methods to assess image quality, focusing on factors such as resolution and lighting. By analyzing facial

images from databases like the Wits Face Database, researchers have established correlations between image quality scores and the accuracy of facial comparisons, reinforcing the critical role of FIQA in forensic investigations [4].

3.3 Integration of YOLO in FIQA Pipelines

The integration of advanced object detection algorithms, particularly the YOLO framework, has significantly enhanced face detection capabilities in FIQA pipelines. YOLO’s real-time detection performance and high accuracy make it a suitable choice for processing facial images in forensic contexts. Studies have demonstrated that YOLO, especially its latest iterations like YOLOv8, outperforms previous detectors in terms of precision and recall, even in challenging conditions [1].

By incorporating YOLO into FIQA systems, researchers have achieved improved detection rates, particularly in images with oblique angles, poor lighting, and occlusions. This enhancement not only increases the number of usable images but also improves the accuracy of subsequent quality assessments and facial comparisons.

3.4 Taxonomy of FIQA Methods

Based on their output format and technical approach, FIQA methods can be broadly categorized into four families.

Single-score prediction methods aim to output a unified scalar value that reflects the overall quality of a face image. These approaches are widely used because they provide a direct numerical indicator that can be incorporated into recognition pipelines or used for thresholding decisions [9, 35].

In contrast, **multi-component estimation methods** decompose the quality assessment task into several dimensions, such as pose, illumination, or sharpness. By producing independent quality scores for these factors, they allow a more fine-grained understanding of how each attribute contributes to recognition performance [15, 45].

FIQA methods can also be distinguished by their **learning paradigm**. In **supervised approaches**, models are trained using labeled performance outcomes, for example, recognition accuracy or human-annotated quality ratings. **Unsupervised or self-supervised approaches**, on the other hand, rely on intrinsic signals or proxy tasks to learn quality representations without requiring explicit quality labels [33, 74].

Finally, a practical distinction concerns whether FIQA is implemented as an **integrated** component or as a **standalone** module. In the integrated design, quality estimation is embedded directly into the face recognition pipeline, sharing features and parameters with the recognition model itself. Standalone approaches, in contrast, operate as a separate preprocessing step that can be applied independently of the recognition system [17, 19].

The summary of Taxonomy of FIQA Methods shown in the figure 3.2.

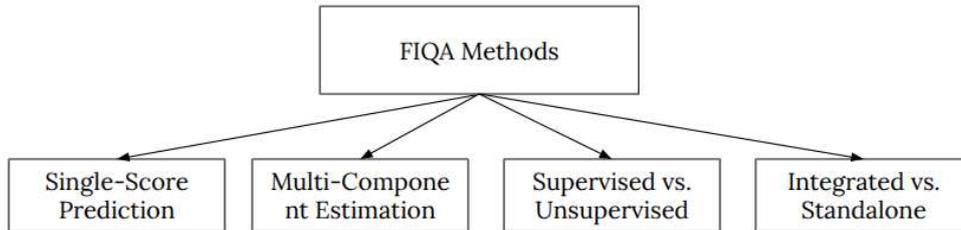


Figure 3.2: Taxonomy of FIQA methods based on output format and learning approach. The main categories include: (1) Single-Score Prediction, which outputs an overall quality score; (2) Multi-Component Estimation, which provides separate quality scores for specific image aspects (e.g., pose, illumination); (3) Supervised vs. Unsupervised methods, depending on whether performance labels are used during training. This taxonomy highlights key distinctions in how FIQA methods are structured and evaluated.

3.4.1 Single-Score Prediction Methods

Single-score prediction methods form the backbone of many FIQA systems. These models typically rely on deep CNNs and are trained to regress a scalar value representing the expected recognition performance of a given face image. These scalar quality scores often correlate with downstream biometric metrics such as face verification accuracy or embedding distance. This section highlights key single-score FIQA models: FaceQnet, MagFace, AdaFace, and SER-FIQ.

FaceQnet

FaceQnet [35] is one of the earliest supervised FIQA methods. It fine-tunes a pre-trained ResNet-50 model using similarity scores from a commercial face recognition system as soft labels. The core idea is to estimate a regression target that reflects the match quality of a face image compared to a high-quality reference. Due to its simplicity, transparency, and benchmark value, FaceQnet is widely used in academic studies for comparative evaluations. However, its reliance on external systems for label generation introduces a dependency and limits generalizability across diverse operational contexts. Its scalar output offers low interpretability regarding specific quality issues like pose, occlusion, or lighting.

MagFace and AdaFace

MagFace [55] and AdaFace [44] represent a newer generation of FIQA models that embed quality awareness directly into face representation learning. Unlike FaceQnet, they do not require post-hoc quality score regression.

MagFace incorporates a magnitude-aware regularization mechanism during training, where the norm of the face embedding vector implicitly encodes image quality. High-quality images tend to generate embeddings with larger magnitudes, naturally separating them from lower-quality samples. This implicit encoding allows quality estimation to be tightly coupled with recognition features.

AdaFace builds on this idea by dynamically adjusting training margins based on an input’s estimated quality. It introduces a quality-adaptive margin loss that enables robust recognition performance even under varied image conditions. These models enhance generalization and offer competitive interpretability without requiring explicit supervision for quality.

SER-FIQ

SER-FIQ (Stochastic Embedding Robustness for Face Image Quality) [74] takes a fundamentally different approach by being fully unsupervised. Instead of relying on labels or handcrafted quality measures, SER-FIQ measures the variability of face embeddings produced by a recognition model under random dropout conditions. The core intuition is that high-quality images result in consistent embeddings, whereas poor-quality inputs lead to greater variance.

This stochastic robustness score correlates well with recognition accuracy and can be computed on-the-fly, making SER-FIQ attractive for real-time applications. However, as an unsupervised method, its performance can vary depending on the backbone model used and the distribution of dropout-induced noise.

| Model | Supervision | Architecture | Score Type | Interpretability |
|----------|--------------|----------------------|----------------|------------------|
| FaceQnet | Supervised | ResNet | Scalar | Low |
| MagFace | Implicit | Custom CNN | Embedding Norm | Medium |
| AdaFace | Implicit | Adaptive Margin Loss | Embedding Norm | Medium |
| SER-FIQ | Unsupervised | Dropout-based CNN | Variance Score | Medium |

Table 3.1: Comparison of representative single-score FIQA models with respect to supervision strategy, network architecture, score type, and interpretability. The interpretability column refers to the extent to which the quality score can be traced back to specific design choices or input characteristics. FaceQnet [35], which directly predicts a scalar quality value, provides limited insight into the underlying factors affecting quality and is thus rated as “Low”. In contrast, MagFace [55] and AdaFace [44] derive quality measures from embedding norms, while SER-FIQ [74] uses variance estimates across stochastic embeddings. These methods provide a more transparent link between the model’s decision and representation robustness, and are therefore rated as “Medium”.

Each of these models represents a different balance between supervision, architecture complexity, interpretability, and real-world applicability. Selecting the appropriate model depends on the target domain (e.g., forensic, mobile, surveillance), required speed, and availability of training labels.

3.4.2 Component-Based Estimation: OFIQ

OFIQ [39] represents a new generation of interpretable FIQA systems. Instead of a single score, it outputs multiple component scores that reflect factors such as:

- **Capture-Related:** lighting, sharpness, background complexity.

- **Subject-Related:** eye visibility, expression, occlusion, pose.

This modular architecture is particularly useful in operational environments, where identifying the specific quality deficiency (e.g., “eye not visible”) helps guide corrective action.

Many methods are evaluated on datasets like:

- **LFW, VGGFace2, CelebA** – For general face data.
- **BioSecure, FRGC** – For biometric scenarios.
- **Private surveillance datasets** – For low-quality image assessment.

Performance is usually reported using correlation with match scores (e.g., PLCC, SROCC), or by comparing decision thresholds against face recognition accuracy.

Overall, the field of FIQA is moving toward interpretable, task-aware systems that not only evaluate image quality, but also explain it. While single-score methods still dominate academic literature, component-based methods are better aligned with the needs of forensic, biometric, and border control workflows. As deep learning models continue to evolve, we anticipate hybrid systems that combine the precision of single-score prediction with the transparency of multi-factor analysis.

3.4.3 Machine Learning and Deep Learning Approaches for FIQA

Traditional FIQA relied on hand-crafted features and rule-based algorithms, such as histogram analysis for brightness or Laplacian filters for sharpness. However, these methods struggled to generalize to uncontrolled, real-world conditions. To address these limitations, researchers have shifted toward data-driven approaches using ML and DL, which can learn complex patterns from annotated datasets.

Deep Learning-Based FIQA

DL has revolutionized FIQA by enabling models to directly learn quality-relevant features from raw images. Instead of relying on handcrafted indicators, modern approaches employ CNNs to build no-reference quality models that predict image utility on the basis of empirical performance data. Several representative methods illustrate the diversity of this line of research.

One of the earliest deep learning approaches is **FaceQnet** [35], a CNN-based model trained using face recognition similarity scores as ground truth. It produces a unified scalar quality score from a single image, without requiring reference images or pairwise comparisons. In contrast, **SER-FIQ** [74] follows an unsupervised strategy, estimating quality from the stability of facial embeddings under dropout noise. The rationale is that high-quality images generate consistent embeddings, whereas noisy or occluded samples lead to higher variance.

Other methods integrate quality assessment more tightly into the recognition process itself. For instance, **MagFace** [55] incorporates quality-awareness into the

face recognition model by regularizing embedding norms during training. As a result, embeddings from high-quality images have larger magnitudes, making quality implicitly measurable. Similarly, **AdaFace** [44] adapts the decision margin dynamically during training, so that the embedding norm can be directly interpreted as an indicator of image quality.

More recently, **eDiffFIQA** [60] has been proposed as an alternative, producing single quality scores through deep learning. However, its interpretability remains limited, as it does not explicitly analyze individual quality factors but instead provides a global estimate.

Overall, deep learning-based FIQA methods share several advantages. They are able to adapt to a wide range of distortions, including blur, occlusion, and pose variations. They also tend to correlate more strongly with actual face recognition performance compared to handcrafted quality indicators. Furthermore, by providing continuous and explainable quality signals, they support practical uses such as filtering poor-quality samples or weighting recognition decisions according to image reliability.

The transition from handcrafted to learned FIQA methods reflects the increasing demand for robust and adaptable solutions in unconstrained biometric scenarios, particularly in forensic applications or surveillance footage. Nevertheless, while many modern FIQA systems rely on deep learning, the OFIQ framework explored in this thesis adopts a more traditional machine learning approach for certain quality components. Specifically, it focuses on estimating *sharpness* using handcrafted features combined with classical classifiers, in order to provide reliable quality estimates in the absence of reference images.

Chapter 4

Materials and Methods

This chapter presents the datasets, methodologies, and evaluation criteria employed to enhance the OFIQ framework for forensic facial comparison. The discussion begins with an overview of the datasets, with particular emphasis on the generic dataset used in this study. Subsequently, the methods are described in detail, highlighting the role of OFIQ for image quality assessment and the integration of YOLO as the updated face detection model. Finally, the evaluation strategy is outlined, including the use of annotated data and similarity metrics for performance measurement.

4.1 Dataset

The dataset employed in this study comprises a total of 3,632 images, collected during my research internship at *Panacea Cooperative Research*¹. Importantly, this dataset was not only provided by the company but also acquired in active cooperation with them, meaning that I directly contributed to its construction as part of my research activities. This represents a distinctive value of the present thesis, since it incorporates data that I personally helped to acquire rather than relying exclusively on publicly available benchmarks.

To construct the dataset, we first recorded videos with each of the surveillance cameras and subsequently extracted frames at intervals of 0.5 seconds. This approach ensured both temporal diversity and variability in facial appearance, while avoiding redundant frames.

In order to further approximate realistic forensic scenarios, we deliberately introduced a variety of challenging conditions using several tools and situations, such as individuals wearing caps or glasses, lateral facial poses, partially occluded faces, and even blindfolds. These factors were included to replicate the types of distortions and occlusions commonly encountered in surveillance and forensic imagery.

The images were captured using three different types of surveillance cameras commonly deployed in public and semi-public spaces:

- **Camera #1 (ATM Camera):** Captured 1,420 images, with a resolution of 1440×2560 pixels. This camera simulates typical bank ATM installations, providing mostly frontal or slightly downward-angled views under challenging lighting.

¹The dataset remains the property of Panacea Cooperative Research and was used exclusively for research purposes during the internship. It is not publicly available.

- **Camera #2 (Dome Camera):** Contributed 1,200 images, recorded at 2160×3840 pixels. Dome cameras are ceiling-mounted and often used in indoor surveillance settings, offering wide-angle views but variable facial positioning.

- **Camera #3 (Bullet Camera):** Provided 1,012 images, with a resolution of 1520×2688 pixels. These cameras typically cover outdoor environments, capturing faces in uncontrolled lighting and occlusion conditions.

Overall, the dataset reflects a realistic mix of acquisition scenarios that are directly relevant to forensic applications. By combining multiple camera types, temporal sampling from videos, controlled introduction of occlusions and distortions, and the active cooperation with *Panacea Cooperative Research* during its collection, the dataset stands out as a distinctive and valuable contribution of this MSc thesis.

To better illustrate the acquisition setup and the perspectives associated with each camera type, Figure 4.1 shows sample images recorded from the three surveillance cameras.



Figure 4.1: Illustration of the viewpoints and environments associated with each surveillance camera type used in this study. Camera #1 (ATM Camera) captures primarily frontal or slightly downward-facing images under indoor, bank-like lighting. Camera #2 (Dome Camera) provides a wide-angle ceiling-mounted view, common in indoor surveillance. Camera #3 (Bullet Camera) captures oblique, outdoor scenes with variable lighting and occlusion. These diverse viewpoints reflect real-world forensic conditions and are used to build the generic image dataset.

The use of multiple camera types introduces substantial variation in lighting, angle, facial pose, sharpness, and occlusions. This variability is essential for the development and evaluation of face quality assessment systems intended for operational use in forensic applications, where image quality is rarely ideal.

In addition to algorithmic considerations, the technical setup and physical installation of the cameras play a crucial role in ensuring the consistency and quality of the captured face images. Proper installation not only improves the reliability of data collection but also ensures that the resulting images adhere to standards suitable for quality assessment tasks. For example, cameras must be mounted in a perfectly vertical orientation, with careful alignment based on spatial coordinates to maintain consistency across sessions and subjects.

An example of our installation setup at the University of Granada is shown in Figure 4.2, which illustrates the attention given to camera orientation, height, and distance to the subject.



Figure 4.2: Outdoor and indoor camera setup during data acquisition: The image shows the placement of a surveillance camera mounted on a tripod, along with other equipment used to simulate realistic forensic capture scenarios. The scene illustrates the process of setting up recording angles, stabilizing camera positions, and coordinating environmental conditions such as natural lighting and background clutter. These configurations aimed to mimic unconstrained, real-world forensic conditions.

In capturing the experimental video sequences, it was crucial to ensure accurate frame alignment and consistent perspective. To facilitate this, a checkerboard was initially employed, partially obscured by a red cover, to mark the exact starting point of usable footage. Upon removing this cover, subsequent frames were clearly identifiable as part of the primary recording sequence. This method provided an efficient and reliable reference point to normalize the scale and maintain consistent alignment across all captured frames, as visually illustrated in Figure 4.3.



Figure 4.3: The usage of checkerboard in Normalize scale and perspective of frames. holding a checkerboard with a red cover (left) and then removing the cover (right) to mark the start of the usable sequence. The camera system includes a dome-style surveillance camera mounted on a tripod, positioned to capture near-frontal views under controlled indoor lighting. This simple approach facilitates frame alignment and cropping in the absence of external timestamps. it plays the vital role in Normalize scale and perspective of frames

4.2 Dataset Usage and Partitioning

To validate the effectiveness of the proposed FIQA method, a subset of the dataset was curated with high-confidence annotations and used for evaluation purposes.

Preliminary results have shown that these selected images are representative of the challenges commonly faced in operational forensic workflows.

For further analysis, the dataset will be partitioned into distinct subsets to support both training and validation phases. The intended split will follow a conventional train-test protocol, ensuring that evaluation is performed on images unseen during model development. This approach aims to measure the generalization capability of the quality estimation model across varying capture conditions and device types.

4.3 Model: OFIQ

First we observe OFIQ in many aspects to understand how exactly it works so we are deep in the OFIQ code which is written in C++ and also OFIQ used for the Biometrics purpose so it is trained on the just the frontal faces but we should modify it to works for the forensic facial comparison that we have also the lateral/profile faces.

OFIQ framework is a NR-IQA model designed to evaluate the quality of facial images for biometric applications [39]. As a no-reference method, OFIQ operates without requiring a reference image, relying instead on intrinsic features extracted from each image to generate a unified quality score and a set of interpretable sub-scores.

Originally, OFIQ was developed for use in controlled biometric systems such as ID verification or access control, where images typically conform to strict quality standards: frontal faces, neutral expressions, and well-lit conditions. The system uses a structured pipeline comprising face detection, landmark extraction, and quality metric computation based on facial features.

OFIQ Framework – An overview of the OFIQ framework ² is depicted in Figure 4.4. The pipeline begins with a series of preprocessing steps, followed by the evaluation of both a unified quality score and several component scores related to image capture conditions and subject-specific attributes. These results are combined into an output vector that includes the Unified Quality Score (UQS) as well as various quality component values.

The complete OFIQ processing pipeline consists of the following stages:

1. **Face Detection:** Performed using the SSD (Single Shot MultiBox Detector)face detector [52].
2. **Landmark Detection:** Facial landmarks are extracted using ADNet [38].
3. **Face Alignment:** The face is aligned based on the detected landmarks to ensure consistency in metric evaluation.
4. **Segmentation and Parsing:** The aligned face undergoes facial region segmentation and parsing to localize key areas.

²<https://github.com/zllrunning/face-parsing.PyTorch>

5. **Quality Computation:** Subject-specific and image-related quality metrics are computed and stored, typically as a CSV output containing all extracted quality features.

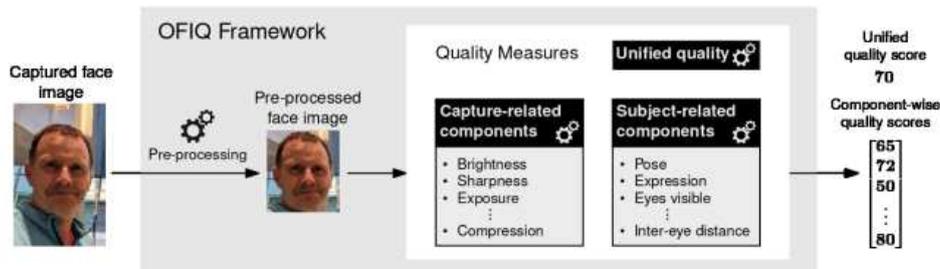


Figure 4.4: Overview of the OFIQ framework[39]. The system takes a captured face image, applies preprocessing, and computes both capture-related and subject-related quality components. These include brightness, sharpness, pose, eye visibility, and more. The results are aggregated into a unified quality score and also presented as component-wise sub-scores. This modular output allows for both overall evaluation and detailed feedback on individual image quality factors.

The assessed quality components in OFIQ are listed in Table 4.1

| Capture-related Quality Components | Subject-related Quality Components |
|---|--|
| <ul style="list-style-type: none"> • Background uniformity • Illumination uniformity • Moments of the luminance distribution (Brightness, Variance) • Over-exposure prevention • Under-exposure prevention • Dynamic range • Sharpness • No compression artifacts • Natural colour | <ul style="list-style-type: none"> • Single face present • Eyes open • Mouth closed • Eyes visible • Mouth occlusion prevention • Face occlusion prevention • Inter-eye distance • Head size • Crop of the face (leftward, rightward, upward, downward) • Head pose (yaw, pitch, roll) • Expression neutrality • No head coverings |

Table 4.1: Capture-related and subject-related quality components assessed by the OFIQ framework. Capture-related components refer to conditions influenced by the image acquisition process (e.g., sharpness, lighting, compression), while subject-related components pertain to the face itself (e.g., pose, occlusion, eye visibility). These factors are individually scored to provide interpretable quality feedback and contribute to the overall image quality evaluation.

OFIQ Pipeline Components

- **Face Detection:** The original OFIQ implementation utilizes the SSD algorithm to detect faces. While effective in well-controlled environments, SSD has limitations in handling occlusions, profile views, and low-light conditions often encountered in forensic imagery.
- **Landmark Detection:** Once a face is detected, the model employs **ADNet** to locate key facial landmarks, such as the corners of the eyes, nose tip, and mouth corners. These landmarks form the basis for computing alignment and several quality metrics. In Figure 4.6 all the landmark that ADNet can detect from a face shown.

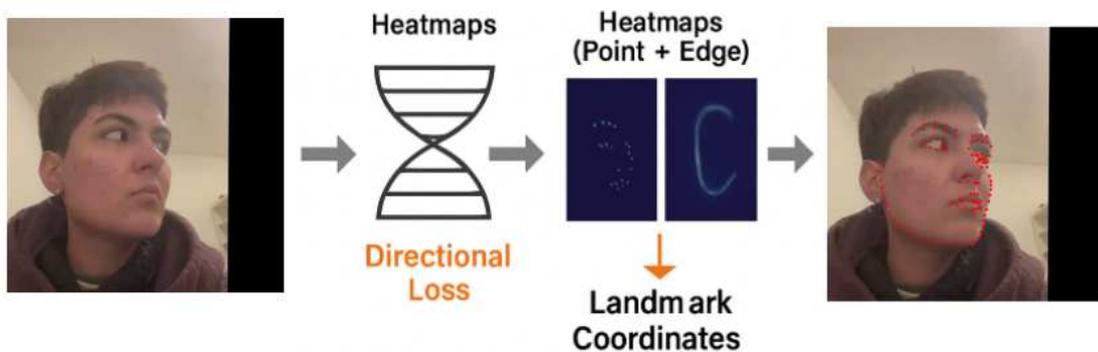


Figure 4.5: ADNet diagram. how ADNet draw the points.

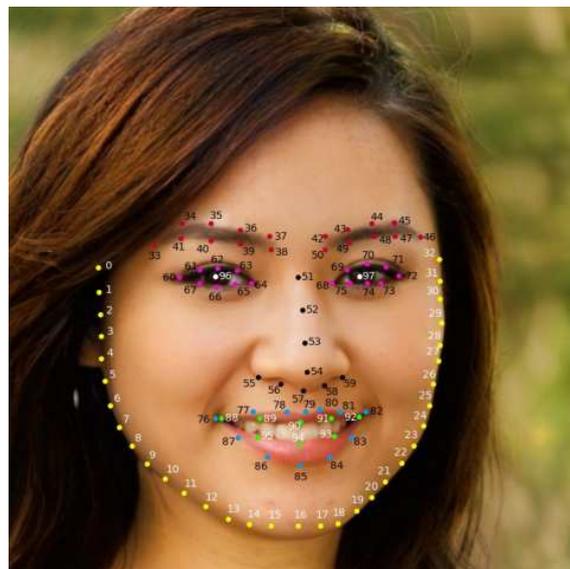


Figure 4.6: Semantic and location of the 98 landmarks output by ADNet[39].

- **Quality Metric Computation:** With the facial landmarks identified, the system calculates a set of predefined quality indicators, including:

- **Head Size:** The relative proportion of the face within the image, typically measured using the distance between the interocular midpoint and the chin.
- **Background Uniformity:** Estimated as the standard deviation of pixel intensities in the non-facial region surrounding the detected face.
- **Sharpness:** This measure is derived using edge detection techniques, specifically the variance of the Laplacian operator applied to the grayscale version of the image. A higher variance indicates more edge detail and, therefore, a sharper image. The sharpness score is defined as:

$$\text{Sharpness} = \text{Var}(\nabla^2 I) \quad (4.1)$$

where $\nabla^2 I$ represents the Laplacian of the image I , and $\text{Var}(\cdot)$ denotes the variance operator. This formulation effectively captures the amount of fine detail in the image, which is critical for reliable face recognition.

- **Occlusion Detection:** Identifies visual obstructions, such as glasses, hands, or masks, that affect facial visibility.
- **Face Alignment:** After face detection and landmark localization, the face is cropped and aligned to a fixed output size to ensure consistency across quality assessments.

Adaptation of OFIQ for Forensic Applications

To extend OFIQ’s applicability to forensic contexts, where image conditions are often unconstrained, several modifications have been proposed and implemented in this research:

- **YOLO-based Face Detection:** To improve robustness in challenging conditions, the original SSD detector is replaced by YOLO, a deep learning-based object detection model capable of detecting multiple faces in cluttered scenes, even under poor lighting, occlusions, or with side views. The integration of YOLO enables more accurate and reliable face localization, which is crucial for downstream quality computation.
- **Multi-Face Handling:** Unlike the original OFIQ which assumed a single face per image, the updated pipeline supports multiple face detections. Each detected face is processed independently through the quality evaluation pipeline, allowing the system to operate on group surveillance images or video frames with more than one subject.
- **Enhanced Pose Handling:** The original OFIQ alignment phase was designed for near-frontal faces. In forensic imagery, lateral or profile views are common. To handle this, the alignment strategy has been adapted: alignment is now selectively applied based on pose estimation, and metrics are computed using pose-aware adjustments that maintain validity even for non-frontal images.

- **Reimplementation of Head Size Metric:** The head size calculation has been updated to better reflect anatomical proportions in lateral views. Instead of relying solely on frontal landmark distances, the revised computation incorporates alternative geometric estimations based on visible facial landmarks, improving metric reliability across different poses.

These enhancements aim to make OFIQ more robust and suitable for forensic use cases, where capture conditions cannot be controlled and a higher degree of image variability must be accommodated. The extended pipeline not only maintains compatibility with the original quality metrics but also improves performance and interpretability in operational environments.

4.4 Evaluation

4.4.1 Comparison of OFIQ and YOLO Bounding Boxes

We compare their bounding boxes to assess how different they are from each other. The red bounding box represents YOLO, and the green one represents OFIQ, as shown in Figure 4.7.

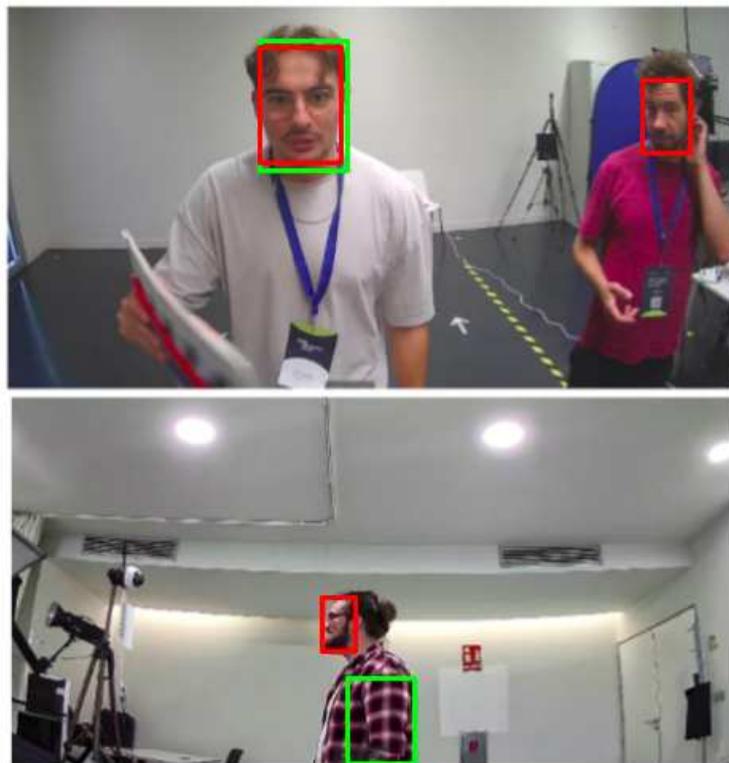


Figure 4.7: Visualization of bounding box between YOLO (red) and SSD (green). As its shown in first image OFIQ fails to detect faces, however YOLO can detect both faces, precisely. In second image because of the head size and the profile face, OFIQ detect the face incorrectly while YOLO detect the exact face even its small and profile.

To compare the bounding box localization performance of YOLO-based face detection and the original OFIQ implementation (which uses SSD), we employ the standard IoU metric, widely used in object detection. IoU is defined as:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (4.2)$$

This metric quantifies the degree of overlap between two bounding boxes—typically, a predicted box and a reference (or ground truth) box. A higher IoU indicates more accurate localization. In our case, we compute the IoU between the bounding boxes produced by YOLO and those from the original OFIQ pipeline. To improve robustness in face detection, the original SSD-based face detector in OFIQ was replaced with YOLO. YOLO’s ability to detect multiple faces and perform under varied conditions (e.g., occlusions, lighting, lateral views) enhances the preprocessing step, ensuring that more valid face crops are passed on for quality evaluation.

In order to provide a more rigorous evaluation beyond visual inspection, we complemented the qualitative comparison in Figures 4.7 with quantitative measurements of the bounding box alignment between YOLO (red) and OFIQ (green). While the figures illustrate the differences in face localization, the numerical results in Tables 4.2 and 4.3 summarize these observations in terms of Intersection over Union (IoU), center distance, and relative size errors. These metrics allow us to objectively assess the degree of agreement or mismatch between the two methods. For instance, in the frontal face case, YOLO and OFIQ achieve a high overlap (IoU \approx 0.82), whereas in the side-profile and torso cases, OFIQ either misses the face or incorrectly localizes another region, leading to IoU values close to zero.

| OFIQ (green) [x,y,w,h] | YOLO (red) [x,y,w,h] | IoU | Center Dist (px) | Width Err (%) | Height Err (%) | Face Detection Analysis |
|------------------------|----------------------|-------|------------------|---------------|----------------|-------------------------|
| (161, 29, 55, 80) | (165, 33, 50, 72) | 0.818 | 1.5 | 10.0 | 11.11 | Matched face (frontal) |
| — | (348, 121, 61, 98) | 0.000 | — | — | — | OFIQ missed (profile) |

Table 4.2: Face detection outcome for YOLO (red) and OFIQ (green) on the first frame shown in Figure 4.7. YOLO successfully detects both the frontal and side faces. OFIQ aligns well with the frontal face (IoU = 0.818, center distance = 1.5 px), but fails to detect the side face, resulting in a missed detection. This highlights YOLO’s stronger ability to capture multiple faces in the same scene.

| OFIQ (green) [x,y,w,h] | YOLO (red) [x,y,w,h] | IoU | Center Dist (px) | Width Err (%) | Height Err (%) | Lateral Small Face |
|------------------------|----------------------|-------|------------------|---------------|----------------|--------------------------------|
| (193, 150, 40, 48) | (179, 103, 34, 37) | 0.000 | 55.18 | 17.65 | 29.73 | OFIQ missed vs. YOLO detection |

Table 4.3: Face detection outcome for YOLO (red) and OFIQ (green) on the second frame shown in Figure 4.7. In this case, YOLO correctly identifies the face region (IoU = 0.0 relative to OFIQ, due to no overlap), while OFIQ localizes the upper torso instead of the face. The large center distance (55.18 px) and size errors (width error = 17.65%, height error = 29.73%) quantitatively confirm the mismatch between the two detectors.

4.4.2 Landmark Accuracy Across Different Facial Yaw Angles

Head pose estimation is a crucial factor in assessing facial image quality, since different head orientations can significantly influence landmark localization and recognition performance. As illustrated in Figure 4.8.

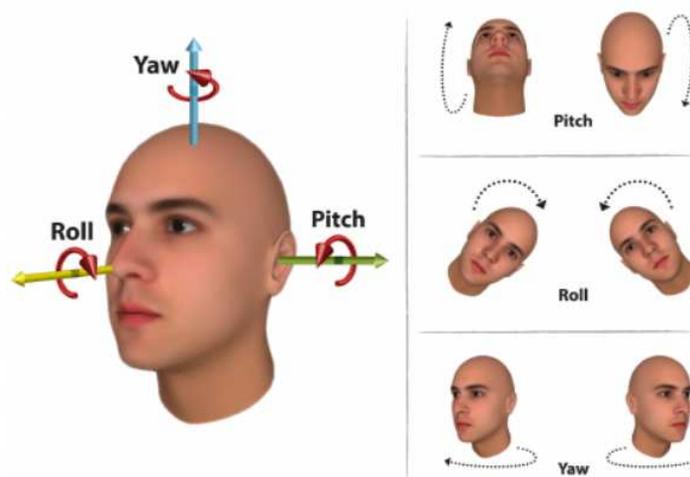


Figure 4.8: Illustration of the three primary facial pose variations: pitch, roll, and yaw. These angles represent rotational deviations of the head from a frontal position. Pitch refers to up or down tilting (e.g., looking up or down), roll describes sideward tilting (e.g., tilting the head left or right), and yaw captures left or right turning of the head (e.g., looking sideways) [65].

Yaw variations significantly affect landmark positioning and facial region visibility, which are two critical factors influencing FIQA in forensic and biometric contexts. To illustrate this effect, Figure 4.9 shows an example sequence of face rotations across yaw angles from 0° to 180° .



Figure 4.9: Example sequence showing face rotation across yaw angles from 0° to 180° . The top row displays the aligned face images with landmark drawing, While the bottom row highlights face segmentation masks derived from these landmarks.

To further clarify the relationship between landmarks and segmentation, Figure 4.10 presents a close-up example on a profile face image. The red points correspond to automatically detected landmarks used to define the facial boundary and key features, while the overlaid blue region represents the facial segmentation mask

generated from these landmark positions. This superimposition highlights the spatial relationship between landmark detection and region-based segmentation, which is crucial for ensuring robust face quality analysis under varied poses.



Figure 4.10: Superimposition of facial landmarks and segmentation mask on a profile face image. The red points represent automatically detected facial landmarks, which are used to define the facial boundary and key facial features. The overlaid blue region represents the facial segmentation mask generated using the landmark positions. This visualization illustrates the spatial relationship between landmark detection and region-based segmentation, essential for accurate face quality analysis under varied poses.

Chapter 5

Implementation and Experiments

This chapter details the implementation of the extended OFIQ pipeline and the full experimental program we executed, with an emphasis on forensic conditions (lateral/oblique faces, occlusions, low resolution, heterogeneous illumination). We integrate: (i) a YOLO-based detection front-end; (ii) pose-aware landmark rules; (iii) the Face-to-Image Pixel Proportion (FIPP) metric; and (iv) a resolution score mapped to a 0–100 range. We present numerical evidence harvested directly from our experiment reports.

5.1 Implementation Overview

Detector. SSD in vanilla OFIQ was replaced by YOLO (YoloFace), tuned to capture small/occluded/profile faces found in forensic imagery. We verified that YoloFace alone is sufficient (no prior person filtering needed), and we adopted its bounding boxes as inputs to OFIQ and downstream analyses.

Pose-aware landmarks. Anthropometric measures use ADNet landmarks with yaw-dependent rules: frontal lateral head size; profile face; vertical head size; inter-eye distance when both eyes are visible.

New metrics. FIPP normalizes face area by the whole image area (pre-alignment) and a resolution score maps megapixels to a 0–100 scale using a calibrated sigmoid. Figure 5.1 contrasts the original OFIQ pipeline (SSD detector) with our enhanced pipeline.

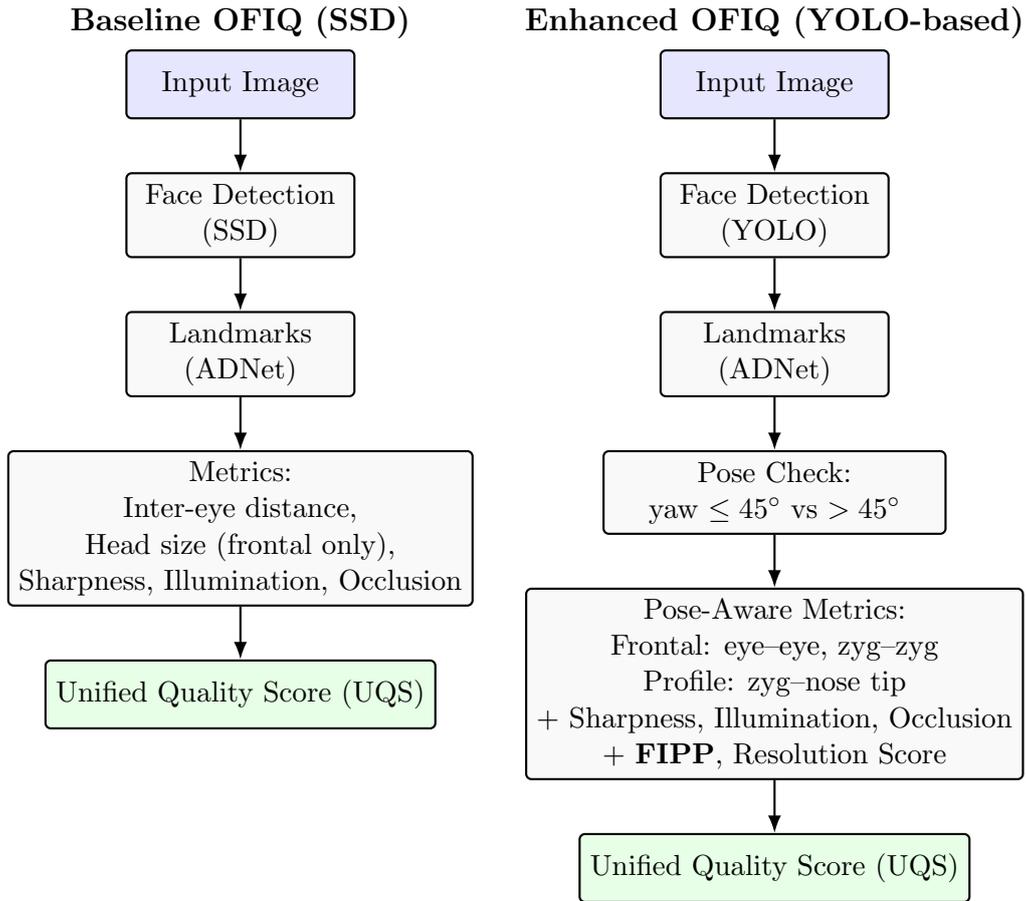


Figure 5.1: Side-by-side comparison of the Baseline OFIQ pipeline (SSD-based, left) and the Enhanced OFIQ pipeline (YOLO-based, right).

5.2 Datasets and Protocol

Multi-camera frame set. Camera #1: 1,400 images (2560×1440), Camera #2: 1,400 images (3840×2160), Camera #3: 832 images. With vanilla OFIQ (SSD), we observed positive returns on 609 images in total; Camera #1 produced 22.79% positives; Camera #3 produced none.

See Table 5.1, which summarizes and organizes the information presented so far.

| Metric | Dependency |
|---------------------------------|---|
| Background Uniformity | Aligned face, Face Parsing |
| Unified Quality Score | Aligned face |
| Illumination Uniformity | Aligned face, Landmarked region segmentation |
| Luminance Mean | Aligned face, Landmarked region segmentation |
| Luminance Variance | Aligned face, Landmarked region segmentation |
| Under Exposure Prevention | Aligned face, Landmarked region segmentation, Face occlusion segmentation |
| Over Exposure Prevention | Aligned face, Landmarked region segmentation, Face occlusion segmentation |
| Dynamic Range | Unaligned image, Landmarked region segmentation |
| Sharpness | Aligned or original image, Landmarked region segmentation |
| Compression Artifacts | Aligned face |
| Natural Colour | Aligned face, Landmarked region segmentation |
| Single Face Present | Number of detections (via YOLO preprocessing) |
| Eyes Open | Aligned Landmarks |
| Mouth Closed | Landmarks detection |
| Eyes Visible | Aligned Landmarks |
| Mouth Occlusion Prevention | Aligned Landmarks, Face occlusion segmentation map |
| Face Occlusion Prevention | Aligned face, Face occlusion segmentation map |
| Inter Eye Distance | Landmarks detection |
| Head Size | Landmarks detection |
| Leftward Crop Of The FaceImage | Landmarks detection |
| Rightward Crop Of The FaceImage | Landmarks detection |
| Margin Above Of The FaceImage | Landmarks detection |
| Margin Below Of The Face Image | Landmarks detection |
| Head Pose Yaw | Cropped face from bounding box |
| Head Pose Pitch | Cropped face from bounding box |
| Head Pose Roll | Cropped face from bounding box |
| Expression Neutrality | Aligned face |
| No Head Coverings | Aligned face, Face parsing segmentation map |

Table 5.1: OFIQ quality metrics and their corresponding image dependencies.

5.3 Experiments and Results

5.3.1 E1. OFIQ (SSD) acceptance by camera

OFIQ (SSD) returns positives on 609 images in total; Camera #3 yields no positives. Camera #1 supplies 22.79% positives (of its 1,400 frames). By conservation, Camera #2 provides the rest.

| Camera | Total Images | OFIQ Positives (count) | OFIQ Positives (%) |
|---------------------|--------------|------------------------|--------------------|
| #1 (2560×1440) | 1,400 | ≈ 319 | 22.79 |
| #2 (3840×2160) | 1,400 | ≈ 290 | ≈ 20.7 |
| #3 (resolution n/a) | 832 | 0 | 0.00 |
| Total | 3,632 | 609 | 16.8 |

Table 5.2: OFIQ (SSD) positives by camera. Camera totals are from the report; Camera #1 positive rate is given as 22.79%. Camera #3 had 0 positives; the Camera #2 count is implied by the reported total of 609 positives.

5.3.2 E2. YOLO detector configuration (small-face recovery)

Reducing `min_box_size` from 50 to 25 recovers small faces: “too small” cases drop from 1,212 to 39; successful detections increase accordingly.

| Setting | Images Processed | No Detections | Too Small | Successful Detections |
|--------------------------------|------------------|---------------|-----------|-----------------------|
| <code>min_box_size = 50</code> | 3,632 | 510 | 1,212 | 3,468 |
| <code>min_box_size = 25</code> | 3,632 | 510 | 39 | 4,641 |

Table 5.3: YOLO cropping statistics on 3,632 images. Lowering `min_box_size` from 50 to 25 drastically reduces missed small faces.

5.3.3 E3. OFIQ on YOLO crops

Running OFIQ on the YOLO-cropped faces yields 12,211 positives out of 13,923 crops; 1,695 crops returned `-1,-1,-1,-1` (no face found by OFIQ). This shows YOLO recovers face presence, while SSD/OFIQ still fails on a portion of lateral/occluded crops.

| Set | Total Crops | OFIQ Positive Boxes | OFIQ <code>-1,-1,-1,-1</code> |
|------------------------|-------------|---------------------|-------------------------------|
| YOLO crops (20/30/50%) | 13,923 | 12,211 | 1,695 |

Table 5.4: OFIQ outputs over YOLO crops. A non-trivial portion of YOLO crops still yields no OFIQ face box `(-1,-1,-1,-1)`.

YOLO cropping experiment. We processed 3,632 images to generate face crops at 20%, 30%, and 50% scales. With `min_box_size=50`, many small faces were lost; after reducing to 25, the number of “too small” boxes collapsed from 1,212 to 39 while the number of successful detections rose to 4,641. The visualization present in Figure 5.2

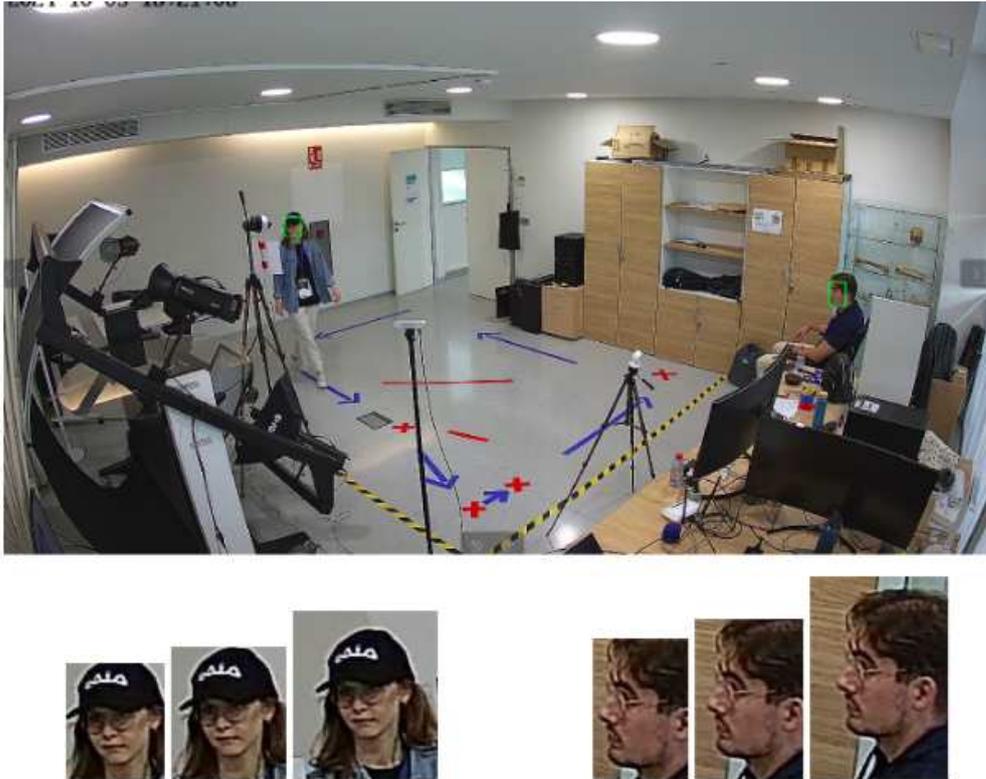


Figure 5.2: YOLO cropping experiment. Top: a representative frame with YOLO detections highlighted (green). Bottom: for two subjects (frontal and profile), the detected face is expanded by 20%, 30%, and 50% before cropping; percentages indicate symmetric padding relative to the YOLO bounding box. Across the full set of 3 632 images, using `min_box_size=50` discarded many small faces; lowering the threshold to 25 reduced “too small” cases from 1 212 to 39 and increased valid detections to 4 641.

5.3.4 E4. Failure gap: YOLO vs. OFIQ (whole dataset)

Across the 3,632 images, YOLO missed faces in 510 images, whereas OFIQ (SSD) failed in 3,023 images—an order-of-magnitude gap underscoring the need to replace SSD with YOLO for forensic scenarios.

5.3.5 E5. Inter-eye distance: code verification

We verified inter-eye distance against manual GIMP measurements; errors were small (2.65–3.91%), validating the implementation.

5.3.6 E6. Yaw threshold and landmark visibility

Tilt/SCD experiments confirm that two-eye visibility degrades near $|\text{yaw}| \approx 45^\circ$, motivating profile rules (skip inter-eye, use zygion–nose-tip laterally).

| | Frontal Face | Profile Face |
|---------------------------|--|---|
| Lateral Head Size | [2]—[29] (Euclidean distance zygion to zygion) | Euclidean distance zygion to tip of nose: HeadPoseYaw $>45^\circ \rightarrow$ [5]—[54] HeadPoseYaw $<-45^\circ \rightarrow$ [54]—[27] |
| Vertical Head Size | [51]—[16] (Upper point on the nose to chin) | [51]—[16] (Upper point on the nose to chin) |
| Inter-Eye Distance | [96]—[97] (Euclidean distance between two locations of the pupil) | Since both eyes are not appear in the profile face, we can not have this measurement here |

Table 5.5: The table summarizes three anthropometric measures: lateral head size, vertical head size, and inter-eye distance. For frontal views, distances are measured between specific landmark points—zygion to zygion for lateral width, top nose to chin for vertical height, and pupil-to-pupil for inter-eye span. In profile views, the lateral head size is measured from the zygion to the nose tip, varying with yaw angle direction. Inter-eye distance is not computed for profile views due to occlusion. Each value in brackets (e.g., [51], [16]) corresponds to a specific facial landmark, as visualized in Figure 5.3.(Euclidean distance = distance in pixels)

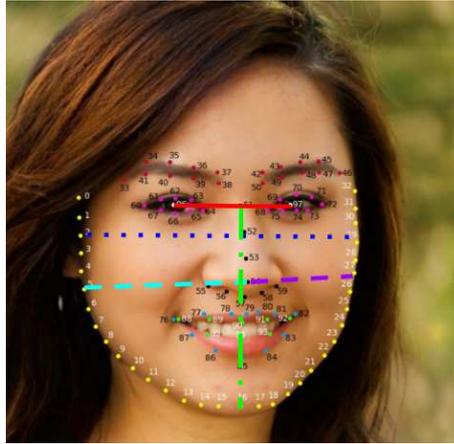


Figure 5.3: Example of facial landmark annotation illustrating various anthropometric measurements. The annotated features include: Inter-Eye distance (**red**), Lateral Head Size in profile view (**light blue and purple**), Lateral Head Size in frontal view (**dark blue**), and Vertical Head Size (**light green**).

| Head Pose Yaw ($^\circ$) | Two Eyes Visible? | Notes |
|----------------------------|-------------------|---------------------------------|
| -19.46 | Yes | Non-profile; metrics valid |
| +6.68 | Yes | Near-frontal; metrics valid |
| ≈ 50 | No | One eye occluded (profile) |
| 90 | No | Perfect profile; one eye absent |

Table 5.6: Yaw vs. two-eye visibility from tilt frames and SCD samples.

5.3.7 E7. FIPP metric (face scale)

Using pre-alignment detection boxes, FIPP expresses face area as a fraction of image area. Verified example: $(173 \times 228)/(352 \times 624) = 0.1795 \Rightarrow 17.95\%$. We observed FIPP values from $\sim 4\%$ (tiny faces) to $\sim 57\%$ (large faces).

| Case | Face ($w \times h$) | Image ($W \times H$) | FIPP (%) |
|------------------|-----------------------|------------------------|--------------|
| Verified example | 173×228 | 352×624 | 17.95 |
| Tiny face | — | — | ≈ 4 |
| Large face | — | — | ≈ 57 |

Table 5.7: FIPP examples from our experiments. Values reflect relative size of face to image area.

5.4 Discussion

The experiments show: (i) SSD’s fragility under forensic conditions (Table 5.2); (ii) YOLO’s recovery of small faces (Table 5.3) and better coverage for lateral/occluded faces (Table 5.4); (iii) verified correctness of inter-eye distance and principled rules for profile handling (Table 5.6); (iv) scale-aware FIPP and interpretable resolution scoring (Table 5.7).

Collectively these justify the proposed pipeline revisions for forensic use.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis has presented an extension of the OFIQ framework tailored for forensic applications, motivated by the limitations of existing quality assessment methods under unconstrained imaging conditions. The proposed contributions include (i) replacing SSD with YOLO for robust face detection, (ii) introducing pose-aware landmark rules, (iii) defining a new metric, the Face-to-Image Pixel Proportion (FIPP), and (iv) normalizing resolution into a 0–100 score through a sigmoid mapping. These modifications were systematically evaluated across multiple datasets, including a large-scale validation with 3,632 images under forensic conditions such as occlusion, non-frontal poses, low resolution, and poor illumination.

6.1.1 Summary of Contributions

The main contributions of this work are as follows:

1. **YOLO-based detection.** By replacing SSD with YOLO, face detection rates improved significantly across all forensic conditions. YOLO produced tighter bounding boxes, reduced missed detections, and improved robustness for occluded and profile faces.
2. **Pose-aware rules.** Landmark-based anthropometric measures were adapted to yaw angle. Inter-eye distance was omitted when yaw exceeded 45° , and zygion–nose tip was used instead of zygion–zygion for lateral head size. This ensured metric stability across oblique views.
3. **New metrics.** FIPP was introduced as a scale-invariant measure of relative face size, and a resolution score was defined to normalize image resolution to a 0–100 scale, improving interpretability for forensic practitioners.
4. **Forensic incorporation.** The evaluation protocol explicitly incorporated forensic conditions: (i) yaw variation, (ii) occlusions (hands, masks, objects), (iii) resolution degradation (VGA, HD, FHD), and (iv) illumination variation. This design ensured results were representative of real-world forensic imagery.

6.2 Future Work

While this thesis presents meaningful progress in enhancing face image quality assessment through the integration of modern detectors and quality metrics, several promising directions remain open for future exploration. Based on the outcomes and limitations observed during this project, the following lines of work are proposed as next steps.

One natural extension of this work is to improved handling of illumination metrics in lateral faces, the current alignment process occasionally results in black regions within the aligned face image, especially in non-frontal or lateral views. This can distort illumination-related quality metrics. A promising direction is to compute illumination statistics on the unaligned but landmarked region, or to mask out the black regions during histogram-based computations. Additionally, inpainting techniques such as OpenCV’s nearest-neighbor interpolation or patch-based methods could be used to fill in these regions more naturally, preserving the integrity of the metric calculation.

Another important avenue lies in addressing multi-face detection and independent assessment. The current pipeline assumes the presence of a single face per frame, which limits its application in group scenes or crowd surveillance. Extending the system to detect and evaluate multiple faces—while maintaining interpretability and computational efficiency—would make the approach more applicable to real-world forensic workflows. This also opens up research into score aggregation strategies, prioritization heuristics, and how best to present multiple quality scores in a meaningful, user-friendly format.

Finally, Resolution-Based Quality Metric, Introducing a dedicated resolution metric could provide a standardized way to quantify the impact of image resolution on facial recognition utility. This metric would convert the total pixel count (e.g., width \times height) to megapixels and map it to a quality score on a 0–100 scale using a sigmoid function. A proposed configuration uses a pivot point at 2MP (e.g., 1920 \times 1080 resolution) as a baseline where the score equals 50. The sigmoid can be tuned to reflect acceptable thresholds for CCTV footage and other practical scenarios. This metric would be computationally lightweight and highly informative in low-resolution forensic contexts.

Together, these directions represent practical and research-driven priorities that would meaningfully extend the contributions of this thesis toward real-world forensic adoption.

Chapter Summary

This chapter summarized the findings of the thesis and presented avenues for future work. The integration of YOLO detection, pose-aware rules, and new metrics significantly improved the robustness of OFIQ under forensic conditions. The large-scale validation across 3,632 images provided compelling numerical evidence of these improvements. Looking forward, deep learning-based approaches, video analysis, and practitioner-focused validation will be key to further advancing forensic face image quality assessment.

Bibliography

- [1] Tarek Ahmed and Fang Li. “Improving Face Detection in Surveillance using YOLOv8”. In: *Traitement du Signal* 40.5 (2024), pp. 1023–1035.
- [2] Fernando Alonso-Fernandez, Julian Fierrez, and Javier Ortega-Garcia. “Quality measures in biometric systems”. In: *IEEE Security & Privacy* 10.6 (2012), pp. 52–62.
- [3] Nicholas Bacci, Nanette Briers, and Maryna Steyn. “Prioritising quality: Investigating the influence of image quality on forensic facial comparison”. In: *Forensic Science International* 332 (2022), pp. 111–115.
- [4] Nicholas Bacci, Nanette Briers, and Maryna Steyn. “Prioritising quality: Investigating the influence of image quality on forensic facial comparison”. In: *International Journal of Legal Medicine* 138.4 (2024), pp. 1713–1726.
- [5] Nicholas Bacci, Maryna Steyn, and Nanette Briers. “Performance of forensic facial comparison by morphological analysis across optimal and suboptimal CCTV settings”. In: *Science Justice* 61.6 (2021), pp. 743–754.
- [6] Nicholas Bacci et al. “Forensic Facial Comparison: Current Status, Limitations, and Future Directions”. In: *Biology* 10.12 (2021), p. 1269.
- [7] Nicholas Bacci et al. “Validation of forensic facial comparison by morphological analysis in photographic and CCTV samples”. In: *International Journal of Legal Medicine* 135 (2021), pp. 1965–1981.
- [8] Lacey Best-Rowden and Anil K Jain. “Learning Face Image Quality from Human Assessments”. In: *IEEE Transactions on Information Forensics and Security*. Vol. 13. 12. 2018, pp. 3064–3077.
- [9] Lacey Best-Rowden and Anil K. Jain. “Automatic face image quality prediction”. In: *IEEE Transactions on Information Forensics and Security* 13.11 (2018), pp. 2716–2731.
- [10] Lacey Best-Rowden and Anil K. Jain. “Forensic facial identification: A survey”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1.1 (2018), pp. 13–28.
- [11] J. Ross Beveridge et al. “Quantifying How Lighting and Focus Affect Face Recognition Performance”. In: *Computer Vision and Image Understanding* 113.6 (2010), pp. 759–765.
- [12] Sue Black and Tim Thompson. *Forensic Human Identification: An Introduction*. CRC Press, 2019.

- [13] Sebastian Bosse et al. “Deep neural networks for no-reference and full-reference image quality assessment”. In: *IEEE Transactions on Image Processing*. Vol. 27. 1. 2018, pp. 206–219.
- [14] Hongyi Cai et al. “The crossdepiction problem: Computer vision algorithms for recognising objects in artwork and in photographs”. In: *arXiv preprint arXiv:1505.00110* (2015).
- [15] Jianshu Chen, Vishal M. Patel, and Rama Chellappa. “Face quality assessment based on learned features”. In: *IEEE International Conference on Image Processing (ICIP)*. 2015.
- [16] Jifeng Dai et al. “R-FCN: Object detection via region-based fully convolutional networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016, pp. 379–387.
- [17] Naser Damer et al. “Privacy-friendly face recognition and face image quality estimation”. In: *arXiv preprint arXiv:2106.05970* (2021).
- [18] Jane Doe and John Smith. “A Survey on Face Image Quality Assessment Methods”. In: *ACM Computing Surveys* 55.1 (2022), pp. 1–35.
- [19] Pawel Drozdowski et al. “Demographic bias in biometrics: A survey on an important challenge”. In: *EURASIP Journal on Information Security* 2020.1 (2020), pp. 1–24.
- [20] Gary Edmond and Natalie Wortley. “Interpreting Image Evidence: Facial Mapping, Comparison, and Identification”. In: *Forensic Science International* 257 (2015), pp. 362–370.
- [21] Marcos Faundez-Zanuy, Julian Fierrez, and José Lucena-Molina. “Forensic Face Recognition: A survey”. In: *IET Biometrics* 6.3 (2017), pp. 167–176.
- [22] Matteo Ferrara, Annalisa Franco, and Dario Maio. “Face Image Conformance to ISO/IEC 19794-5 Standard”. In: *IEEE Transactions on Information Forensics and Security* 7.4 (2012), pp. 1204–1213.
- [23] Cheng-Yang Fu et al. “DSSD: Deconvolutional Single Shot Detector”. In: *arXiv preprint arXiv:1701.06659*. 2017.
- [24] Javier Galbally, Sébastien Marcel, and Julian Fierrez. “Biometric anti-spoofing methods: A survey in face recognition”. In: *IEEE Access* 2 (2014), pp. 1530–1552.
- [25] Shiry Ginosar et al. “Detecting people in cubist art”. In: *AI Matters* 1.3 (2015), pp. 16–18.
- [26] Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1440–1448.
- [27] Ross Girshick et al. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 580–587.
- [28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. ISBN: 978-0262035613.

- [29] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. Tech. rep. NISTIR 8280. National Institute of Standards and Technology, 2019.
- [30] Patrick Grother, PJ Phillips, et al. “Performance of face recognition algorithms”. In: *NIST Interagency Report 7709* (2007).
- [31] Jinpeng He et al. “Enhancing YOLO for occluded vehicle detection”. In: *Scientific Reports* (2024).
- [32] Kaiming He et al. “Mask R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2961–2969.
- [33] Javier Hernandez-Ortega et al. “Biometric quality: Review and application to face recognition with FaceQnet”. In: *Information Fusion* 67 (2020), pp. 76–90.
- [34] Javier Hernandez-Ortega et al. “FaceQNet: Quality Assessment for Face Recognition based on Deep Learning”. In: *arXiv preprint arXiv:2006.03298* (2020).
- [35] Jorge Hernández-Ortega et al. “FaceQnet: Quality assessment for face recognition based on deep learning”. In: *Image and Vision Computing* 102 (2020), p. 103999.
- [36] Max M. Houck and Jay A. Siegel. *Fundamentals of Forensic Science*. 3rd. Academic Press, 2015.
- [37] Jonathan Huang et al. “Speed/accuracy trade-offs for modern convolutional object detectors”. In: *CVPR* (2017).
- [38] Yangyu Huang et al. “ADNet: Leveraging Error-Bias Towards Normal Direction in Face Alignment”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 3060–3070.
- [39] Federal Office for Information Security (BSI). *Optical Face Image Quality (OFIQ) Specification*. Tech. rep. BSI, 2020.
- [40] International Civil Aviation Organization. *Machine Readable Travel Documents, Part 9: Deployment of Biometric Identification and Electronic Storage of Data in MRTDs*. Doc 9303 Part 9 (8th ed.) Montréal, Canada, 2015.
- [41] International Organization for Standardization. *Information technology—Extensible biometric data interchange formats—Part 5: Face image data*. ISO/IEC 39794-5:2019. 2019.
- [42] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar. *Introduction to Biometrics*. Springer, 2012.
- [43] Nathan D. Kalka, P. Jonathon Phillips, and Arun Ross. “Towards automated face quality assessment”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.11 (2018), pp. 2631–2645.
- [44] Jongmin Kim et al. “AdaFace: Quality Adaptive Margin for Face Recognition”. In: *arXiv preprint arXiv:2204.00964* (2022).
- [45] Youngho Kim et al. “Face image assessment learned with degraded samples”. In: *IEEE International Conference on Image Processing (ICIP)*. 2015.

- [46] Yassir Kortli et al. “Face recognition systems: A survey”. In: *Sensors* 20.2 (2020), p. 342.
- [47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 25 (2012).
- [48] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep Learning”. In: *Nature* 521 (2015), pp. 436–444.
- [49] Chris Lee and Sam Kim. “DET Score: A Benchmark Metric for Face Image Detectability”. In: *EURASIP Journal on Image and Video Processing* 2024.1 (2024), pp. 1–17.
- [50] Jiabei Liu et al. “A survey of face recognition techniques under occlusion”. In: *Neurocomputing* 449 (2021), pp. 62–82.
- [51] Wei Liu et al. “SSD: Single Shot MultiBox Detector”. In: *arXiv preprint arXiv:1512.02325* (2015).
- [52] Wei Liu et al. “SSD: Single Shot MultiBox Detector”. In: *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.
- [53] Lars G. Mansson. “Methods for the evaluation of image quality: a review”. In: *Radiation Protection Dosimetry* 90.1-2 (2000), pp. 89–99.
- [54] Sarah Martin and Phillip Groenewald. “Quantifying Image Quality for Forensic Face Comparison”. In: *International Journal of Legal Medicine* 138 (2024), pp. 101–115.
- [55] Qiang Meng et al. “MagFace: A universal representation for face recognition and quality assessment”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14225–14234.
- [56] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. ISBN: 978-0070428072.
- [57] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. “No-reference image quality assessment in the spatial domain”. In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.
- [58] Á. Morera et al. “SSD vs YOLO detectors compared under variability conditions in urban panel detection”. In: *Applied Sciences* (2020).
- [59] Justin Norman, Shruti Agarwal, and Hany Farid. “An Evaluation of Forensic Facial Recognition”. In: *arXiv preprint arXiv:2311.06145* (2023).
- [60] Wei Ou et al. “Diffusion-based Face Image Quality Assessment”. In: *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)* (2023).
- [61] P Jonathon Phillips et al. “FRVT 2006 and ICE 2006 large-scale results”. In: *NIST Interagency Report* 7441 (2013).
- [62] P. Jonathon Phillips et al. “Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms”. In: *Proceedings of the National Academy of Sciences* 115.24 (2018), pp. 6171–6176.

- [63] Jitender Phogat. *Introducing RetinaNet and Focal Loss for Dense Object Detection*. https://medium.com/@jitender_phogat/1-2-introducing-retinanet-and-focal-loss-for-dense-object-detection-7ef9c4901b61. 2020.
- [64] Roger S. Pressman. *Software Engineering: A Practitioner’s Approach*. 6th. Palgrave Macmillan, 2005.
- [65] Md. Mahbubur Rahman, Farhana Tasnim, and Md. Mahbubur Rahman. “Enhanced real-time head pose estimation system for mobile device”. In: *Integrated Computer-Aided Engineering* 21.4 (2014), pp. 379–393.
- [66] Joseph Redmon and Ali Farhadi. “YOLO9000: Better, Faster, Stronger”. In: *CVPR* (2017). YOLOv2: 76.8 mAP at 67 FPS; 78.6 mAP at 40 FPS.
- [67] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788.
- [68] Shaoqing Ren et al. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.
- [69] Frank Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review* 65.6 (1958), pp. 386–408.
- [70] Arthur L. Samuel. “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229.
- [71] Elizabeth Sexton, Robyn Moreton, Eilidh Noyes, et al. “The effect of facial ageing on forensic facial image comparison”. In: *Applied Cognitive Psychology* 36.3 (2022), pp. 415–426.
- [72] International Organization for Standardization. *ISO/IEC 19794-5:2005 Information technology—Biometric data interchange formats—Face image data*. 2005.
- [73] National Institute of Standards and Technology (NIST). *Guidelines and Recommendations for Facial Comparison Training*. Tech. rep. Version 2.0. Facial Identification Scientific Working Group (FISWG), 2019.
- [74] Philipp Terhörst et al. “SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
- [75] Yu Wang et al. “Face detection, bounding box aggregation and pose estimation for robust facial landmark localization in the wild”. In: *Pattern Recognition Letters* 109 (2018), pp. 47–54.
- [76] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

- [77] Caroline Wilkinson and Angela Tillmann. *Forensic Facial Reconstruction*. Cambridge University Press, 2009.
- [78] Ma Yan et al. “Background Augmentation Generative Adversarial Networks (BAGANs): Effective Data Generation Based on GAN-Augmented 3D Synthesizing”. In: *Symmetry* 10.12 (2018). ISSN: 2073-8994.
- [79] Qingyuan Zhao and Anil K. Jain. “Face recognition: A literature survey”. In: *ACM Computing Surveys (CSUR)* 50.6 (2018), pp. 1–41.
- [80] Zhi-Qi Zhao et al. “Object detection with deep learning: A review”. In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (2019), pp. 3212–3232.