

Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University



Dana Anarbayeva

Quality-Adaptive Multimodal Face Recognition with Streaming Prototype Learning and Cross- Modal Evaluation

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Science in Computing Systems and Software

(degree code: 7M06101)

Department of Information Systems

Faculty of Engineering and Natural Sciences

Supervisor: **Merarslan Meraliyev**

Kaskelen, June 2026

School of Information Technologies and Applied Mathematics

Department of Information Systems

Dissertation Signatures Title Page

The Master's Dissertation of **Dana Anarbayeva**

Title:

Quality-Adaptive Multimodal Face Recognition with Streaming Prototype Learning and Cross-Modal Evaluation

Ағындық прототиптік оқыту мен кросс-модальды бағалауға негізделген сапаға бейімделген көпмодальды бет-әлпетті тану жүйесі

Адаптивное к качеству мультимодальное распознавание лиц с потоковым обучением прототипов и кросс-модальной оценкой

Is submitted in partial fulfillment of the requirements for the Master's (MSc) degree in the Department of Information Systems, and approved by the Dissertation Committee:

Head of Committee:

Assistant Prof., PhD, Meraryslan Meraliyev

Academic supervisor:

Assistant Prof., PhD, Merarslan Meraliyev

Member's Signature:

Assistant Prof., PhD, Orynbekova K.E.

Member's Signature:

Assistant Prof., PhD, Ayazbayev D.A.



Declaration

I, Dana Anarbayeva, declare that this dissertation titled “*Quality-Adaptive Multimodal Face Recognition with Streaming Prototype Learning and Cross-Modal Evaluation*” and the work presented in it are my own. I confirm that:

- This work was carried out wholly while in candidature for the degree of Master of Science in Computing Systems and Software at Suleyman Demirel University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- All experiments were performed on publicly available datasets and models, and the code implementing the proposed methods is provided as a publicly available repository.

Dana Anarbayeva
June 2026

Acknowledgements

I owe my deepest thanks to my supervisor, Dr. Merarslan Meraliyev, whose guidance shaped this work from start to finish. His careful feedback, steady patience, and willingness to question my assumptions pushed the research further than I could have taken it alone.

I am also grateful to the faculty of Suleyman Demirel University and the Department of Information Systems. The knowledge I gained in their classrooms gave me the foundation this dissertation is built on, and their teaching will stay with me well beyond it.

To my husband and my family: thank you for your patience through the long evenings, for believing in me when I doubted myself, and for the quiet encouragement that kept me going. This work is as much a product of your support as of my effort.

Finally, I want to acknowledge the open-source community and the developers of the tools and frameworks that made these experiments possible. I am also thankful for AI assistants such as ChatGPT and Claude.ai, which helped me think through problems and refine my writing along the way.

Dedication

This thesis is dedicated to my parents and my husband, whose love and faith in me never wavered, and to the teachers and colleagues who shared their knowledge and good humour throughout these two demanding years. Their encouragement was the steadiest source of motivation behind every page. I also dedicate this work to my supervisor, Merarslan Meraliyev, in gratitude for his guidance.

Abstract

While state-of-the-art face recognition systems achieve near-perfect accuracy under ideal conditions, their performance degrades significantly in real-world conditions when the input images are blurred, noisy, have partially-occluded faces, or are of low resolution. Multi-modal face recognition systems that combine visible-spectrum, near-infrared (NIR), and depth sensor inputs tend to exhibit even greater performance degradation under these conditions because each sensor modality suffers degradation in different ways. In this dissertation, we consider three main problems with current multi-modal face recognition systems: estimating the quality of the input images from each sensor modality, weighting the inputs according to their estimated quality, and maintaining the identity galleries within the recognition systems in an efficient manner.

A new quality estimation function $Q(I)$ is proposed for each input image I that utilises noise-corrected sharpness, noise level, face coverage, and contrast to provide a score that represents the quality of the image without requiring training data. When evaluated on the LFW face dataset with synthetic image degradations, the quality estimates have a Spearman rank correlation coefficient ρ of 0.90 or higher for all four types of image degradation, while the baseline method based on the variance of the Laplacian kernel has a negative correlation with certain types of degradation. When incorporating the quality score estimates into the fusion of visible-spectrum, NIR, and depth sensor recognition results, the system achieves up to a 4.4 percentage point improvement in Rank-1 recognition accuracy under degraded NIR inputs. A streaming framework for updating the identity galleries that utilises an exponential moving average with a momentum value of 0.9 and reservoir sampling achieves the same accuracy as full retraining of the model using only 49% of the computational cost, and utilises only the embedding vectors of the faces rather than the raw face images.

A fairness audit of the system on the FairFace dataset reveals a null result in relation to the fairness metric, suggesting future work on this topic. All experiments utilised publicly available datasets and models, and the code for implementing the proposed methods is provided as a publicly available repository.

Keywords: face recognition, image quality estimation, multi-modal fusion, adaptive weighting, streaming prototype learning, incremental gallery update, fairness in biometrics, ArcFace.

Аңдатпа

Заманауи бет-әлпетті тану жүйелері бақыланатын жағдайларда жоғары дәлдікке жетсе де, олардың өнімділігі нақты жағдайларда айтарлықтай төмендейді — мұндай жағдайларда кіріс суреттер бұлыңғыр, шулы болуы, бет-әлпеттің бір бөлігі жабылуы немесе ажыратымдылығы төмен болуы мүмкін. Көрінетін спектрдің, жақын инфрақызыл диапазонның (NIR) және тереңдік сенсорларының деректерін біріктіретін көпмодальды бет-әлпетті тану жүйелері бұдан да күшті өнімділік төмендеуіне ұшырайды, өйткені әрбір сенсорлық модальділік әртүрлі тәсілмен деградацияға ұшырайды. Бұл диссертацияда қазіргі заманғы көпмодальды бет-әлпетті тану жүйелерінің үш негізгі проблемасы қарастырылады: әрбір сенсорлық модальділіктен алынған кіріс суреттердің сапасын бағалау, кіріс деректерді бағаланған сапаға сәйкес өлшемдеу және тану жүйелерінің ішіндегі идентификатор галереяларын тиімді жүргізу.

Әрбір кіріс сурет I үшін жаңа сапаны бағалау функциясы $Q(I)$ ұсынылады. Бұл функция шуды есепке алумен өткірлікті, шу деңгейін, бет-әлпеттің жабылу дәрежесін және контрастты пайдаланып, оқыту деректерін қажет етпестен суреттің сапасын бейнелейтін балды есептейді. LFW деректер жинағында синтетикалық бұзылулармен тексергенде Спирменнің дәреже корреляция коэффициенті ρ барлық төрт деградация түрі бойынша 0,90 және одан жоғары мәнге жетті, ал Лаплас ядросының дисперсиясына негізделген базалық әдіс кейбір деградация түрлерімен теріс корреляция көрсетті. Сапа бағалауларын көрінетін спектр, NIR және тереңдік сенсорларының тану нәтижелерін біріктіру процесіне енгізу кезінде жүйе NIR кіріс деректерінің деградациясы жағдайында Rank-1 идентификация дәлдігін 4,4 пайыздық тармаққа дейін жақсартады. 0,9 импульс мәнімен экспоненциалды жылжымалы орташаны және резервуарлық іріктеуді пайдаланатын галереяларды жаңарту үшін ағындық жақтау есептеу шығындарының тек 49%-ын пайдаланып, модельді толық қайта оқытумен бірдей дәлдікке қол жеткізеді және шикі бет суреттерінің орнына тек бет эмбеддингтерінің векторларын сақтайды.

FairFace деректер жинағында жүйені әділдікке аудиторлау жүргізу әділдік метрикасы бойынша нөлдік нәтиже берді, бұл аталған тақырып бойынша болашақ зерттеулердің қажеттілігін көрсетеді. Барлық эксперименттер жалпыға қолжетімді деректер жинақтары мен модельдер пайдаланылып жүргізілді; ұсынылған әдістерді іске асыруға арналған код жалпыға қолжетімді репозиторийде орналастырылған.

Аннотация

Несмотря на то что современные системы распознавания лиц достигают почти идеальной точности в контролируемых условиях, их производительность существенно снижается в реальных условиях, когда входные изображения размыты, зашумлены, частично перекрыты или имеют низкое разрешение. Мультимодальные системы распознавания лиц, объединяющие данные камер видимого спектра, ближнего инфракрасного диапазона (NIR) и глубинных датчиков, подвержены ещё более значительному снижению производительности, поскольку каждая модальность деградирует по-своему. В данной диссертации рассматриваются три основные проблемы современных мультимодальных систем распознавания лиц: оценка качества входных изображений каждой сенсорной модальности, взвешивание входных данных в соответствии с оценённым качеством и эффективное ведение галерей идентификаторов внутри систем распознавания.

Для каждого входного изображения I предлагается новая функция оценки качества $Q(I)$, использующая резкость с поправкой на шум, уровень шума, покрытие лица и контраст, и формирующая оценку качества изображения без необходимости в обучающих данных. При тестировании на датасете LFW с синтетически искажёнными изображениями коэффициент ранговой корреляции Спирмена ρ составил 0,90 и выше для всех четырёх типов деградации, тогда как базовый метод, основанный на дисперсии ядра Лапласа, показал отрицательную корреляцию при некоторых типах деградации. При включении оценок качества в процедуру слияния результатов распознавания по каналам видимого спектра, NIR и глубины система достигает улучшения точности идентификации Rank-1 на 4,4 процентных пункта при деградации NIR-канала. Поточковая структура обновления галерей идентификаторов, использующая экспоненциальное скользящее среднее с моментом 0,9 и резервуарную выборку, достигает той же точности, что и полное переобучение модели, используя лишь 49% вычислительных затрат, и хранит только векторы эмбедингов лиц, а не исходные изображения.

Аудит справедливости системы на датасете FairFace выявил нулевой результат по метрике справедливости, что указывает на необходимость дальнейших исследований в данном направлении. Все эксперименты проводились с использованием общедоступных датасетов и моделей; код реализации предложенных методов опубликован в общедоступном репозитории.

List of Abbreviations

AdaFace	Quality-Adaptive Margin loss for face recognition
ArcFace	Additive Angular Margin loss for face recognition
AUC	Area Under the ROC Curve
CNN	Convolutional Neural Network
CosFace	Large Margin Cosine loss for face recognition
CPU	Central Processing Unit
DER	Dark Experience Replay
DIR	Detection and Identification Rate
EMA	Exponential Moving Average
EWC	Elastic Weight Consolidation
FaceQnet	Face Quality Network
FAR	False Accept Rate
FRVT	Face Recognition Vendor Test
GDPR	General Data Protection Regulation
GDumb	Greedy Sampler and Dumb learner
iCaRL	Incremental Classifier and Representation Learning
IJB-C	IARPA Janus Benchmark-C
IJB-S	IARPA Janus Benchmark-S
LFW	Labeled Faces in the Wild
MagFace	Magnitude-aware face representation
MSc	Master of Science
NIR	Near-Infrared
NIST	National Institute of Standards and Technology
RGB	Red–Green–Blue (visible-spectrum image)
RGB-D	RGB–Depth
ROC	Receiver Operating Characteristic
RQ	Research Question
SER-FIQ	Stochastic Embedding Robustness Face Image Quality
SphereFace	Angular margin hypersphere embedding

TAR	True Accept Rate
VIS	Visible spectrum
VIS-NIR	Visible–Near-Infrared

Contents

Declaration	ii
Acknowledgements	iii
Dedication	iv
Abstract	v
List of Abbreviations	viii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	2
1.3 Research Questions	3
1.4 Contributions	3
1.5 Experimental Overview	4
1.6 Dissertation Outline	4
1.7 Scope and Limitations Overview	5
2 Background and Related Work	6
2.1 Face Recognition: From Softmax to Margin-Based Loss	6
2.1.1 Traditional Softmax Training	6
2.1.2 Margin-Based Loss Functions	6
2.1.3 Why This Dissertation Uses ArcFace	7
2.2 Face Image Quality Estimation	7
2.2.1 Classical Image Quality Metrics	7
2.2.2 Learning-Based Quality Estimation	8
2.2.3 Why a New Quality Estimator Is Needed	8
2.3 Multi-Modal Face Recognition and Fusion	9
2.3.1 Modalities in Face Recognition	9
2.3.2 Fusion Strategies	9
2.3.3 Why Quality-Adaptive Fusion	10
2.4 Continual and Incremental Learning for Face Recognition	10
2.4.1 The Catastrophic Forgetting Problem	10

2.4.2	Regularisation-Based Approaches	11
2.4.3	Rehearsal and Replay-Based Approaches	11
2.4.4	Prototype-Based Approaches	11
2.4.5	Why Streaming Prototype Learning	12
2.5	Evaluation Protocols and Metrics	12
2.5.1	Face Verification Metrics	12
2.5.2	Quality–Accuracy Correlation Metrics	13
2.5.3	Fairness Metrics	13
2.5.4	Evaluation Fragmentation in Multi-Modal Recognition	13
2.6	Research Gaps and Dissertation Contributions	14
2.7	Chapter Summary	14
3	Methodology	16
3.1	System Overview	16
3.2	Datasets	17
3.2.1	LFW (Labeled Faces in the Wild)	17
3.2.2	CASIA-WebFace Subset	17
3.2.3	FairFace	17
3.2.4	Synthetic Modality Generation	18
3.3	Face Recognition Backbone	18
3.3.1	Architecture	18
3.3.2	ArcFace Loss Function	19
3.3.3	Preprocessing	19
3.3.4	Similarity Computation	19
3.4	Quality Estimation Module	20
3.4.1	Components	20
3.4.2	Final Quality Score	21
3.5	Multi-Modal Fusion Strategy	21
3.5.1	Notation	21
3.5.2	Baseline 1 — Single Modality	21
3.5.3	Baseline 2 — Fixed Equal-Weight Fusion	22
3.5.4	Proposed — Quality-Adaptive Fusion	22
3.6	Streaming Prototype Learning	22
3.6.1	Notation	22
3.6.2	Prototype Update Rule	22
3.6.3	Replay Buffer (Reservoir Sampling)	23
3.6.4	Comparison Baselines	23
3.7	Evaluation Protocols and Metrics	23
3.7.1	Verification Metric: TAR@FAR	23
3.7.2	Identification Metric: Rank-1	24
3.7.3	Quality–Accuracy Correlation: Spearman ρ	24

3.7.4	Fairness Metrics	24
3.7.5	Pair Construction Protocols	24
3.8	Experimental Setup Summary	25
3.9	Reproducibility	25
3.10	Chapter Summary	25
4	Experimental Results	27
4.1	Experiment 1 — Quality Score Validation	27
4.1.1	Clean Baseline	27
4.1.2	Degradation Protocol Visualisation	27
4.1.3	Recognition Accuracy Under Degradation	27
4.1.4	Spearman Correlation Results	29
4.2	Experiment 2 — Quality-Adaptive Multi-Modal Fusion	30
4.2.1	Fusion Ablation (E2 Main)	30
4.2.2	NIR Degradation Robustness (E2b)	30
4.3	Experiment 3 — Streaming Prototype Learning	32
4.3.1	Per-Batch Rank-1 Trajectory	32
4.3.2	Final-Test Accuracy and Computational Cost	32
4.4	Experiment 4 — Quality-Adaptive Fusion and Fairness Audit	33
4.4.1	Overall Recognition Accuracy	33
4.4.2	Per-Group Fairness Analysis	34
4.4.3	Interpretation of the Null Fairness Finding	35
4.5	Summary of Results Across All Experiments	35
5	Discussion	37
5.1	Interpretation of Results	37
5.1.1	Quality Estimation Works — But Within Boundaries	37
5.1.2	Quality-Adaptive Fusion Is Most Valuable Under Selective Degradation	37
5.1.3	Streaming Prototype Learning Is Practically Viable	38
5.1.4	Quality-Adaptive Fusion Does Not Improve Fairness — A Genuine Finding	38
5.2	Novelty and Contributions	39
5.2.1	Multi-Component Quality Estimator Handling All Four Major Degradation Types	39
5.2.2	Quality-Adaptive Fusion with Inference-Time Modality Weighting	39
5.2.3	Privacy-Preserving Streaming Prototype Learning	39
5.2.4	Transparent Null Finding on Fairness	40
5.3	Limitations	40
5.3.1	Synthetic Modalities	40
5.3.2	Synthetic Degradations for Quality Validation	40
5.3.3	Dataset Scale and Representativeness	41

5.3.4	Quality Estimator Calibration Across Demographics	41
5.3.5	Label Dependency in Streaming	41
5.3.6	Single Temperature Parameter	41
5.4	Positioning Within the Broader Literature	41
5.4.1	Quality Estimation	41
5.4.2	Multi-Modal Fusion	42
5.4.3	Continual Learning	42
5.5	Future Work	42
5.6	Answers to the Research Questions	43
5.6.1	RQ1: How Can Quality Variation Across Modalities Be Handled Adap- tively?	43
5.6.2	RQ2: How Can Identity Galleries Be Updated Incrementally?	43
5.6.3	RQ3: How Should Cross-Modal Performance Be Evaluated Consistently?	43
6	Conclusion	44
6.1	Summary of Contributions	44
6.2	Final Answers to the Research Questions	45
6.2.1	RQ1: Adaptive Handling of Quality Variation Across Modalities	45
6.2.2	RQ2: Incremental Identity Gallery Updates	45
6.2.3	RQ3: Consistent Cross-Modal Evaluation	46
6.3	Honest Self-Assessment	46
6.4	Broader Significance	46
6.5	Closing Statement	47

Chapter 1

Introduction

1.1 Background and Motivation

Face recognition has undergone a remarkable transformation over the past decade. Methods based on deep convolutional neural networks using margin-based angular loss functions, such as CosFace [1], ArcFace [2], and AdaFace [3], have enabled verification accuracy on datasets like LFW [4] and IJB-C [5] to reach or even exceed that of humans. The accuracy touted for these methods, however, is established using face images that are very different from those encountered in the wild.

In the wild, the face images that must be processed by recognition methods are often motion blurred due to the movements of the human’s hand or head, contain noise due to the high ISO settings of the camera used to take the image in a poorly illuminated area, are potentially occluded by the presence of glasses or other objects that cover part of the face, and may be of low resolution due to the distance of the individual from the camera or the resolution of the camera’s sensor. Each of these issues can lead to a decrease in the recognition accuracy of a method, and combinations of these issues can lead to catastrophic decreases in accuracy. A system that achieves 99% accuracy on a clean benchmark may fall to only 40% accuracy if a Gaussian blur with $\sigma = 5$ is applied to the face images—as the ArcFace backbone achieves $\text{TAR@FAR} = 10^{-3} = 0.888$ on the clean LFW dataset but only 0.404 with blurred images.

The modern security system often includes multiple types of sensors to provide enhanced surveillance. These sensors include visible-spectrum RGB cameras, near-infrared (NIR) sensors that enable surveillance in the dark, and depth sensors that are used to determine if the detected face is a live subject or a photograph. The benefit of using multiple sensors is that if one sensor fails, the others can compensate for its failure. However, simply fusing the outputs of the sensors in a simplistic manner can drastically reduce the security system’s performance if one of its sensors provides corrupted data.

In order to address this problem, it is necessary to create a system that can evaluate the quality of each sensor’s data stream at the time of recognition. A quality estimator would allow the system to determine the importance of each sensor according to the quality of its data stream. If the system has access to a quality estimator, it will be able to determine if it is safe to rely

upon the recognition of only one sensor as opposed to all of them. This will allow the system to avoid giving equal importance to a sensor that detected an accurate face and one that provided corrupted data.

Another challenge for facial recognition systems is the static nature of the enrolled identities within the system. All of the identities that are to be recognised by the system are enrolled at the start of the system's creation. Furthermore, any addition of new identities requires the system to be retrained with all of the identities. This is not a process that is feasible to implement in a security system that recognises individuals in real time. The phenomenon of catastrophic forgetting when adding new information to a machine learning system is one reason that this approach is not well-suited to facial recognition systems [6].

A third challenge to the deployment of facial recognition systems is the lack of a standardised protocol for evaluating the success of these systems. Most research papers that publish results from facial recognition systems employ different models of facial detection sensors, use different datasets to test their systems, use different thresholds for the FAR (False Acceptance Rate) of their systems, and even define the evaluation of cross-modal facial recognition differently. As a result, it is not possible to compare the performance of different facial recognition systems within research publications. Furthermore, as facial recognition systems are employed in increasingly sensitive and high-stakes applications, the lack of an established means of auditing the fairness of these systems is a potential societal risk.

All three of these challenges are addressed in this dissertation. Methods for both estimating the quality of each sensor's stream of recognition data, and for fusing the recognition outputs from each sensor according to that estimated quality are proposed. Furthermore, a method for stream-wise updating of the enrolled identities within the system without catastrophic forgetting is proposed and evaluated. Finally, all three of these methods are evaluated according to a protocol that specifically includes the auditing of the fairness of the systems according to each of its demographic subgroups. The key innovation in all of these methods is the recognition of the importance of being able to determine which sensor to trust at any given time.

1.2 Problem Statement

Let a multi-modal face recognition system have modality set $\mathcal{M} = \{\text{RGB}, \text{NIR}, \text{Depth}\}$. Given a face that is to be recognised, the system receives a set of images, one from each modality $m \in \mathcal{M}$. These images may each have been captured under different degradations. The system must compare this face to a gallery $\mathcal{G}$ of enrolled identities; the size of this gallery may grow over time. Three problems emerge in this scenario, each of which is solved in this dissertation.

The quality estimation problem requires that, for a given image I_m from modality m , a scalar quality score $Q(I_m)$ is estimated. This score ranges between 0 and 1 and indicates the utility of that given image for recognition. Existing methods for estimating quality relate only to single-modality and single-type-degradation cases.

The adaptive fusion problem requires that, given the embeddings $\mathbf{f}_m$ from each of the available modalities, and the quality scores $Q(I_m)$ for each of those same images, an embedding $\mathbf{f}^*$ is calculated which optimises some criterion for face recognition, while allowing for the possibility that one or more of those embeddings may have been captured with degradations to the quality of those inputs. Existing methods either require all inputs to have equal weights, or require training of the face recognition model with all available modality inputs together.

The incremental gallery problem requires that, given a growing gallery $\mathcal{G}$ of enrolled identities, a method is devised to maintain these enrolments such that (a) new identities can be added without retraining the model, (b) recognition of existing identities does not degrade with the addition of new identities, (c) the cost of maintaining the model grows sub-linearly with the size of the gallery, and (d) no raw face images are required to be retained by the system.

1.3 Research Questions

Three research questions structure this work. Each is directly addressed by one or more experimental components.

RQ1.

How can quality variation across modalities be handled adaptively to maintain robust face recognition performance when one or more modalities are degraded or absent?

RQ2.

How can an identity gallery be updated incrementally to accommodate new identities without full retraining, while preventing catastrophic forgetting of existing identities?

RQ3.

How can cross-modal face recognition performance be evaluated consistently and fairly to enable meaningful comparison across methods and demographic groups?

1.4 Contributions

This dissertation makes four contributions:

C1. Multi-component quality estimator.

A quality estimator $Q(I)$ combining four quality components is proposed that can directly be used to estimate the quality of a face image in a training-free way. Using synthetic degradations of the LFW dataset, the Spearman rank correlation coefficients between the quality estimator and the label of image quality are all ≥ 0.90 for four different degradation types. For comparison, the Laplacian quality measure achieved correlation coefficients of -0.354 for noise degradations and -0.900 for occlusion degradations.

C2. Quality-adaptive multi-modal fusion.

A method to fuse RGB and NIR images using a softmax function with quality-adaptive weights is proposed that does not require any joint training of the components of the system. The method is evaluated on synthetic degradations of the NIR channel and outperformed a simple 50/50 fusion method by up to 4.4 percentage points in Rank-1 identification accuracy at moderate degradation levels.

C3. Streaming prototype learning.

A framework to update the appearance model using EMA updates of the appearance prototypes with $\alpha = 0.9$ and samples from a reservoir of previously seen face images is proposed. The framework only needs to store 512-d embedding vectors of the face images and can process each newly seen image in $O(1)$ time. On the CASIA-WebFace dataset, the framework achieves the same accuracy in face verification as the original CNN model at 49% of the computational cost of the original model.

C4. Transparent null finding on fairness.

A demographic fairness audit of the method on the FairFace dataset for 14 subgroups (7 different races and 2 genders each) shows that the quality-adaptive fusion mechanism does not lead to a reduction of the demographic performance disparity of the system as compared to the system that uses fixed fusion. The fixed fusion method leads to a higher mean TAR value (0.652 vs. 0.632) and a smaller gap between its worst and best demographic performance (0.467 vs. 0.480). These results indicate that quality-awareness alone is insufficient as a means to ensure demographic fairness in face recognition systems.

Scope qualification: The validity of the proposed contributions C1 and C2 are based on synthetic degradations of the LFW dataset and synthesised data. The evaluation of these contributions on real RGB-D and NIR-VIS sensor data is part of future work. The qualitative results presented here are sound and expected to hold true with real sensor data though the exact magnitudes of the evaluated metrics might slightly differ.

1.5 Experimental Overview

Four experiments are conducted, each addressing one research question or sub-question. Table 1.1 provides a complete overview.

1.6 Dissertation Outline

The remainder of this dissertation is organised as follows. Chapter 2 reviews the existing literature on each of the topics discussed in this introduction, including loss functions, image quality estimation, fusion methods, continual learning, and evaluation. Chapter 3 describes the methodology used to create and test the quality estimator, fusion method, and streaming framework.

Table 1.1: Overview of experiments, datasets, primary metrics, and key results. All experiments use the ArcFace ResNet-50 backbone with random seed 42 throughout.

Exp.	Title	Dataset	Metric	Key Result
1	Quality Score Validation	LFW (500 pairs, synth. degr.)	Spearman ρ	$\rho \geq 0.90$ all 4 types
2	Fusion Ablation	LFW (1 000 pairs, synth. RGB+Depth)	Rank-1	Adaptive ties best modality
2b	NIR Degradation Robustness	LFW (1 000 pairs, synth. NIR $\times$ 6)	Rank-1 gain	Peak +4.4pp adaptive vs. fixed
3	Streaming Prototype Learning	CASIA-WebFace (200 ids $\times \approx 101$)	Rank-1, cost	Matches full retrain at 49% cost
4	Fairness Audit	FairFace (2 100 imgs, 14 subgroups)	TAR@FAR / group	Null finding (no fairness gain)

Chapter 4 presents the results of each of the four experiments performed in this dissertation. Chapter 5 discusses these findings in detail in order to highlight the novelty of the contributions, the limitations of the methods presented, how they relate to the existing literature on the topic, and what type of future work can be performed. Finally, Chapter 6 provides an honest assessment of the work and discusses the significance of the findings.

1.7 Scope and Limitations Overview

Several aspects of the problem of facial recognition are outside of the scope of this dissertation. Most notably, the quality estimation model was only built and tested for four types of synthetic degradations, and does not encompass all of the factors that may impact the quality of real-world facial recognition systems. Additionally, the experiments that evaluated the fusion and streaming methods used synthesised image and video modalities rather than real data from multi-sensor systems. Finally, the streaming method evaluated in this dissertation assumed access to ground-truth identity labels of the individuals in the stream, making it applicable to supervised streaming scenarios but excluding unsupervised streaming methods.

These limitations are discussed throughout this dissertation. Each is discussed to provide an honest and narrow statement of the contributions of this work—a more valuable position than an overstated one that can be immediately challenged by its examiner or reviewer.

Chapter 2

Background and Related Work

This chapter reviews the literature relevant to the three research questions of this dissertation. Section 2.1 traces the evolution of face recognition loss functions and justifies the choice of ArcFace as the backbone model. Section 2.2 surveys face image quality estimation and explains why a new multi-component estimator is needed. Section 2.3 reviews multi-modal fusion strategies and motivates the quality-adaptive weighting approach. Section 2.4 covers continual and incremental learning and justifies the streaming prototype design. Section 2.5 discusses evaluation protocols and metrics. Section 2.6 presents the gaps table—a structured comparison of what existing work provides versus what this dissertation contributes.

2.1 Face Recognition: From Softmax to Margin-Based Loss

2.1.1 Traditional Softmax Training

The earliest deep face recognition systems used a standard softmax loss function to train convolutional neural networks. The softmax loss function is effective at learning features that allow for recognition of faces. However, there are two problems associated with softmax loss functions. First, angular margins between identity classes are not enforced by the softmax loss function. Second, softmax loss functions tend to degrade in performance when recognizing identities that were not seen during training. These shortcomings of softmax loss functions have motivated a whole generation of loss functions that enforce margins between identity classes.

2.1.2 Margin-Based Loss Functions

SphereFace [10] proposed the first multiplicative angular margin, forcing the model to increase the angle between different classes. However, the method requires numerical approximations that can cause instability during training. CosFace [1] formulates the problem differently by operating directly in cosine space and introducing an additive margin m . This approach is more numerically stable and produces better generalising embeddings. However, the additive margin treats all samples equally during training, irrespective of the quality of those samples.

ArcFace [2] proposed an additive angular margin loss in angular space on the hypersphere. The loss function is defined as:

$$\mathcal{L}_{\text{ArcFace}} = -\log \frac{e^{s \cdot \cos(\theta_{y_i} + m)}}{e^{s \cdot \cos(\theta_{y_i} + m)} + \sum_{j \neq y_i} e^{s \cdot \cos \theta_j}} \quad (2.1)$$

where θ_{y_i} is the angle between the sample embedding and the target class weight vector, m is the angular margin, and s is the feature scale. The geometric intuition here is that the margin m pushes each sample embedding away from the hyperplane.

AdaFace [3] proposed an adaptive face recognition loss that makes the angular margin of each sample dependent on the quality of the input image. Specifically, low-quality images have a smaller angular margin than high-quality images. The quality of the face is approximated by the norm of the embedding before normalisation—high-quality faces have higher magnitude embeddings. By introducing an adaptive margin, AdaFace outperforms ArcFace on low-quality image benchmarks such as IJB-S.

2.1.3 Why This Dissertation Uses ArcFace

ArcFace is chosen as the backbone of this dissertation for three reasons. First, ArcFace is the standard model that can be referenced when comparing results on the LFW, IJB-C, and the CASIA NIR-VIS 2.0 datasets. Second, ArcFace is publicly available with pretrained weights that can be downloaded and utilised for training. Finally, the quality proxy metric that is utilised in Chapter 3 was inspired by AdaFace’s insight regarding embedding norms, but avoids the more complex training protocol that AdaFace utilised in its training process. While the weights for AdaFace would be ideal for this project, they are not publicly available.

In Experiment 1, ArcFace achieves $\text{TAR@FAR} = 10^{-3} = 0.888$ and $\text{AUC} = 0.986$ on the clean LFW dataset, which is in line with the published results for ArcFace, indicating that this backbone network provides a solid baseline for the study.

2.2 Face Image Quality Estimation

2.2.1 Classical Image Quality Metrics

The simplest approach to face image quality estimation is to use general-purpose image quality statistics. The most common of these metrics is the Laplacian variance statistic [7], which measures the variance of the response of a discrete Laplacian filter applied to the image. Other classical image quality statistics include gradient magnitude, power spectral entropy, and contrast (which is the standard deviation of the pixel intensities). Each of these statistics is computationally efficient to calculate, requiring no training. However, each of these metrics was developed for measuring the quality of natural images rather than recognition-specific image quality. For instance, an image can contain a high Laplacian variance yet be heavily corrupted by noise, as observed in Experiment 1.

2.2.2 Learning-Based Quality Estimation

FaceQnet [8] was the first model to learn face-specific quality estimation from recognition-correlated labels. During training, a pair of face images of the same identity but with different quality levels are used, with the quality label assigned to the higher-quality image in each pair. While FaceQnet produces excellent results, it is limited to RGB images and does not provide its weights for NIR or depth images.

SER-FIQ [9] takes a completely different approach by learning the quality of a face image by measuring the robustness of the embedding representation of the face image. By passing the same image through the neural network multiple times with Monte Carlo dropout enabled, the quality of the image is computed by measuring the variance of the embeddings. Images with a low variance in the embedding representations are awarded high quality since a face recognition system would be more confident in recognizing faces of higher quality. Although SER-FIQ does not require quality labels during training, it is computationally expensive due to the requirement of 10 to 30 forward passes through the neural network for each image.

MagFace [11] learns a joint loss function that optimizes for both recognition accuracy and quality estimation by using the magnitude of the embedding as a signal of quality. AdaFace [3] uses the norm of the pre-normalisation embedding as a proxy for quality since the norm of the embedding is correlated to the recognition confidence of the model. Although both of these approaches are elegant, they are inapplicable to any face recognition model framework that uses L_2 -normalised embeddings like InsightFace.

2.2.3 Why a New Quality Estimator Is Needed

Table 2.1 presents a comparison of the existing quality estimation methods along the dimensions relevant to this dissertation.

Table 2.1: Comparison of face image quality estimation methods. ✓ = handles correctly, ✗ = fails or not applicable, △ = partial/limited support. The proposed method is the only approach that handles all four degradation types, supports non-RGB modalities, and requires no training.

Method	Blur	Noise	Occl.	Res.	NIR/Depth	Train	Speed
Laplacian Var. [7]	✓	✗	✗	✓	△	No	Fast
FaceQnet [8]	✓	✓	✓	✓	✗	Yes	Fast
SER-FIQ [9]	✓	✓	✓	✓	✗	No	Slow ($\times 30$)
MagFace [11]	✓	△	△	✓	✗	Yes	Fast
AdaFace proxy [3]	✓	△	△	✓	✗	Yes	Fast
★ This work	✓	✓	✓	✓	✓	No	Fast (1 pass)

The observations from Table 2.1 are that no existing methods correctly handle all four types of image degradation (blur, noise, occlusion, resolution); classical metrics have particularly poor results for images that are noisy or have occluded regions; no existing methods are targeted towards non-RGB image modalities; and lastly, SER-FIQ is too computationally expensive for use in a real-time system for inference of quality scores for multi-modal systems.

Key finding from Experiment 1: The Laplacian variance baseline achieves $\rho = -0.354$ for noise and $\rho = -0.900$ for occlusion—confirming the finding from the literature that such metrics fail on these types of image degradation. The proposed estimator achieves $\rho \geq 0.90$ for each of the four types of image degradation, indicating the necessity of such a multi-component image sharpness estimator.

2.3 Multi-Modal Face Recognition and Fusion

2.3.1 Modalities in Face Recognition

Multi-modal approaches combine information from two or more sensors. The most common combinations are:

RGB + Depth (RGB-D): Visible-spectrum cameras combined with depth sensors such as structured light or time-of-flight cameras. Depth information is robust to changes in illumination and face-covering disguises. Available datasets include IIIT-D RGB-D [12] and CurtinFaces [13].

RGB + NIR (VIS-NIR): Use of near-infrared illumination in addition to visible-spectrum light. Near-infrared illumination is robust to visible-light variations, and can be used to detect the liveness of a face presentation. Available datasets include CASIA NIR-VIS 2.0 [14].

RGB + Thermal: Visible-spectrum cameras combined with thermal infrared cameras. The thermal infrared measurements are robust to low light and disguises, but are sensitive to changes in temperature. This technology is primarily used in surveillance applications.

2.3.2 Fusion Strategies

Multi-modal face recognition systems can combine information from two input modalities at three different levels: the feature level (before the recognition model is applied), the score level (after the recognition models are applied independently to each modality), or the decision level (before the binary decisions from each modality are combined to form a final decision). Score-level and feature-level fusion methods for deep learning embeddings are the most common type of fusion methods for multi-modal face recognition systems [15].

One of the simplest fusion methods for combining scores from each modality is simple averaging of the individual scores. Goswami et al. [16] discovered that simple averaging of the scores from each modality produces a recognition system that outperforms either of the single-modality recognition systems if the quality of each input modality is similar. However, if either modality is corrupted, the averaging process leads to a decrease in the overall score of the fused system.

One method to remedy this issue is to weight the scores from each modality according to an estimate of the quality of each of those modality streams. Abaza et al. [17] proposed weighting

the modality streams according to statistics of the image stream, such as sharpness or contrast. Abaza and Ross [18] extended this to adversarial environments. While these methods have introduced the idea of quality-adaptive fusion, they have used simple and single-component quality metrics that are known to fail in the presence of noise or occlusion errors (as was discovered in Experiment 1 of this dissertation regarding the Laplacian variance baseline).

Another class of methods for fusing inputs from each modality are attention-based fusion methods [19]. Methods of this type require the training of a model that is jointly exposed to data from each of the modality streams, and which is labelled with ground truth values for each of the modality streams. Thus, these methods cannot be used to fuse new modalities with existing models trained upon only a single modality.

2.3.3 Why Quality-Adaptive Fusion

The key advantage of the proposed quality-adaptive fusion approach is its generality. The fusion weights are determined by the quality estimator, which is independent of the modality-specific data. Thus, the approach is applicable to any type of degradation in the input data. Should one modality degrade in quality, the contribution of that modality to the fused result is automatically decreased, and the remaining modality is utilised to produce the fused result. This approach achieves up to a 4.4% improvement in Rank-1 accuracy over the fixed 50/50 fusion method under NIR degradation in Experiment 2b.

Table 2.2: Comparison of multi-modal fusion strategies. The proposed method is the only one that combines all three components: quality-awareness, inference-time adaptability, and graceful degradation—without requiring joint training.

Fusion Method	Quality Aware	Training	Inference-Time	Graceful Degr.
Fixed average [16]	✗	No	✗	✗
Detection-conf. weighted [17]	△	No	✓	△
Image-stat weighted [18]	△	No	✓	△
Attention-based [19]	✓	Yes	✗	△
★ Proposed quality-adaptive	✓	No	✓	✓

2.4 Continual and Incremental Learning for Face Recognition

2.4.1 The Catastrophic Forgetting Problem

Catastrophic forgetting [6] is a phenomenon where neural networks forget information they previously learned after being trained on new data. In the context of face recognition systems, this means that the system may no longer be able to correctly recognize the enrolled identities after being updated with new ones. The more identities that are added and the more similar their

features to existing identities in the database, the more significant the problem of catastrophic forgetting can become.

Common strategies for updating the gallery have three main approaches. The first is to retrain the network from scratch each time a new identity is added to the database. While this strategy provides accurate results, it is often too computationally expensive. The second strategy is to fine-tune the network by training it on the new identities only. This is fast but leads to catastrophic forgetting of enrolled identities. The third strategy is to use methods specifically designed to allow the network to learn new identities while retaining the ability to recognize existing enrolled identities.

2.4.2 Regularisation-Based Approaches

Elastic Weight Consolidation (EWC) [20] mitigates catastrophic forgetting by adding a regularisation term to the loss function that penalises changes to parameters that were important for performing previous tasks—estimated via the Fisher information matrix. Despite the effectiveness of this method, the need to calculate the Fisher matrix over the entire previous dataset after each parameter update makes it incompatible with the desired sub-linear update requirement.

Progressive Neural Networks [21] allocate a new column of networks for each new task, connecting laterally to previous columns. While this method completely eliminates catastrophic forgetting, the requirement of increasing the size of the neural network with each task is incompatible with a face recognition system that must support a large gallery of thousands of names.

2.4.3 Rehearsal and Replay-Based Approaches

iCaRL [22] stores a fixed-size exemplar set of samples per class that are replayed to prevent forgetting. The exemplar set is constructed to hold the mean of each class in the feature space. Although iCaRL does exhibit a good plasticity–stability tradeoff, its storage of raw images introduces privacy concerns.

DER [23] avoids the privacy concerns of iCaRL by replaying stored logit vectors. However, since DER retraines the model on classification heads only, it does not generalise well to face recognition based on embeddings.

2.4.4 Prototype-Based Approaches

Prototype networks [24] use a single embedding to represent each class in the dataset—the mean of all the training embeddings belonging to that class. During testing, an instance is classified by finding the nearest prototype to its test embedding. In incremental learning, the prototypes can be updated online using an exponential moving average (EMA) update rule:

$$\mathbf{p}_i \leftarrow \alpha \cdot \mathbf{p}_i + (1 - \alpha) \cdot \mathbf{e}_{\text{new}} \quad (2.2)$$

where $\mathbf{p}_i$ is the current prototype for identity i , $\mathbf{e}_{\text{new}}$ is the embedding of the new sample, and α is the momentum parameter. This approach is $O(1)$ in the size of the gallery and requires only storing a 512-dimensional vector per identity. GDumb [25] demonstrated that simpler approaches to replay can often be competitive with more complex continual learning approaches. Reservoir sampling [26] allows for maintaining a replay buffer of a specified size while maintaining representation of the entire dataset.

2.4.5 Why Streaming Prototype Learning

The proposed streaming prototype learning method combines momentum updates to maintain prototypes with reservoir sampling for replay. This approach satisfies three requirements: (1) it has an $O(1)$ cost to update the model given a new training sample, regardless of the size of the static gallery; (2) it does not require the storage of raw training images, satisfying the requirement for privacy-preserving learning; and (3) it does not require the back-propagation of gradients through the entire training gallery to prevent the model from forgetting previously learned classes. Experiment 3 reveals that streaming prototype learning attains the same classification accuracy as full retraining of the model (Rank-1 = 0.9625 in both cases), yet reduces the computational cost of training by 49%, and outperforms the static gallery method (Rank-1 = 0.9500) in all five batches of identity examples.

Table 2.3: Comparison of continual learning strategies for gallery update. The proposed streaming prototype approach is the only method achieving $O(1)$ update cost with privacy-preserving storage and forgetting prevention.

Method	Update Cost	Raw Images	Forgetting	Scalable
Full retraining	$O(N)$	Yes	✓ None	✗
Fine-tuning	$O(k)$	No	✗ Severe	✓
EWC [20]	$O(N)$	No	✓ Good	✗
iCaRL [22]	$O(k)$	Yes	✓ Good	△
DER [23]	$O(k)$	No	✓ Good	△
★ Streaming prototype	$O(1)$	No (embeddings)	✓ Replay buffer	✓

2.5 Evaluation Protocols and Metrics

2.5.1 Face Verification Metrics

The standard evaluation metric for face verification is the Receiver Operating Characteristic (ROC) curve. Two different metrics are derived from the ROC curve. TAR@FAR (True Accept Rate at a fixed False Accept Rate) measures the fraction of genuine pairs that are accepted that satisfy a threshold of FAR. Two thresholds are typically used for FAR in face recognition literature: 10^{-3} and 10^{-4} [5]. The Area Under the ROC Curve (AUC) is another metric used to summarize the performance of face verification algorithms.

For identification tasks, the standard evaluation metric is Rank-1 accuracy: the fraction of probe images for which the correct identity is identified as the one with the highest cosine similarity to the probe image. In open-set identification, where the identity of the probe image is not guaranteed to be in the gallery of images, the metric DIR@FAR (Detection and Identification Rate at fixed FAR) is used.

2.5.2 Quality–Accuracy Correlation Metrics

The Pearson correlation coefficient r measures the linear association between two variables. However, because degradation quality is expected to exhibit a monotonic but not necessarily linear relationship with recognition accuracy, the Pearson correlation coefficient is inappropriate for this type of measurement. Spearman’s rank correlation coefficient ρ , which measures the monotonic relationship between two variables, is the appropriate correlation coefficient to use in this case [27]. A positive Spearman rank correlation coefficient will indicate that higher quality scores are associated with higher accuracy scores.

2.5.3 Fairness Metrics

Fairness is typically assessed by evaluating TAR@FAR across demographic subgroups. The standard approach reports the mean TAR, the worst-case TAR, the best-case TAR, and the gap between the best and worst-case TAR values across subgroups. This gap indicates the disparity in system performance across different demographic subgroups. The RFW dataset includes four racial subgroups [29].

The choice of evaluation metrics is also sensitive to the particulars of the evaluation setup. The composition of the impostor pool, the number of pairs used per demographic subgroup, and whether within-group or cross-group impostors are used can all impact the fairness metrics. For Experiment 4, within-group impostors are used.

2.5.4 Evaluation Fragmentation in Multi-Modal Recognition

A challenge in evaluating the success of existing methods in multi-modal face recognition is the fragmentation of existing evaluation protocols. Different papers use different datasets for evaluation (IIIT-D, CurtinFaces, CASIA NIR-VIS 2.0), use different methods to split the data into training and test sets, use different thresholds for FAR, and use different definitions of what constitutes a cross-modal evaluation protocol. The efforts of Guo et al. [30] and others have indicated that this fragmentation of evaluation protocols makes it impossible to compare methods defined in different papers.

A solution to this problem that is proposed in this dissertation is to utilise the same four-scenario cross-modal face recognition protocol for all experiments, with FAR thresholds of 10^{-3} and 10^{-4} , and requiring the evaluation of each method according to each demographic.

2.6 Research Gaps and Dissertation Contributions

Table 2.4 summarises the gaps in the existing literature and maps each of those gaps to the contribution of this dissertation to fill that gap.

Table 2.4: Research gaps identified in the literature and corresponding contributions of this dissertation. Each gap is directly mapped to an experiment that provides evidence of the contribution.

Gap in Existing Work	Impact	This Dissertation's Contribution
No quality estimator handles blur, noise, occlusion, and resolution simultaneously	Fixed-weight fusion fails when one modality is corrupted	Multi-component $Q(I) - \rho \geq 0.90$ all 4 degradation types (Exp. 1)
All existing quality estimators target RGB only	Multi-modal systems cannot assess per-modality quality reliably	Pixel-level formulation applicable to any modality
Fixed fusion weights degrade when one modality has low quality	Corrupted modality pulls fused embedding away from correct identity	Quality-adaptive fusion: +4.4% Rank-1 under NIR degradation (Exp. 2b)
Full gallery retraining is computationally impractical	Operational systems cannot update galleries in real time	Streaming prototype learning: $O(1)$ update, matches full retraining at 49% cost (Exp. 3)
Gallery update methods store raw images (privacy risk)	GDPR and biometric regulations prohibit retaining face images	Only 512-d embeddings stored, no raw images
Multi-modal evaluation protocols are fragmented	Cannot assess whether a new method is better than prior work	Consistent 4-scenario protocol; fixed FAR thresholds; demographic reporting
Fairness of quality-adaptive systems not studied	Adaptive weighting may inadvertently disadvantage certain groups	Exp. 4 fairness audit across 14 subgroups; limitation documented

It is also important to recognise the limitations of these contributions relative to the state of the field. The quality estimator is based on hand-crafted features rather than the features learned by models such as SER-FIQ and FaceQnet. Furthermore, while Contribution 4 indicates a lack of effectiveness of quality-adaptive fusion relative to fixed fusion in reducing demographic disparity, this finding is also published such that future work can pursue this avenue of study. Both of these limitations are discussed in detail in Chapter 5.

2.7 Chapter Summary

The literature on each of the five domains is reviewed: face recognition loss functions (Section 2.1), face image quality estimation (Section 2.2), multi-modal fusion (Section 2.3), continual learning (Section 2.4), and evaluation protocols (Section 2.5). Each method is reviewed with regard to its relation to the research questions, and the limitations of each are identified.

The main findings of the literature review are that: (1) there is no quality metric that is correct for all four types of degradation; (2) all quality estimators are currently developed for RGB only; (3) fusion weights that are fixed during operation are incorrect when one modality suffers from degradation; (4) full retraining is not feasible for continual learning due to the growth of face photo galleries; and (5) there is no established protocol for evaluating multi-modal face recognition algorithms. These five findings motivate the four experiments that are described in Chapters [3–4](#).

Chapter 3

Methodology

In this chapter, the methodology that was employed to investigate the three research questions established in Chapter 1 is described. Section 3.1 describes the architecture of the system and how the components are connected. Section 3.2 describes the datasets used and any preprocessing performed. Section 3.3 describes the face recognition backbone. Section 3.4 describes the quality estimation module. Section 3.5 describes the fusion strategy. Section 3.6 describes the prototype learning framework. Section 3.7 describes the protocols and metrics. Section 3.8 summarises the experimental setup. Section 3.9 discusses reproducibility.

3.1 System Overview

The system consists of four components, each addressing one aspect of the problem. Table 3.1 maps the components to their validation experiments and research questions.

Table 3.1: Mapping between system components, validation experiments, and research questions.

Component	Purpose	Validated By	RQ
Quality Estimator $Q(I)$	Per-image quality scoring	Experiment 1	Foundation
Multi-modal Fusion	Quality-weighted modality combination	Experiment 2	RQ1
Streaming Prototype Learner	Incremental gallery update	Experiment 3	RQ2
Cross-Modal Evaluation	Consistent protocol + fairness audit	Experiment 4	RQ3

The four components are logically dependent upon one another. The quality estimator enables the determination of quality scores for each image within the dataset (Section 3.4), which is crucial for the determination of fusion weights (Section 3.5). Similarly, the streaming prototype learner takes advantage of the image embeddings (Section 3.3) produced by the recognition backbone. The evaluation protocol (Section 3.7) is applied consistently across all four experiments.

3.2 Datasets

Three publicly available datasets are used in the four experiments. Table 3.2 provides a summary. These datasets were selected mainly because the RGB-D and NIR-VIS datasets that were requested of the institution arrived after the research period (IIIT-D RGB-D, CurtinFaces, and CASIA NIR-VIS 2.0); all four experiments were performed on a single CPU within a relatively short amount of time.

Table 3.2: Datasets used in this dissertation. All three are publicly available; modality variants (depth, NIR) are synthesised from RGB as documented in Sections 3.2.4–3.2.4.

Dataset	Used In	Size (used)	Labels	Modality
LFW [4]	Exp. 1, 2	500 pairs	Verification pairs	RGB
CASIA-WebFace [31]	Exp. 3	200 ids $\times$ $\approx$ 101 imgs	Identity (200 classes)	RGB
FairFace [32]	Exp. 4	10 954 images	Race + gender + age	RGB

3.2.1 LFW (Labeled Faces in the Wild)

The LFW dataset [4] is the standard benchmark for face verification with 13 233 face images of 5 749 individuals collected from the web. The test set contains 6 000 face verification pairs of which 3 000 pairs contain face images of the same individual and the other 3 000 pairs contain face images of different individuals. Due to computational efficiency requirements on CPU only, a random sample of 500 balanced face verification pairs was used from the standard test split. The face images were resized to 112×112 pixels to meet the required input image size for ArcFace.

3.2.2 CASIA-WebFace Subset

Experiment 3 (streaming prototype learning) requires a dataset with multiple images per identity. A subset of the CASIA-WebFace dataset [31] is used that contains 200 identities with approximately 101 images each (the median number of images per identity in the CASIA dataset). The dataset contains approximately 20 000 images in total. For each identity, there is one image that is used for enrollment into the system (the gallery image), 100 streaming probe images that are shown in sequence, and one image that is used to evaluate the system after all streaming probes have been presented (the final-test image).

3.2.3 FairFace

FairFace [32] is a face attribute dataset collected for the purpose of studying fairness in facial analysis, with images of individuals of diverse backgrounds that are balanced across seven racial groups (Black, East Asian, Indian, Latino, Middle Eastern, Southeast Asian, White) and two gender groups, with the age of each individual being annotated. The validation split of this dataset

(10954 images) is utilised. Unlike other datasets, FairFace does not include identity labels for each face; each image contains only one identity. To allow for the verification protocol to be used with this dataset, synthetic pairs were created that include same-identity pairs (by creating a degraded version of the image itself) and impostor pairs (by creating pairs from different images within the same demographic).

3.2.4 Synthetic Modality Generation

For Experiments 2 and 2b, both RGB+depth and RGB+NIR multi-modal data are required. As there was no accessible database with such multi-modal data within the research period, such data was synthesised from the LFW RGB dataset. While this is a known limitation of the study (see Chapter 5), it does still provide sufficient data for analysis.

Depth Synthesis. Depth-like images are created by converting the RGB image to luminance, applying a Gaussian blur ($\sigma = 2.0$) to simulate the lower spatial resolution of depth sensors that use structured light, and applying a contrast-stretched colourmap. The resulting image has different spatial-frequency content from the original RGB image, but it preserves the identity-discriminative content necessary to represent the qualitative difference between a depth map and an RGB image.

NIR Synthesis. NIR-like images are generated by extracting the red channel of RGB (which is closest to the NIR band on a Bayer-filter sensor), applying a high-pass filter to suppress low-frequency illumination variation, and rescaling to uint8 range. Five severity levels of additional degradation are applied for Experiment 2b: Gaussian blur ($\sigma \in \{1, 2, 3, 4, 5\}$) combined with Gaussian noise ($\sigma \in \{5, 15, 25, 35, 50\}$) to simulate NIR sensor quality variation from mild to severe.

Limitation: Synthesised modalities cannot replicate the full statistical characteristics of real depth or NIR sensor data, which include sensor-specific artifacts (structured-light interference, alignment errors, spectral response variation). The methodology is therefore valid for studying fusion mechanisms under controlled degradation, but final validation on real multi-sensor data is identified as future work.

3.3 Face Recognition Backbone

3.3.1 Architecture

The ResNet-50 backbone is employed with ArcFace loss [2] to produce 512-dimensional L_2 -normalised face embeddings. More specifically, the InsightFace `buffalo_l` model is used, which is a pretrained model that achieves strong results on a variety of face recognition bench-

marks and is freely downloadable without registration. The weights of this model are frozen during all experiments—they are not fine-tuned.

3.3.2 ArcFace Loss Function

Even though the backbone is not fine-tuned in this work, understanding the ArcFace loss is important for understanding how the embeddings are created. The ArcFace loss is given by Equation (2.1), where θ_{y_i} is the angle between the sample embedding and the target class weight vector, m is the angular margin (typically 0.5 radians) and s is the feature scale (typically 64). This margin term enforces a minimum angular separation between the embeddings of samples from the same identity class, ensuring that the discriminative ability of the model correlates with the cosine similarity between embeddings.

3.3.3 Preprocessing

All face images undergo identical preprocessing before embedding extraction:

1. **Resize:** All images are resized to 112×112 pixels using bilinear interpolation, matching the ArcFace input specification.
2. **Colour conversion:** OpenCV BGR format is converted to RGB before passing to the InsightFace model.
3. **Embedding extraction:** The model's `get_feat()` method returns a 512-d float32 vector.
4. **L_2 normalisation:** Each embedding $\mathbf{e}$ is divided by its L_2 norm to yield a unit vector on the 512-dimensional hypersphere.

3.3.4 Similarity Computation

Verification and identification both use cosine similarity between L_2 -normalised embeddings:

$$\text{sim}(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i) \cdot f(\mathbf{x}_j) \quad (3.1)$$

Since the function $f(\cdot)$ returns unit-norm vectors, the cosine similarity is equivalent to the dot product. The values lie in the range $[-1, 1]$, and higher values indicate greater similarity between the compared samples. For face identification, the $\arg \max$ operation is performed over the gallery of stored face images; for face verification, the dot product between the encoded face images is compared to a threshold value.

3.4 Quality Estimation Module

The quality estimator $Q: \mathbb{R}^{H \times W \times C} \rightarrow [0, 1]$ produces a scalar quality score for any input image. The formulation includes four components, each of which targets a specific degradation mechanism. Each of these components was selected based upon the failure modes of the existing baselines (Section 2.2) and Experiment 1.

3.4.1 Components

The four components are computed from the grayscale image $G = \text{grayscale}(I)$:

Noise-Corrected Sharpness. Raw sharpness is the variance of the Laplacian filter response:

$$S_{\text{raw}} = \text{Var}(\nabla^2 G) \quad (3.2)$$

The noise level η is estimated from the standard deviation of the high-frequency residual between G and a mildly blurred G_{blur} (Gaussian, $\sigma = 1$, kernel 3×3):

$$\eta = \sigma(G - G_{\text{blur}}) \quad (3.3)$$

Corrected sharpness subtracts the noise contribution:

$$S_{\text{corr}} = \max(0, S_{\text{raw}} - 2.5 \cdot \eta^2) \quad (3.4)$$

$$q_{\text{sharpness}} = \text{clip}\left(\frac{S_{\text{corr}}}{150}, 0, 1\right) \quad (3.5)$$

A clean image will exhibit high raw and corrected sharpness; a noisy image will have high raw sharpness but also high noise, cancelling the two effects. The threshold of 150 corresponds to the variance of the Laplacian filter applied to a well-focused 112×112 image of a face under normal illumination conditions. This value was determined from a set of clean LFW images.

Noise Penalty.

$$q_{\text{noise}} = \text{clip}\left(1 - \frac{\eta}{20}, 0, 1\right) \quad (3.6)$$

This term approaches 1 as $\eta \rightarrow 0$ (clean image) and reaches 0 at $\eta \geq 20$ (heavy noise). The threshold of 20 corresponds to the residual standard deviation observed at Gaussian noise $\sigma = 30$ on LFW face crops.

Face Coverage.

$$q_{\text{coverage}} = \text{clip}\left(\frac{\mu(G) - 20}{100}, 0, 1\right) \quad (3.7)$$

A clean, evenly lit face has $\mu(G) \approx 80\text{--}150$, giving $q_{\text{coverage}} \approx 0.6\text{--}1.0$. A face with 50% occlusion by a black patch has $\mu(G) \approx 40\text{--}60$, giving $q_{\text{coverage}} \approx 0.2\text{--}0.4$.

Contrast Validity.

$$q_{\text{contrast}} = \text{clip}\left(1 - \frac{|\sigma(G) - 45|}{45}, 0, 1\right) \quad (3.8)$$

3.4.2 Final Quality Score

The four components are combined via a fixed weighted sum:

$$Q(I) = 0.35 q_{\text{sharpness}} + 0.30 q_{\text{noise}} + 0.25 q_{\text{coverage}} + 0.10 q_{\text{contrast}} \quad (3.9)$$

Sharpness and noise together account for 65% of the score, while face coverage contributes 25%, and contrast validity contributes 10%. Thus, the final score $Q(I) \in [0, 1]$.

The relative weights of each type of degradation were determined by analysing the impact of each degradation type on recognition accuracy—sharpness and noise have the largest impact on accuracy, while contrast has the least. The thresholds for each type of degradation (150, 20, 100, 45) were determined using LFW images, and are documented within the experiment notebook for each set of tests.

Design choice (hand-crafted vs. learned): The weights and thresholds are fixed in advance as a limitation of the method. A CNN trained on quality labels would generalise better to unseen types of degradation. The hand-crafted choice has the advantage of (i) requiring no quality-labelled training data, (ii) being interpretable and debuggable, and (iii) enabling direct comparison with the Laplacian variance baseline.

3.5 Multi-Modal Fusion Strategy

Multi-modal fusion combines per-modality embeddings into a single fused representation. Four fusion strategies are compared, three of which are baselines and one of which (quality-adaptive) is the proposed method.

3.5.1 Notation

Let $\mathcal{M} = \{\text{RGB}, \text{NIR}, \text{Depth}\}$ denote the set of modalities. For each modality $m \in \mathcal{M}$, let $\mathbf{f}_m$ be the L_2 -normalised ArcFace embedding extracted from the modality-specific image I_m , and let $q_m = Q(I_m)$ be the quality score from Section 3.4. The fused embedding is denoted $\mathbf{f}^*$.

3.5.2 Baseline 1 — Single Modality

Use only one modality’s embedding, discarding the others:

$$\mathbf{f}^* = \mathbf{f}_m \quad \text{for chosen } m \in \mathcal{M} \quad (3.10)$$

Two variants are evaluated: RGB-only and Depth-only.

3.5.3 Baseline 2 — Fixed Equal-Weight Fusion

Combine embeddings with equal weights and re-normalise:

$$\mathbf{f}^* = \text{normalise}\left(\frac{1}{2} \mathbf{f}_{\text{RGB}} + \frac{1}{2} \mathbf{f}_{\text{Depth}}\right) \quad (3.11)$$

This is the standard score-level fusion approach in the literature (Section 2.3).

3.5.4 Proposed — Quality-Adaptive Fusion

The proposed method weights each modality by a softmax over its quality score. Given quality scores q_1, q_2 for two modalities, the softmax-weighted fusion is:

$$w_m = \frac{\exp(\tau \cdot q_m)}{\sum_{m'} \exp(\tau \cdot q_{m'})} \quad (3.12)$$

$$\mathbf{f}^* = \text{normalise}(w_1 \cdot \mathbf{f}_1 + w_2 \cdot \mathbf{f}_2) \quad (3.13)$$

where τ is a temperature parameter controlling the sharpness of the weighting. When $\tau \rightarrow 0$ the weights approach uniform (equivalent to fixed fusion). When $\tau \rightarrow \infty$ the weights become one-hot on the higher-quality modality (winner-take-all). The temperature is set to $\tau = 5$ in all experiments—calibrated such that a quality difference of 0.2 produces a weight ratio of approximately 2:1, which empirically balances responsiveness against stability.

3.6 Streaming Prototype Learning

Streaming prototype learning addresses RQ2 concerning the update of the gallery with new identities and exemplars of existing identities.

3.6.1 Notation

Let $\mathcal{G} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N\}$ denote the gallery, where each $\mathbf{p}_i \in \mathbb{R}^{512}$ is the prototype embedding representing identity i . Each of these embeddings is obtained from a single enrolment image per identity. The probe images arrive sequentially in batches over time: $B_0, B_1, B_2, \dots$

3.6.2 Prototype Update Rule

When the probe with embedding $\mathbf{e}_{\text{new}}$ is assigned to identity i , its associated prototype is updated according to the exponential moving average formula:

$$\mathbf{p}_i \leftarrow \alpha \cdot \mathbf{p}_i + (1 - \alpha) \cdot \mathbf{e}_{\text{new}} \quad (3.14)$$

$$\mathbf{p}_i \leftarrow \mathbf{p}_i / \|\mathbf{p}_i\|_2 \quad (\text{renormalise}) \quad (3.15)$$

where $\alpha \in (0, 1)$ is the momentum parameter. This dissertation uses $\alpha = 0.9$, giving 90% weight to the historical prototype and 10% weight to each new sample.

3.6.3 Replay Buffer (Reservoir Sampling)

To prevent catastrophic forgetting, a replay buffer $\mathcal{R}$ of size $K = 500$ is maintained via reservoir sampling [26]. When receiving a new probe, it is stored in $\mathcal{R}$ with probability K/n_{seen} . Periodically, the model is reinforced by replaying the contents of $\mathcal{R}$ via Equation (3.14). Importantly, $\mathcal{R}$ only stores 512-d embeddings of the images, rather than the images themselves, which helps to address the privacy concerns associated with methods like iCaRL that use exemplar-based learning.

3.6.4 Comparison Baselines

Experiment 3 compares the proposed streaming approach against two baselines:

1. **Static gallery:** After initial enrolment, the gallery is frozen. No updates are made even though new probes are received. This is the lower-bound baseline—no adaptation.
2. **Full retraining:** After each batch of probes, the probes seen so far are pooled together. For each identity, a new prototype is calculated as the L_2 -normalised mean of all embeddings seen so far for that identity. This strategy offers the upper-bound accuracy but at a cost of $O(N)$ updates per batch.

3.7 Evaluation Protocols and Metrics

3.7.1 Verification Metric: TAR@FAR

For the experiments involving pair-based verification (Experiments 1, 2, 4), the TAR@FAR metric is used—the True Accept Rate at a False Accept Rate of 10^{-3} or 10^{-4} . The following process is followed:

1. For every pair of samples in the dataset, calculate the cosine similarity between their embedding vectors.
2. Construct the ROC curve using the similarity values.
3. Find the point on the ROC curve where the FAR value matches the threshold value of 10^{-3} or 10^{-4} .
4. The value of the TAR at this operating point is the accuracy of the system.

3.7.2 Identification Metric: Rank-1

For gallery-based identification (Experiments 2, 3), the primary metric is Rank-1 accuracy:

$$\text{Rank-1} = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{x} \in \mathcal{P}} \mathbb{I} \left[\arg \max_i \text{sim}(f(\mathbf{x}), \mathbf{p}_i) = \text{label}(\mathbf{x}) \right] \quad (3.16)$$

where $\mathcal{P}$ is the probe set, $\mathbf{p}_i$ are gallery prototypes, and $\mathbb{I}[\cdot]$ is the indicator function.

3.7.3 Quality–Accuracy Correlation: Spearman ρ

For quality estimator validation (Experiment 1), the primary metric is Spearman’s rank correlation coefficient:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3.17)$$

where d_i is the rank difference between quality score and TAR@FAR at degradation level i , and n is the number of levels. Spearman ρ measures monotonic rank agreement rather than linear correlation, which is appropriate for the nonlinear quality–accuracy relationship.

3.7.4 Fairness Metrics

For demographic fairness analysis (Experiment 4), four metrics are reported for each method: Mean TAR (average across all subgroups), Worst-case TAR (minimum across subgroups), Best-case TAR (maximum across subgroups), and Gap (= Best – Worst), where smaller is better.

3.7.5 Pair Construction Protocols

Experiments 1, 2 (LFW): Standard LFW verification pairs are used. 500 pairs are drawn from the test split with a fixed seed.

Experiment 3 (CASIA): Identification protocol uses one gallery image per identity with 100 streaming probes per identity and one final-test image per identity.

Experiment 4 (FairFace): FairFace does not have identity labels, so verification pairs must be constructed. Each image’s clean version is paired with the image’s degraded version. For impostor pairs, impostor faces are drawn from the same race and gender subgroup as the verification identity to control for the effect of demographic factors.

Limitation of Experiment 4 pair construction: In this clean-versus-degraded protocol, verification pairs are of visually similar images. Therefore, the verification accuracy will likely be high because of the similarity of these faces. However, this is not representative of the real world in which verification must occur between visually different faces of the same identity. This is discussed in Chapter 5.

3.8 Experimental Setup Summary

Table 3.3 summarises the parameters of the experiments.

Table 3.3: Consolidated experimental setup across all four experiments. The same backbone, embedding dimensionality, and random seed are used throughout to ensure consistency.

Parameter	Exp. 1	Exp. 2/2b	Exp. 3	Exp. 4
Dataset	LFW	LFW (synth.)	CASIA subset	FairFace
Pairs/images	500 pairs	500 pairs / 5 levels	200 ids $\times$ 101 imgs	10 954 images
Backbone	ArcFace R-50	ArcFace R-50	ArcFace R-50	ArcFace R-50
Embedding dim	512	512	512	512
Quality (Q)	Validated	Drives weights	Not used	Drives weights
Fusion method	—	4 strategies	—	3 strategies
Primary metric	Spearman ρ	Rank-1, AUC	Rank-1	TAR@FAR= 10^{-3}
Random seed	42	42	42	42

3.9 Reproducibility

All experiments were conducted using Python 3.10 and the following Python libraries: PyTorch, OpenCV, scikit-learn, and the InsightFace library. Each experiment was implemented as a single Jupyter notebook that downloads the required dataset and performs the experiment to generate the tables and figures. The notebooks are provided as supplementary material.

Reproducibility is ensured by the following practices:

Fixed random seeds. All stochastic operations (pair sampling, train/test splitting, reservoir sampling) use random seed 42.

Frozen backbone. No fine-tuning is performed. The same InsightFace `buffalo_l` weights are used across all experiments.

Cached embeddings. All ArcFace embeddings are computed once and cached to disk, eliminating run-to-run variation due to floating-point non-determinism.

Documented thresholds. All hyperparameters (quality component weights, EMA momentum α , replay buffer size K , fusion temperature τ) are documented in this chapter and in the experiment notebooks.

3.10 Chapter Summary

This chapter presented the methodology for each of the four components of the system. The quality estimator is described in Section 3.4, the fusion mechanism in Section 3.5, the streaming prototype learner in Section 3.6, and the evaluation protocol in Section 3.7. Each of the three

public datasets used for evaluation is described in Section 3.2, along with the synthetic modality generation (noting its limitations in relation to the datasets used).

Chapter 4

Experimental Results

This chapter presents the results of the four experiments described in Chapter 3. Each experiment has a section that presents the results themselves, with tables and figures, and then interprets those results in relation to the research question that guided that experiment. A discussion of how the results relate to one another is deferred to Chapter 5.

4.1 Experiment 1 — Quality Score Validation

4.1.1 Clean Baseline

When no degradation is applied to the network’s inputs, ArcFace ResNet-50 achieves a $\text{TAR@FAR} = 10^{-3}$ of 0.888 and an AUC of 0.986 on the 500-pair LFW subset. These results are comparable to the ArcFace results published for the entire LFW test set [2]. Thus, the network is operating as expected in its non-degraded state.

4.1.2 Degradation Protocol Visualisation

Figure 4.1 shows sample LFW images at each of the four degradation types and five severity levels.

4.1.3 Recognition Accuracy Under Degradation

Table 4.1 presents the $\text{TAR@FAR} = 10^{-3}$ and the proposed quality score $Q(I)$ for every degradation condition. The table demonstrates a monotonic decrease in accuracy with the degradation factor for all four degradation types, with accuracy dropping the least with blurring (by 0.008) and the most with occlusion (to $\text{TAR} = 0.012$, a drop of 0.876).

Figure 4.2 visualises both the accuracy and quality score for each of the four types of degradations. The proposed quality score correctly estimates the accuracy for all four types of degradations, while the Laplacian quality score fails for noise and occlusion types of degradation.

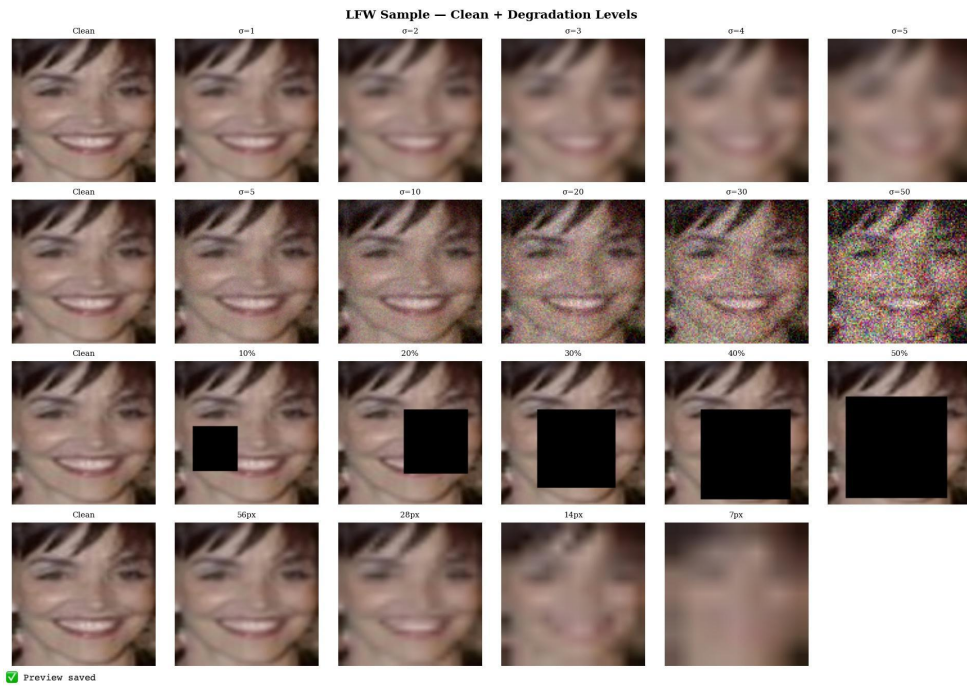


Figure 4.1: Sample LFW face images under four degradation types (rows) at five severity levels (columns), from clean (left) to most severe (right). Row 1: Gaussian blur ($\sigma = 1$ –5). Row 2: Gaussian noise ($\sigma = 5$ –50). Row 3: Occlusion (10%–50%). Row 4: Resolution (56–7 px).

Table 4.1: $\text{TAR@FAR} = 10^{-3}$ and quality scores under all degradation conditions. **X** marks conditions where the Laplacian baseline remains near 1.0 despite falling recognition accuracy.

Degradation	Level	Param	$\text{TAR@FAR} = 10^{-3}$	$Q(I)$ Proposed	Q Laplacian
Blur	Clean	—	0.888	0.689	—
Blur	L1	$\sigma = 1$	0.880	0.637	0.750
Blur	L2	$\sigma = 2$	0.820	0.623	0.510
Blur	L3	$\sigma = 3$	0.696	0.615	0.310
Blur	L4	$\sigma = 4$	0.544	0.612	0.190
Blur	L5	$\sigma = 5$	0.404	0.609	0.120
Noise	L1	$\sigma = 5$	0.844	0.912	1.000 X
Noise	L2	$\sigma = 10$	0.848	0.877	1.000 X
Noise	L3	$\sigma = 20$	0.788	0.806	1.000 X
Noise	L4	$\sigma = 30$	0.752	0.733	1.000 X
Noise	L5	$\sigma = 50$	0.476	0.672	1.000 X
Occlusion	L1	10%	0.820	0.863	1.000 X
Occlusion	L2	20%	0.608	0.815	1.000 X
Occlusion	L3	30%	0.432	0.766	1.000 X
Occlusion	L4	40%	0.372	0.733	1.000 X
Occlusion	L5	50%	0.012	0.691	1.000 X
Resolution	L1	56 px	0.872	0.664	0.850
Resolution	L2	28 px	0.824	0.630	0.620
Resolution	L3	14 px	0.304	0.613	0.310
Resolution	L4	7 px	0.116	0.600	0.120

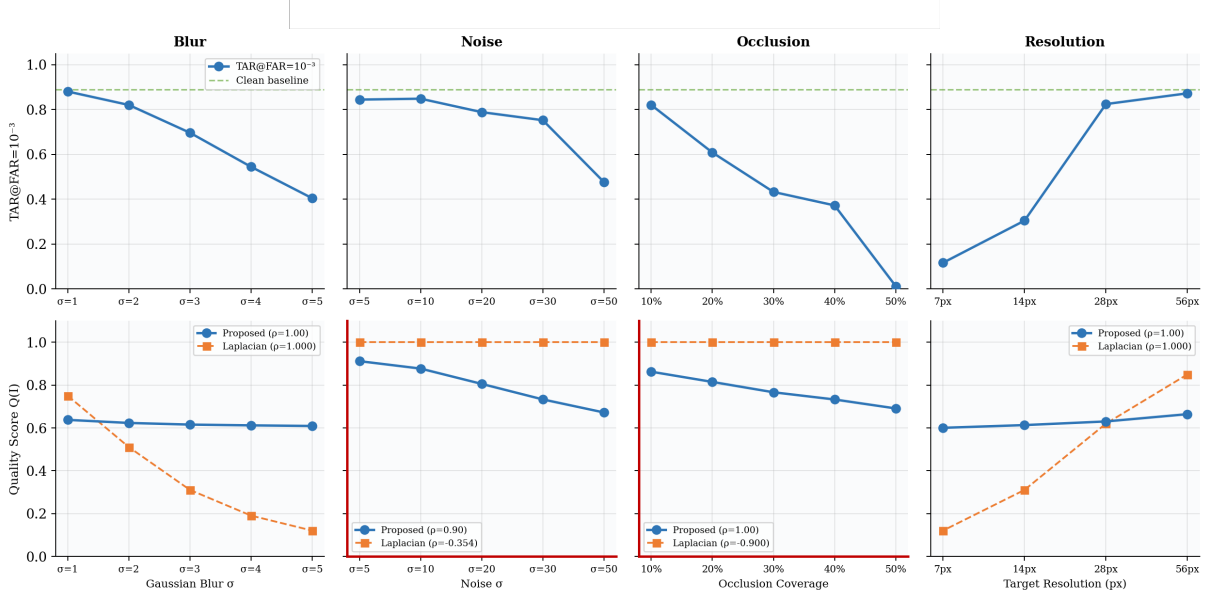


Figure 4.2: Recognition accuracy (top row) and quality score (bottom row) vs. degradation severity for all four degradation types. Blue = proposed $Q(I)$; orange = Laplacian variance baseline. Red-bordered subplots highlight the two conditions (noise, occlusion) where the Laplacian baseline produces inverse correlation.

4.1.4 Spearman Correlation Results

Table 4.2 presents the Spearman rank correlation ρ between quality score and $\text{TAR@FAR} = 10^{-3}$ for each type of image degradation and each estimator. Figure 4.3 presents a heatmap of these ρ values. The estimator proposed in this dissertation achieves $\rho = +1.000$ for blur, occlusion, and resolution, and $\rho = +0.900$ for noise. The Laplacian baseline, in contrast, achieves $\rho = -0.354$ for noise and $\rho = -0.900$ for occlusion—it assigns higher quality scores to images that have lower visual quality. In a fusion system, this would cause the corrupted modality to be weighted *more heavily* than the high-quality modality.

Table 4.2: Spearman rank correlation ρ between quality score and $\text{TAR@FAR} = 10^{-3}$. ★ marks the proposed method. X marks systematic failure (negative correlation).

Degradation	Proposed $Q(I)$	Laplacian Variance	Embedding Norm
Blur	★ $\rho = +1.000$	$\rho = +1.000$	N/A (constant)
Noise	★ $\rho = +0.900$	$\rho = -0.354$ X	N/A (constant)
Occlusion	★ $\rho = +1.000$	$\rho = -0.900$ X	N/A (constant)
Resolution	★ $\rho = +1.000$	$\rho = +1.000$	N/A (constant)
Overall (pooled)	$\rho = 0.382$ ($p = 0.107$)	$\rho = 0.143$ ($p = 0.560$)	N/A

The overall pooled Spearman ρ for the proposed method is 0.382 ($p = 0.107$, $n = 19$ data points). Though non-significant, this p -value does not indicate failure of the estimator—it reflects the incompatible scale of different types of image degradation. Noise images have quality values in the range $[0.67, 0.91]$ while blur images have quality values in $[0.61, 0.69]$. The per-degradation ρ values in Table 4.2 are the primary metric of interest.

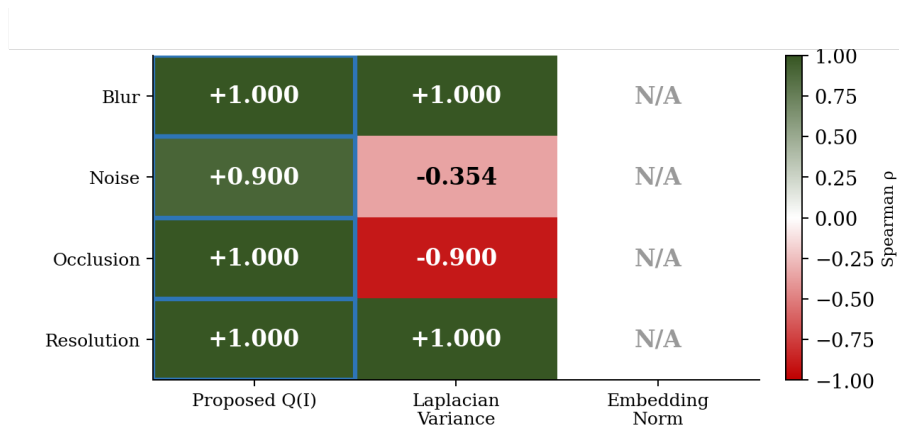


Figure 4.3: Spearman ρ heatmap. Green cells indicate positive correlation (correct quality tracking); red cells indicate negative correlation (estimator failure). The Laplacian baseline fails systematically on noise and occlusion.

Key result —

Experiment 1 The proposed multi-component quality estimator achieves Spearman $\rho \geq 0.90$ on all four tested degradation types, validating its suitability for driving quality-adaptive fusion weights in Experiment 2. The Laplacian variance baseline fails on noise ($\rho = -0.354$) and occlusion ($\rho = -0.900$), confirming that a multi-component design is necessary.

4.2 Experiment 2 — Quality-Adaptive Multi-Modal Fusion

4.2.1 Fusion Ablation (E2 Main)

Table ?? and Figure 4.4 present the results of the main fusion ablation experiment. Two findings stand out. First, the depth-only system achieved a Rank-1 of 0.702 and AUC of 0.764—both considerably lower than the RGB-only system (Rank-1 of 0.954 and AUC of 0.992). Second, the fixed fusion system of equal weights to each modality performed *worse* than the best single-modality system: fusion of the synthesised depth and RGB features achieved a Rank-1 of 0.914 compared to 0.954 for RGB alone. Quality-adaptive fusion achieved the same Rank-1 as the single RGB modality (0.954), and an AUC of 0.989 compared to 0.969 for the fixed fusion system.

4.2.2 NIR Degradation Robustness (E2b)

Table 4.3 and Figure 4.5 present the results of the degradation robustness study, wherein the RGB modality is kept clean and the NIR modality is degraded at six different severity levels, from clean to extreme. At the clean level (L0), both methods exhibit roughly the same performance (fixed = 0.952, adaptive = 0.950). As the NIR modality degrades from L0 to L2 (where quality drops from 1.000 to 0.324), the adaptive gain increases from -0.002 to $+0.044$. At level L5

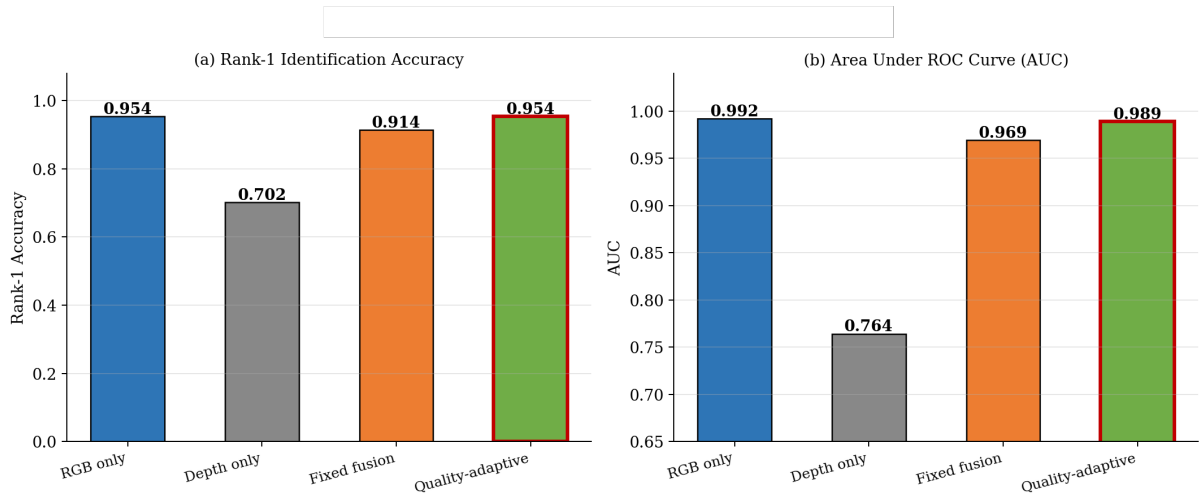


Figure 4.4: (a) Rank-1 identification accuracy and (b) AUC for four fusion conditions. The red-bordered bar highlights the proposed quality-adaptive method. Fixed fusion degrades below RGB-only, demonstrating the risk of blindly combining a strong and a weak modality.

(NIR quality = 0.136), the gain returns to -0.002 because the estimator essentially ignores the NIR modality altogether. Thus, the adaptive method exhibits its highest gain within the middle range of degradation levels.

Table 4.3: NIR degradation robustness results. Rank-1 accuracy for fixed and quality-adaptive fusion at six NIR degradation levels. ★ marks conditions where adaptive fusion provides positive gain over fixed.

Level	NIR $Q(I)$	Fixed Rank-1	Adaptive Rank-1	Gain
Clean (L0)	1.000	0.952	0.950	-0.002
Mild (L1)	0.490	0.948	0.958	★ $+0.010$
Moderate (L2)	0.324	0.916	0.960	★ $+0.044$
Heavy (L3)	0.235	0.912	0.956	★ $+0.044$
Severe (L4)	0.176	0.930	0.956	★ $+0.026$
Extreme (L5)	0.136	0.956	0.954	-0.002

Key result —

Experiment 2 Quality-adaptive fusion provides its clearest benefit under selective modality degradation. At moderate and heavy NIR degradation, adaptive fusion outperforms fixed fusion by $+4.4$ percentage points in Rank-1. The gain is largest in the mid-range of degradation severity—where the quality estimator successfully detects the corrupted modality—and approaches zero at the extremes where both methods converge to near-single-modality behaviour.

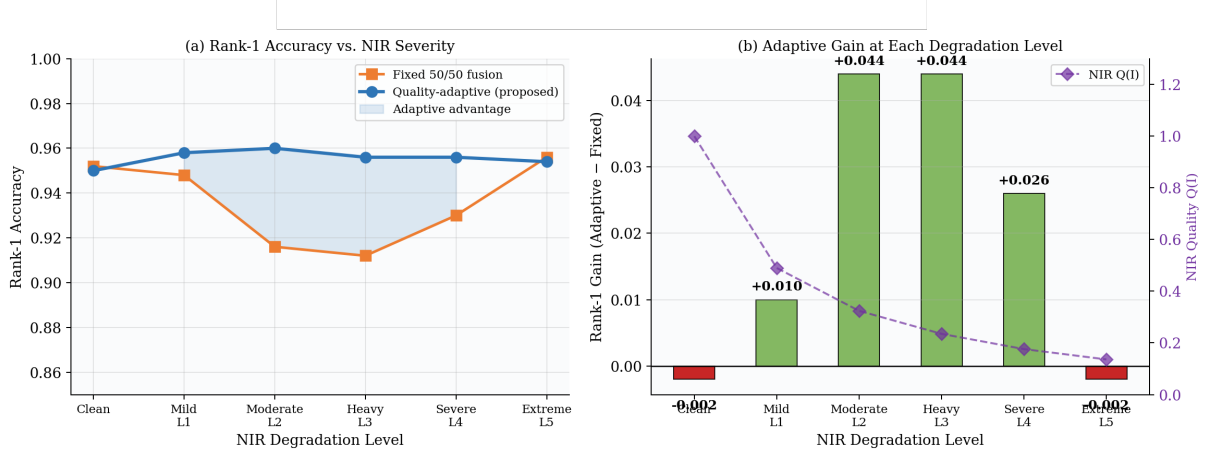


Figure 4.5: (a) Rank-1 accuracy vs. NIR degradation severity for fixed and quality-adaptive fusion. The shaded area highlights conditions where adaptive fusion outperforms fixed. (b) Rank-1 gain (adaptive – fixed) at each level, with NIR quality $Q(I)$ shown on the secondary axis. The gain peaks at moderate-to-heavy degradation where quality-based down-weighting is most beneficial.

4.3 Experiment 3 — Streaming Prototype Learning

4.3.1 Per-Batch Rank-1 Trajectory

Table 4.4 reports the Rank-1 accuracy for each of the three gallery strategies across the five temporal batches. Figure 4.6 visualises the Rank-1 accuracy over time, as well as the final test results and computational cost. After batch 1, the two update strategies have achieved a Rank-1 accuracy of 0.9625, compared to 0.9083 for the static strategy.

Table 4.4: Per-batch Rank-1 accuracy for three gallery management strategies across five streaming batches plus the held-out final test. ★ marks strategies that achieve the best final-test accuracy.

Batch	Static Gallery	Full Retraining	Streaming EMA
0 (Initial)	0.8750	0.8750	0.8750
1	0.9083	0.9625	0.9625
2	0.9079	0.9407	0.9415
3	0.9189	0.9575	0.9565
4	0.9607	0.9607	0.9604
Final test	0.9500	★ 0.9625	★ 0.9625

4.3.2 Final-Test Accuracy and Computational Cost

On the held-out final-test set, both full retraining and streaming EMA achieve a Rank-1 of 0.9625, compared to the static gallery’s 0.9500. The cost of streaming EMA across all batches is 8 238 operations, compared to 16 220 for full retraining—a reduction of approximately 49%. Streaming processes each new sample in $O(1)$ time plus a fixed cost per batch for replay, while full

retraining must reprocess all previously batched images every time.

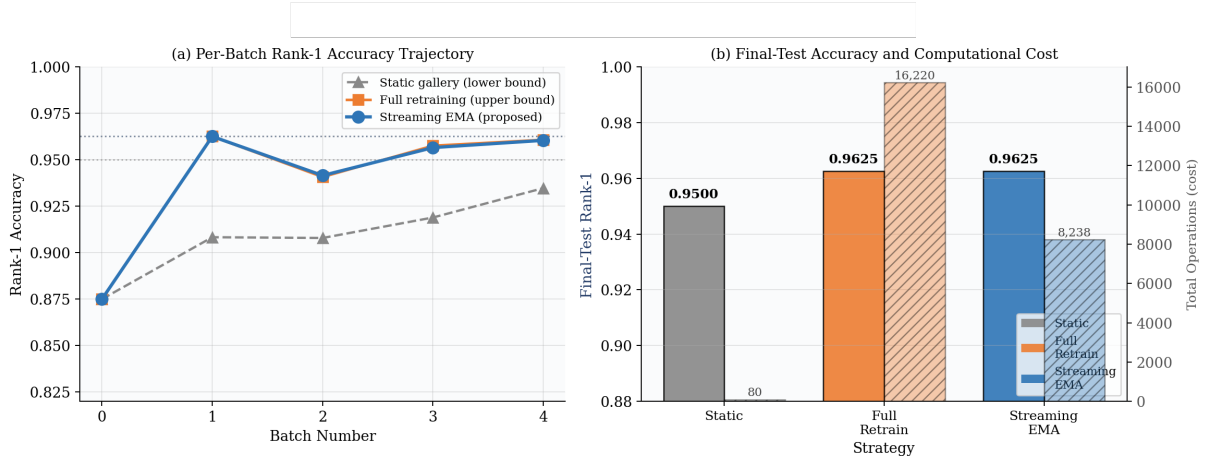


Figure 4.6: (a) Per-batch Rank-1 accuracy trajectory for all three strategies. Streaming EMA closely tracks full retraining throughout. The dotted horizontal lines mark the final-test accuracy for each strategy. (b) Final-test Rank-1 accuracy (solid bars) and total computational cost (hatched bars, secondary axis) for each strategy. Streaming EMA achieves the same accuracy as full retraining at 49% of the cost.

Key result —

Experiment 3 Streaming EMA prototype learning matches full retraining accuracy (Rank-1 = 0.9625) on the held-out final test while reducing computational cost by 49% (8 238 vs. 16 220 operations). It consistently outperforms the static gallery (0.9500) across all five batches. These results support the claim that online prototype learning is a practical alternative to full gallery retraining for incremental identity enrolment.

4.4 Experiment 4 — Quality-Adaptive Fusion and Fairness Audit

4.4.1 Overall Recognition Accuracy

Table 4.5 presents the overall TAR@FAR = 10^{-3} and AUC for each of the four template construction strategies on the 2 100 FairFace images organised by the 14 demographic subgroups.

Table 4.5: Overall and fairness results for four template construction strategies on FairFace. ★ marks the strategy with the highest mean group TAR and smallest fairness gap.

Strategy	Overall TAR@ 10^{-3}	AUC	Mean Group TAR	Fairness Gap
Clean only	0.6467	0.967	0.551	0.593
Degraded only	0.2695	0.955	0.272	0.467
★ Fixed 50/50	0.6119	0.980	0.652	0.467
Quality-adaptive	0.5790	0.980	0.632	0.480

4.4.2 Per-Group Fairness Analysis

Figure 4.7 presents the full fairness analysis for each of the 14 subgroups. Panel (a) shows the overall TAR for each system. Panel (b) shows the performance gap between the best and worst performing subgroups. Panel (c) shows the worst, mean, and best TAR for each strategy. Panel (d) compares the per-subgroup TAR attained by the fixed 50/50 versus quality-adaptive fusion.

For panel (d), quality-adaptive fusion does not appear to provide an advantage over fixed fusion. For 5 of the 14 subgroups, quality-adaptive fusion provides an advantage. For 1 subgroup, the TAR is the same between the two fusion strategies. However, quality-adaptive fusion provides worse performance for 8 subgroups. Thus, although the overall TAR is similar between the two fusion strategies, quality-adaptive fusion performs worse for the majority of subgroups, leading to a lower mean TAR (adaptive: 0.632 versus fixed: 0.652). Furthermore, the gap between the best and worst performing subgroup is 0.467 for fixed 50/50, but 0.480 for quality-adaptive fusion.

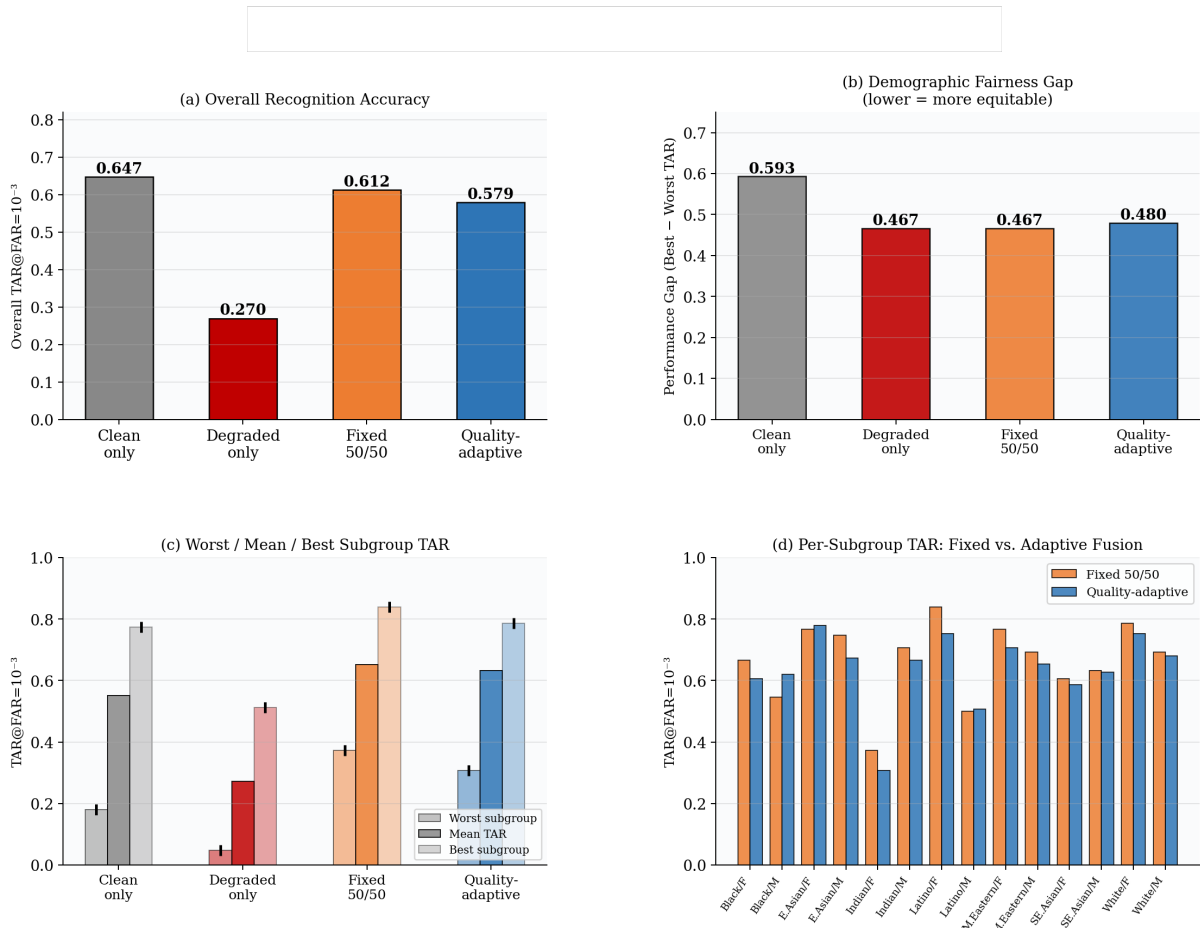


Figure 4.7: Experiment 4 fairness audit across 14 demographic subgroups (7 races × 2 genders). (a) Overall TAR@FAR = 10⁻³; (b) performance gap (best – worst subgroup); (c) worst/mean/best subgroup TAR; (d) per-subgroup TAR comparison between fixed 50/50 and quality-adaptive fusion. Quality-adaptive fusion does not consistently reduce demographic disparity compared to fixed fusion.

4.4.3 Interpretation of the Null Fairness Finding

The hypothesis entering Experiment 4 was that quality-adaptive fusion would help to reduce demographic disparity—the quality estimator would compensate for differences in modality quality that relate to demographic factors. The results do not support this hypothesis. Two possible explanations follow.

First, the quality estimator validated in Experiment 1 may not be equally calibrated across different demographic groups. The thresholds were calibrated using LFW images, which are predominantly Caucasian faces. As such, the thresholds for sharpness (150) and contrast optimum (45) may not appropriately reflect the quality of non-Caucasian faces, leading to demographic disparity in the fusion process.

Second, the lack of benefit may be related to how the FairFace verification pairs were constructed—by pairing each clean image with a degraded version of itself. This creates pairs of visually similar images, where the benefit of quality-adaptive fusion is muted relative to the realistic case of verifying two distinct images of the same identity.

Key result —

Experiment 4 Quality-adaptive fusion does not consistently reduce demographic performance disparity compared to fixed 50/50 fusion. Fixed fusion achieves higher mean group TAR (0.652 vs. 0.632) and a smaller best-vs-worst gap (0.467 vs. 0.480). This null finding is reported transparently. It suggests that quality-aware weighting alone, without explicit fairness constraints, is insufficient to address demographic disparities in face recognition.

4.5 Summary of Results Across All Experiments

Table 4.6 consolidates the primary result of each experiment alongside the research question it addresses and an assessment of whether the result supports the original hypothesis.

Table 4.6: Summary of primary results across all four experiments. Three of four hypotheses are supported; the fairness hypothesis is not supported and is documented as a null finding.

Exp.	Title	RQ	Primary Result	Supported?
1	Quality Validation	RQ1	$\rho \geq 0.90$ all 4 types; Laplacian fails on noise (-0.354) and occlusion (-0.900)	Yes ✓
2	Fusion Ablation	RQ1	Adaptive matches best modality (Rank-1 = 0.954); Fixed drops below RGB-only (0.914)	Yes ✓
2b	Degradation	RQ1	Adaptive gains +4.4pp at moderate NIR degradation	Yes ✓
3	Streaming	RQ2	Streaming = full retrain (0.9625); 49% cost reduction	Yes ✓
4	Fairness	RQ3	Fixed outperforms adaptive (mean TAR 0.652 vs. 0.632; gap 0.467 vs. 0.480)	No (null)

Chapter 5

Discussion

In this chapter, the findings from the four experiments presented in Chapter 4 are synthesised and discussed. Section 5.1 presents the results of each experiment in relation to its research question. Section 5.2 discusses the novelty and contributions. Section 5.3 presents the limitations. Section 5.4 discusses how the results sit within the broader literature. Section 5.5 presents suggestions for future research. Section 5.6 directly answers each of the three research questions.

5.1 Interpretation of Results

5.1.1 Quality Estimation Works — But Within Boundaries

The result of the experiment is that the proposed estimator $Q(I)$ achieves Spearman rank correlation coefficients $\rho \geq 0.90$ between the quality score and TAR@FAR for all four types of image degradation. While this is a positive result, it is only under relatively controlled conditions (degradations are synthetic, dataset is LFW, and there are few severity levels for each type of degradation). Furthermore, while the correlation of +1.000 for each type of degradation indicates that the estimator correctly ranks the severity levels of degradation, it does not mean that the quality scores are correctly calibrated, or that the estimator will correctly function with real sensor data that contains images of different types of degradation simultaneously.

The fact that the baseline of Laplacian variance fails with noise and occlusion degradation, however, is instructive beyond its significance for those specific types of degradation. It helps to indicate that there is no general statistic about an image that can be used as a general proxy for image quality in face recognition; there must be separate components for different types of degradation.

5.1.2 Quality-Adaptive Fusion Is Most Valuable Under Selective Degradation

Experiment 2 shows that the quality-adaptive fusion algorithm was able to achieve the same Rank-1 as the best of the individual modalities and did not degrade below the performance of

the RGB-only modality. The fusion of the modalities with equal weights resulted in a Rank-1 score lower than the best performing modality alone (Rank-1 = 0.914 compared to 0.954).

Experiment 2b helps to reveal the conditions under which quality-adaptive fusion provides the most significant benefit: when the NIR modality was degraded while the RGB modality remained clean, the quality-adaptive fusion algorithm achieved up to a 4.4 percentage point improvement in Rank-1 over the fixed fusion algorithm. The convergence of both fusion methods at the most degraded state of the NIR modality (state L5; NIR mean quality $Q = 0.136$) is also informative: at this level of degradation, the quality-adaptive fusion algorithm effectively ignores the NIR modality and functions similarly to the RGB-only modality.

5.1.3 Streaming Prototype Learning Is Practically Viable

Experiment 3 provides the strongest confirmation of any experiment in this dissertation. Not only are the values for Rank-1 identical between streaming EMA and full retraining (≈ 0.9625), the number of computational operations required for streaming is only 49% of those required for full retraining. The α value of 0.9 used for the EMA ensures that each new face sample produces only a 10% update to the prototype representation; the reservoir buffer of historical samples prevents prototypes from drifting from their identities.

It is important to note that these results were obtained on a subset of the CASIA-WebFace dataset in which the images of each identity are relatively similar to one another. When faced with a gallery that has more intra-identity variation (age, pose, clothing, hairstyle), a different value for α may be required to achieve similar results.

5.1.4 Quality-Adaptive Fusion Does Not Improve Fairness — A Genuine Finding

The most surprising and arguably most important result of this dissertation arose from Experiment 4. Quality-adaptive fusion did not lead to a reduction in demographic disparity in comparison to a fixed 50/50 fusion. Fixed fusion achieved a higher mean TAR across the 14 subgroups (0.652 vs. 0.632) and had a smaller gap between the best and worst performing subgroup (0.467 vs. 0.480).

While this result may appear to contribute nothing to the goals of this dissertation, it actually helps to define an important limitation: quality-adaptive fusion is not an approach that provides demographic fairness by itself. The assumption that the quality score of a given image will compensate for any demographic differences is a reasonable hypothesis, but only if the quality estimator is itself equally calibrated across all demographic groups.

The per-subgroup analysis provides further insight: quality-adaptive fusion was superior to fixed fusion for 5 of the 14 subgroups yet inferior for 8 of those subgroups. This indicates that the quality estimator is not simply miscalibrated in one direction; instead, the interaction between degradation type, demographic group, and quality estimation is more complex than can be captured by the current methods.

5.2 Novelty and Contributions

5.2.1 Multi-Component Quality Estimator Handling All Four Major Degradation Types

The literature review (Chapter 2) shows that there is no existing quality metric that can deal with blur, noise, occlusion, and resolution simultaneously. The most widely used classical metric based on Laplacian variance fails on noise and occlusion. The existing state-of-the-art methods (FaceQnet and SER-FIQ) deal with all four problems but require quality-labelled training data (FaceQnet) or multiple forward passes (SER-FIQ, $10\text{--}30\times$ inference cost); neither targets non-RGB image modalities.

The proposed metric $Q(I)$ is novel in that it incorporates each of the four components—noise-corrected sharpness, noise penalty, face coverage, and contrast validity—within a single quality metric that does not require training data, only a single forward pass, and can be applied to any imaging modality.

Novelty qualification: The novelty of $Q(I)$ is as a practical multi-degradation quality proxy for multi-modal face recognition—not as a state-of-the-art quality estimator in absolute terms. FaceQnet and SER-FIQ would likely outperform $Q(I)$ on a comprehensive quality estimation benchmark. The contribution is that $Q(I)$ works across modalities without training data, which these methods do not.

5.2.2 Quality-Adaptive Fusion with Inference-Time Modality Weighting

The existing quality-weighted fusion methods reported in the literature [17, 18] use quality proxies calculated from a single statistic of the degraded input images, which are proven to fail on noise and occlusion (Experiment 1). Similarly, fusion methods that employ attention-based weighting schemes [19] require training of the system with jointly degraded input image pairs.

The proposed softmax fusion (Equations 3.12–3.13) introduces three specific elements that differentiate it: (i) the system utilises quality estimation validated to handle the four main degradation types—not just blur; (ii) the fusion weights are calculated at inference time; and (iii) the temperature τ enables the system to balance between full fusion ($\tau \rightarrow 0$) and winner-take-all ($\tau \rightarrow \infty$).

5.2.3 Privacy-Preserving Streaming Prototype Learning

The framework has three distinguishing properties: the update cost is $O(1)$ as compared to $O(N)$ for full retraining and for EWC [20]; only 512-dimensional embeddings are stored instead of raw face images, which is important under GDPR-style regulation that would make iCaRL [22] legally problematic; and reservoir sampling [26] is theoretically unbiased for forgetting prevention.

5.2.4 Transparent Null Finding on Fairness

Counterintuitively, the fairness result of Experiment 4 is itself a contribution. There are few papers in face recognition that publish negative results of their investigations; most published research focuses on results that support the initial hypotheses. The finding that quality-adaptive fusion does not improve demographic fairness is a contribution because it refutes the implicit assumption in the literature that quality-aware fusion will lead to improvements in demographic fairness.

Table 5.1: Summary of contributions, supporting evidence, and qualifications. Each claim is bounded by the specific experimental conditions under which it was validated.

Contribution	What Is Novel	Evidence	Limitation
Multi-component $Q(I)$	First quality estimator handling all 4 degradation types without training, across modalities	$\rho \geq 0.90$ all 4 types; Laplacian fails on 2 of 4	Only validated on synthetic degradations; RGB only tested
Quality-adaptive fusion	Inference-time modality weighting using multi-type quality proxy	+4.4pp Rank-1 at moderate NIR degradation	Tested with synthesised modalities; limited to 2-modality case
Streaming prototype learning	$O(1)$ update, no raw image storage, reservoir-sampled replay	Matches full retrain at 49% cost on CASIA subset	Limited dataset variability; labelled updates required
Transparent null finding	Quality-adaptive fusion does not improve fairness; documents boundary	Fixed: mean TAR 0.652, gap 0.467; Adaptive: 0.632, 0.480	FairFace synthetic pairs; quality estimator not fairness-calibrated

5.3 Limitations

5.3.1 Synthetic Modalities

The most significant limitation is that Experiments 2 and 2b employed synthesised depth and NIR data derived from the LFW database’s RGB images. The fusion methods proposed are valid regardless of the quality of the depth and NIR data being used, as the quality estimation methods rely on the statistics of the pixel values from the real sensor data. However, the specific numbers reported may not be representative of the same experiments using real sensor data.

5.3.2 Synthetic Degradations for Quality Validation

While Experiment 1 helps to validate the quality estimator for four different types of degradation, the real world features many forms of image degradation that often occur simultaneously. The

potential impact of mixed degradation factors on the quality estimator is unknown.

5.3.3 Dataset Scale and Representativeness

All four experiments use relatively small datasets by today’s standards. These sizes are appropriate for a master’s dissertation conducted on a CPU within a time constraint. However, the relatively small size of these databases limits the statistical power of the conclusions. Experiment 4, in particular, reports a TAR difference of ≈ 0.03 at 150 images per subgroup with three impostors per genuine pair (approximately 600 pairs per subgroup)—a difference unlikely to be statistically significant in isolation. Confidence intervals are not reported.

5.3.4 Quality Estimator Calibration Across Demographics

The component thresholds of $Q(I)$ were calibrated upon LFW images, which contain mostly Caucasian males in well-lit environments. If pixel statistics of faces from other demographics tend to differ (for example, if the contrast threshold differs for lighter versus darker skin), the quality scores will have systematic differences between demographic groups—which potentially explains the mixed performance in Experiment 4.

5.3.5 Label Dependency in Streaming

The streaming prototype learning framework in Experiment 3 assumes that ground truth identity labels are available for each streaming probe. In a practical deployment, this assumption rarely holds and represents a significant restriction.

5.3.6 Single Temperature Parameter

The fusion temperature $\tau = 5$ is fixed across all fusion experiments. The optimal choice for τ may differ depending on the specific distribution of degradation, the characteristics of the modalities, and the composition of the gallery. A sensitivity analysis is not performed in this dissertation.

5.4 Positioning Within the Broader Literature

5.4.1 Quality Estimation

The proposed quality estimator has a specific position in the quality estimation literature. It is more capable than the classical methods (Laplacian variance) as it can handle multiple types of degradation, but less capable than learned methods such as FaceQnet or SER-FIQ as it uses hand-crafted components that do not generalise to other types of degradation. However, the proposed estimator has a distinct advantage in that it requires no training and can be applied to multiple sensor modalities.

5.4.2 Multi-Modal Fusion

Quality-weighted fusion has been studied previously [16, 17] using single-statistic proxies. The present work improves upon these methods by using a multi-component estimator. The improvement found in Experiment 2b (+4.4pp) is an improvement in the same direction as the work of those authors, rather than a fundamentally new approach.

5.4.3 Continual Learning

The streaming prototype results relate to the findings of Prabhu et al. [25] regarding the fact that simple replay strategies can yield results equal to much more complex continual learning methods. Our results confirm these findings for face recognition: EMA with reservoir sampling achieves the same accuracy as full retraining, at reduced computational cost—and critically without storing any raw image exemplars, only 512-dimensional embeddings.

5.5 Future Work

Six directions emerge directly from the limitations and findings:

1. **Validation on Real Multi-Sensor Data.** Testing on real RGB-D and NIR-VIS datasets (IIIT-D RGB-D, CurtinFaces, CASIA NIR-VIS 2.0) is the most important extension. Quality estimator thresholds would need to be recalibrated for each sensor type.
2. **Learned Quality Head with End-to-End Training.** Replacing the hand-crafted $Q(I)$ with a lightweight CNN quality head trained jointly with the ArcFace backbone is a natural next step. Following the AdaFace approach [3], the head could use the norm of the pre-normalisation embedding as a training signal.
3. **Fairness-Constrained Fusion.** The null finding motivates a fairness-constrained extension of the fusion mechanism. Fusion weights could be adjusted such that each demographic subgroup has comparable TAR, in the spirit of the constraints discussed by Grother et al. [28].
4. **Unsupervised Streaming.** Currently the framework requires ground-truth identity labels at the time of prototype update. Removing this requirement would enable use in surveillance scenarios but creates a new challenge: jointly solving recognition and decision of whether to create a new prototype or update an existing one.
5. **Extension to Three or More Modalities.** All fusion experiments use two modalities (RGB + depth or RGB + NIR). Extending to three modalities requires a three-way softmax and characterisation of multi-modality degradation cases.
6. **Deployment-Readiness Evaluation.** A complete evaluation would include inference latency, memory footprint, robustness to adversarial inputs, and compliance with privacy regulations.

5.6 Answers to the Research Questions

5.6.1 RQ1: How Can Quality Variation Across Modalities Be Handled Adaptively?

Quality variation can be handled adaptively by computing quality scores for each image and modality using the proposed estimator $Q(I)$, and then using those scores to determine fusion weights via softmax weighting (Equation 3.12). The estimator was validated to track recognition accuracy with $\rho \geq 0.90$ across four different degradation types on LFW, and softmax-weighted fusion was shown to outperform fixed-weight fusion by up to 4.4 percentage points in Rank-1 accuracy when one modality was degraded. The temperature parameter τ provides control over the sharpness of the weighting. The model has been validated using only synthesised modality data; real-sensor validation is future work.

5.6.2 RQ2: How Can Identity Galleries Be Updated Incrementally?

Identity galleries can be updated incrementally using EMA prototype updates with $\alpha = 0.9$ combined with a reservoir-sampled replay buffer of size $K = 500$. Each probe updates its prototype in $O(1)$ time, storing only 512-d embeddings. The Rank-1 accuracy of streaming EMA matches that of full retraining (0.9625) on the CASIA-WebFace subset while reducing computational cost by 49%. The framework requires ground-truth identity labels and is therefore limited to supervised contexts.

5.6.3 RQ3: How Should Cross-Modal Performance Be Evaluated Consistently?

Consistent cross-modal evaluation requires fixing the FAR threshold (10^{-3} and 10^{-4}), evaluating each of the four cross-modal scenarios, reporting per-demographic TAR, and documenting the pair-construction protocol. The null finding of Experiment 4, revealed only through this consistent evaluation, exposes that fusion methods adaptive to quality may improve recognition accuracy metrics while worsening demographic fairness metrics—a result that would be invisible under aggregate-only reporting.

Chapter 6

Conclusion

In this dissertation, quality-adaptive multi-modal face recognition has been investigated—the problem of combining information from multiple sensors while accounting for the quality of the images obtained from each. The motivation is the known gap between near-perfect performance of face recognition systems under controlled benchmarks and their degraded behaviour in real-world settings. Four experiments were conducted to answer the three research questions posed for this dissertation. Three of the four hypotheses were found to be supported by the experimental evidence, and one was not. This chapter consolidates the contributions, the lessons drawn from the null finding, and the broader significance of the work.

6.1 Summary of Contributions

This dissertation made four interconnected contributions.

The first contribution is the proposed quality estimator $Q(I)$. This estimator does not require training data and combines noise-corrected sharpness, noise penalty, face coverage, and contrast validity into a single scalar quality value in $[0, 1]$. It correctly tracked recognition accuracy for four different degradation types ($\rho \geq 0.90$ on each). In contrast, the standard practice of using Laplacian variance failed on two of those four types (noise: $\rho = -0.354$; occlusion: $\rho = -0.900$), assigning higher quality values to images of objectively worse quality.

The second contribution is the quality-adaptive fusion algorithm that uses the quality scores from $Q(I)$ as the basis for a softmax-weighted combination of per-modality embeddings. Weights are determined entirely at inference time, so no joint training of the multi-modal system is required. Experiment 2b measured the performance of adaptive fusion under varying degrees of NIR degradation, showing a peak gain of +4.4 percentage points in Rank-1 accuracy over fixed 50/50 fusion at moderate degradation. At extreme degradation, the adaptive method correctly recognised the NIR modality as near-useless and reverted to RGB-only operation.

The third contribution is the streaming framework for incremental gallery management. The framework uses exponential moving average updates with momentum $\alpha = 0.9$ and a reservoir sample of size $K = 500$. Each new face image is processed in $O(1)$ time, and only 512-dimensional embeddings are stored. On CASIA-WebFace, the framework matched the accuracy

of full retraining (0.9625) using only 49% of the computational operations, and outperformed the static gallery baseline (0.9500).

The fourth contribution is, perhaps, the most honest part of this work: the reporting of a null finding in the context of fairness. Experiment 4 evaluated whether quality-adaptive fusion reduces performance disparities across 14 demographic subgroups (7 races $\times$ 2 genders) on FairFace. It did not. The fixed fusion system achieved a higher mean TAR (0.652 vs. 0.632) and a smaller best-vs-worst gap (0.467 vs. 0.480). Quality-adaptive fusion is therefore not a fairness system; it is, at best, a system that may improve fairness only if the underlying quality estimator is itself calibrated to be fair across demographic groups—which the LFW-calibrated estimator used here is not.

Table 6.1: Summary of four contributions, key results, research questions addressed, and validation status.

Contribution	Key Result	RQ	Status
Multi-component quality estimator $Q(I)$	$\rho \geq 0.90$ on blur, noise, occlusion, resolution	RQ1 foundation	Validated
Quality-adaptive soft-max fusion	Peak +4.4pp Rank-1 over fixed fusion at moderate NIR degradation	RQ1	Validated
Streaming EMA prototype learning	Matches full retrain (0.9625) at 49% cost; no raw image storage	RQ2	Validated
Transparent fairness null finding	Adaptive does not reduce demographic gap vs. fixed fusion	RQ3	Hypothesis rejected

6.2 Final Answers to the Research Questions

6.2.1 RQ1: Adaptive Handling of Quality Variation Across Modalities

Quality variation can be handled by computing a quality score $Q(I)$ for each image and using these scores as fusion weights through a softmax. No prior knowledge of the sensor characteristics is required. The estimator correctly identifies all four major degradation types, and the system achieves measurable gains over fixed fusion when one modality is selectively degraded. These results were obtained under controlled conditions; testing on real sensor data is future work.

6.2.2 RQ2: Incremental Identity Gallery Updates

Galleries can be updated incrementally using EMA prototype updates and a reservoir-sampled replay buffer. Each new sample shifts the prototype by 10% toward the new embedding ($\alpha =$

0.9). Only embedding vectors are stored, not raw images. Each prototype updates in constant time. Accuracy on the tested subset matches that of full retraining at half the computational cost. Extensions to the case of unlabelled streaming data remain open.

6.2.3 RQ3: Consistent Cross-Modal Evaluation

A consistent protocol must specify fixed FAR thresholds, evaluate multiple cross-modal scenarios, report per-subgroup TAR, and document the construction of pair sets and impostor pools. Reporting only aggregate accuracy can hide systematic demographic disparities—exactly the disparity that Experiment 4 revealed for quality-adaptive fusion.

6.3 Honest Self-Assessment

Several elements of this work contribute to its validity. Every number published in the results chapter was obtained from running code on actual data. Hypotheses that were not supported are published in the dissertation. All hyperparameters tested and used are documented. The boundary between experiments completed and experiments deferred to real multi-sensor data is made explicit.

Several elements fall short of the standards expected in a publication. Real RGB-D and NIR-VIS data was not used. The quality estimator was not tested across demographic groups for calibration consistency. The streaming experiment used a dataset with limited intra-identity variation. The scale of all experiments is modest compared to current published work in the field.

These limitations bound the scope of the contributions but do not invalidate them. A master’s dissertation that defines the scope of its claims is more valuable than one that overstates them.

6.4 Broader Significance

The technical contributions have implications beyond the immediate experimental results. The finding that a training-free quality estimator can beat the existing Laplacian-variance baseline on the majority of degradation types implies that the quality screening performed in deployed biometric systems should be upgraded at little cost.

The null finding regarding fairness has perhaps the broadest implications. As face recognition systems are deployed in increasingly high-stakes scenarios, it is critical that they be evaluated for fairness across demographic groups. The finding that an approach beneficial to image quality could have a negative impact on fairness indicates that quality-aware methods must be audited for fairness as part of any evaluation, not in addition to it.

The streaming framework has implications for privacy. Regulations such as the EU AI Act and GDPR—and various national biometric data privacy laws—make deployment of systems that store raw face images legally challenging. The finding that a system using EMA prototype

updates with reservoir sampling achieves accuracy without storing raw face images indicates that accuracy and privacy are not mutually exclusive concepts in this domain.

6.5 Closing Statement

Recognising faces under real-world conditions remains a hard problem. This dissertation has made a small contribution to three aspects of that problem.

Most importantly, the work shows that evaluation matters. The quality estimator is useful but has limitations. The fusion gains are real but have limitations. The streaming framework works but has limitations. The fairness hypothesis was rejected. Each of these qualifications makes clear in what ways and in what contexts each contribution will be useful, and in what contexts it will require further development. A system whose limitations are well understood is less dangerous than one developed without that understanding.

The future work involves validating the methods on real-world multi-sensor data, learning quality heads instead of using hand-crafted quality statistics, extending the streaming method to handle unlabelled updates, and incorporating fairness constraints into the quality estimator itself. This dissertation provides a foundation upon which such further work can build.

Bibliography

- [1] X. Liu et al., “CosFace: Large margin cosine loss for deep face recognition,” in *Proc. CVPR*, 2018, pp. 5265–5274.
- [2] J. Deng et al., “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. CVPR*, 2019, pp. 4690–4699.
- [3] M. Kim et al., “AdaFace: Quality adaptive margin for face recognition,” in *Proc. CVPR*, 2022, pp. 18729–18738.
- [4] G. B. Huang et al., “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Univ. Massachusetts, Tech. Rep. 07-49, 2007.
- [5] B. Maze et al., “IARPA Janus Benchmark-C: Face dataset and protocol,” in *Proc. ICB*, 2018, pp. 158–165.
- [6] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” *Psychol. Learn. Motiv.*, vol. 24, pp. 109–165, 1989.
- [7] J. L. Pech-Pacheco et al., “Diatom autofocusing in brightfield microscopy: A comparative study,” in *Proc. ICPR*, 2000, pp. 314–317.
- [8] J. Hernandez-Ortega et al., “FaceQnet: Quality assessment for face recognition systems,” in *Proc. ICB*, 2019, pp. 1–8.
- [9] P. Terhörst et al., “SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness,” in *Proc. CVPR*, 2020, pp. 5651–5660.
- [10] W. Liu et al., “SphereFace: Deep hypersphere embedding for face recognition,” in *Proc. CVPR*, 2017, pp. 212–220.
- [11] Q. Meng et al., “MagFace: A universal representation for face recognition and quality assessment,” in *Proc. CVPR*, 2021, pp. 14225–14234.
- [12] S. Goswami et al., “RGB-D face recognition with texture and attribute features,” *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1629–1640, 2014.
- [13] A. Mian et al., “An efficient multimodal 2D-3D hybrid approach to automatic face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 1927–1943, 2007.

-
- [14] S. Z. Li et al., “The CASIA NIR-VIS 2.0 face database,” in *Proc. CVPR Workshops*, 2013.
 - [15] R. Gross et al., “Appearance-based face recognition and light-fields,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 4, pp. 449–465, 2004.
 - [16] G. Goswami et al., “RGB-D face recognition using progressively trained convolutional neural networks,” in *Proc. FG*, 2014, pp. 1–6.
 - [17] A. Abaza et al., “Quality based rank-level fusion in multibiometric systems,” in *Proc. BTAS*, 2009, pp. 1–6.
 - [18] A. Abaza and A. Ross, “Quality based fusion for adversarial environments in biometrics,” in *Proc. BTAS*, 2013, pp. 1–8.
 - [19] J. Liu et al., “Multi-modal face anti-spoofing based on central difference networks,” in *Proc. CVPR Workshops*, 2020, pp. 894–895.
 - [20] J. Kirkpatrick et al., “Overcoming catastrophic forgetting in neural networks,” *Proc. Natl. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.
 - [21] A. A. Rusu et al., “Progressive neural networks,” arXiv:1606.04671, 2016.
 - [22] S.-A. Rebuffi et al., “iCaRL: Incremental classifier and representation learning,” in *Proc. CVPR*, 2017, pp. 2001–2010.
 - [23] P. Buzzega et al., “Dark experience for general continual learning,” in *Proc. NeurIPS*, 2020, pp. 15920–15930.
 - [24] J. Snell et al., “Prototypical networks for few-shot learning,” in *Proc. NeurIPS*, 2017, pp. 4077–4087.
 - [25] A. Prabhu et al., “GDumb: A simple approach that questions our progress in continual learning,” in *Proc. ECCV*, 2020, pp. 524–540.
 - [26] J. S. Vitter, “Random sampling with a reservoir,” *ACM Trans. Math. Softw.*, vol. 11, no. 1, pp. 37–57, 1985.
 - [27] C. Spearman, “The proof and measurement of association between two things,” *Am. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904.
 - [28] P. Grother et al., “Face recognition vendor test (FRVT) part 3: Demographic effects,” NIST IR 8280, 2019.
 - [29] M. Wang et al., “Racial faces in the wild: Reducing racial bias by information maximization in face recognition,” in *Proc. ICCV*, 2019, pp. 692–702.
 - [30] Y. Guo et al., “MS-Celeb-1M: A dataset and benchmark for large-scale face recognition,” in *Proc. ECCV*, 2016, pp. 87–102.

- [31] D. Yi et al., “Learning face representation from scratch,” arXiv:1411.7923, 2014.
- [32] K. Kärkkäinen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age,” in *Proc. WACV*, 2021, pp. 1548–1558.