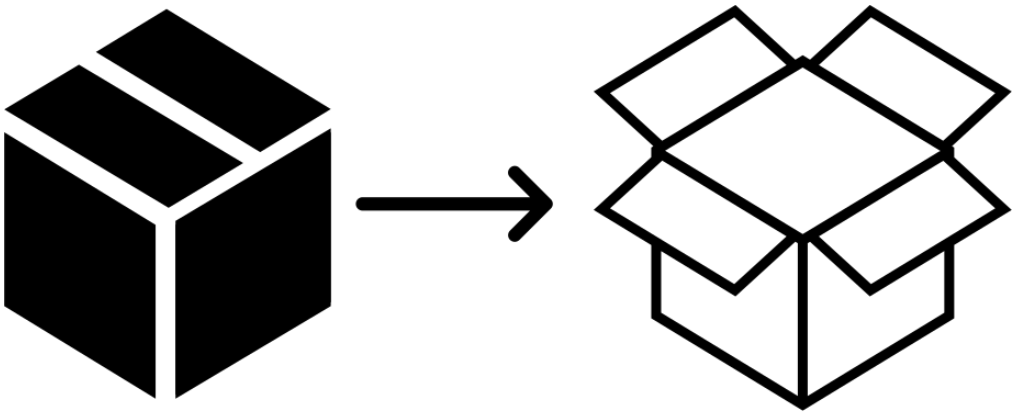
chist-era



XAIface

Measuring and Improving Explainability for AI-based Face Recognition

Progress Report on Implementation of Evaluation Metrics and Protocols (v1)

Deliverable number: D2.2

Version: 1.1

EURECOM Sophia Antipolis · JOANNEUM RESEARCH DIGITAL · universität wien · it · EPFL

Acronym of the project: XAIface
Title of the project: Measuring and Improving Explainability for AI-based Face Recognition.
Grant: CHIST-ERA-19-XAI-011
Web site of the project: https://xaiface.eurecom.fr/

Executive Summary

This deliverable reports the progress of the second task of the second work package in the XAIface project (T2.2). It focuses on the methodology for efficient performance assessment, the design of efficient metrics and protocols, and the overall benchmarking process for face recognition systems.

This document summarizes the activities as follows. A new methodology of performance assessment is firstly proposed. It measures the impact of different types of influencing factors, which is in line with the content in T2.1, on the performance of deep face recognition systems. Then, a benchmarking process is defined to evaluate a face recognition system under two face recognition tasks. Four key elements of a benchmarking process are explained in detail, including two state-of-the-art reference deep face recognition pipelines, multiple face recognition databases, experimental protocols for the two face recognition tasks, and the performance metrics used to report the results. In the end, the benchmarking process is summarized with examples of key instantiations for each dimension.

Table of content

Abbreviations

ACC	Accuracy
AgeDB	Age Database
ArcFace	Additive Angular Margin Loss
AUC	Area Under the Curve
AWGN	Additive White Gaussian Noise
CALFW	Cross-Age Labeled Faces in the Wild
CASIA	Chinese Academy of Sciences' Institute of Automation
CFP-FP	Frontal to Profile Face Verification in the Wild
CMC	Cumulative Match Characteristic
CosFace	Cosine Face
CPLFW	Cross-Pose Labeled Faces in the Wild
DCN	Deformation Convolutional Network
DCNNs	Deep Convolutional Neural Networks
DiveFace	Dataset for Diversity-Aware Face Recognition
DL-Comp	Deep Learning-based Compression
FAR	False Accept Rate
FDDB	Face Detection Data Set and Benchmark
FERET	Face Recognition Technology
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FPR	False Positive rate
FPIR	False Positive Identification Rate
FR	Face Recognition
GB	Gaussian Blur
GN	Gaussian Noise
IJB	IARPA Janus Benchmark
JPEG	Joint Photographic Experts Group
LFW	Labeled Faces in the Wild
LR	Low Resolution
MagFace	Magnitude Face
MTCNN	Multi-task Cascaded Convolutional Networks
Po-Gau-N	Poissonian-Gaussian Noise
ResNet	Residual Neural Network
ROC	Receiver Operating Characteristic
SphereFace	Hypersphere Face
TAR	True Accept Rate
TN	True Negative
TP	True Positive
TPIR	True Positive Identification Rate
TPR	True Positive Rate
YTF	YouTube Faces

1. Introduction

Face recognition has become a prominent biometric technology in our society, widely used in multiple areas, such as access control, video surveillance, automatic annotation etc. With the development of deep convolutional neural networks (DCNNs), some deep learning-based face recognition methods (Chen et al. 2018) (Deng et al. 2019) (Wang et al. 2020) (Sun et al. 2020) (Meng et al. 2021) trained with large-scale face datasets have demonstrated nearly perfect results on popular public face recognition benchmarks. Although the advanced network architecture and discriminative learning approaches successfully boosted the performance, it is critical to fully understand and explain the decisions made by the technology. The objective of the XAIface project is to increase the level of trust of face recognition technology by identifying the influencing factors and better revealing the underlying mechanisms of the current face recognition system.

To accomplish this objective, one of the most important tasks is to identify and to understand the impact and role of different influencing factors in an end-to-end learning-based face recognition system. Deliverable D2.1 has investigated a comprehensive list of possible influencing factors. The goal of this deliverable – D2.2 "Implementation of Evaluation Metrics and Protocols" – is to design suitable metrics and evaluation protocols in order to explicitly and precisely measure the impact of influencing factors in an end-to-end learning-based face recognition system. This report presents the current progress of the design and the implementation of evaluation metrics and protocols.

To systematically measure the impact and the strength of either each influencing factor or the combination of different factors, a new methodology of performance evaluation has been firstly proposed and implemented. Afterward, a comprehensive benchmarking process is developed to give comparative performance assessment for face recognition systems facing different influencing factors.

In this deliverable, the document is structured as follows. Section 1 introduces the general motivation and the objective for the XAIface project and this document. Section 2 presents a novel assessment framework designed for general recognition and detection tasks .The presented framework can quantitatively and comprehensively measure the impact of given influencing factors on the performance of a face recognition system. Then, a benchmarking process is defined to evaluate the impact of given influencing factors in a more rigorous manner. Four key elements of the process will be introduced in the following four sections. Section 3 will introduce the chosen reference face recognition pipeline throughout the XAIface project, i.e., ArcFace (Deng et al. 2019) and MagFace (Meng et al. 2021). Moreover, their advanced performance on multiple public face recognition benchmarks are reported and will serve as a reference baseline for future comparison. Face recognition database is critical for defining the overall face recognition performance. Therefore, Section 4 presents a criteria for dataset selection along with a brief description of the datasets to be used in the project. Section 5 mainly shows the experimental protocol, which often refers to the type of recognition task to be performed with a specific database. The fourth key element is the performance metric, which is described in Section 6. This section summarizes the commonly used performance metrics for face recognition systems and categorizes the metrics

according to the two application scenarios, i.e., verification and identification. Finally, Section 7 summarizes the benchmarking process and the usage of the above four elements.7.

2. Evaluation Framework

Several studies have investigated the vulnerability of CNN-based models to real-world and common image corruptions. (Dodge and Karam 2016) first measured the performance of image classification models on data that suffered from noise, contrast variation and image blur. (Hendrycks and Dietterich 2019) proposed a benchmark to evaluate the robustness of image recognition models towards common corruptions. Extensive work (Michaelis et al. 2019) (Kamann and Rother 2020) has been carried out in object detection and semantic segmentation and applied to safety-critical applications. Current activities in this area mostly study impacts of corruptions during data acquisition and mainly apply to image classification or object detection tasks. Similar analysis in the face recognition community has been reported by (Karahan et al. 2016) (Mehdipour Ghazi and Kemal Ekenel 2016) (Grm et al. 2018), which investigated the robustness of CNN-based FR models. They focused on the impact of face variations caused by standard image processing operation, illumination, occlusion, and misalignment. An assessment framework is proposed to offer a more general solution with a particular focus on face recognition.Additionally, it considers the impact of a wider range of realistic image processing operations in the end-to-end workflow.

In this section, a rigorous and comprehensive assessment framework for face recognition tasks is described, which assesses the impact of various influencing factors. This framework will serve as a broad benchmarking approach for different recognition systems and a wide range of prospective influencing factors. Figure. 2.1 illustrates the architecture of the proposed assessment framework.

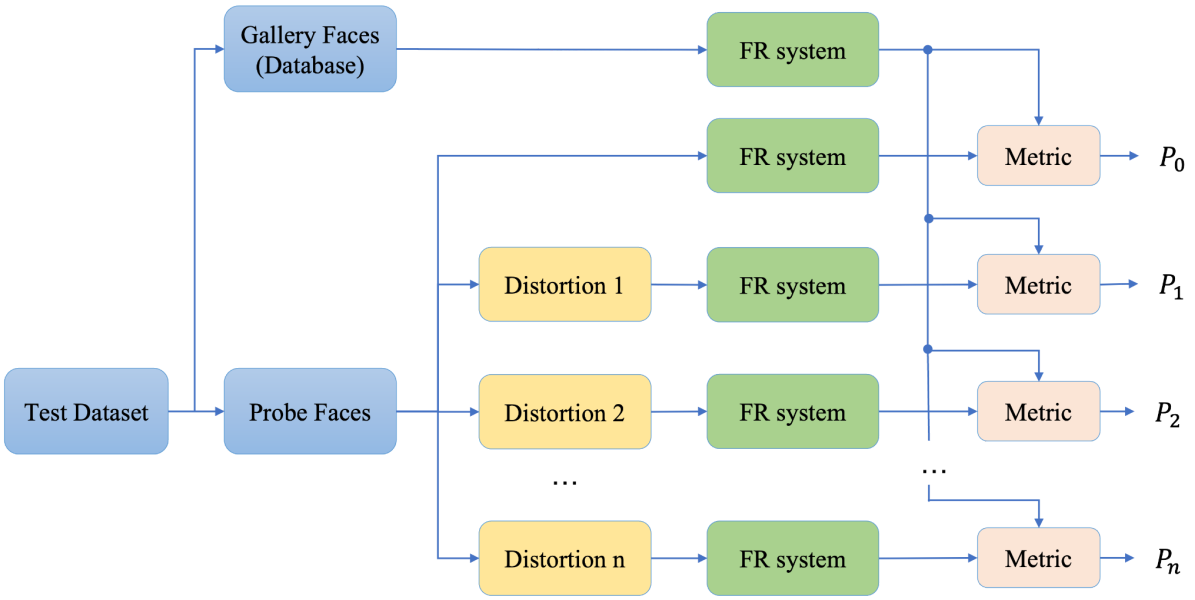


Figure. 2.1: Overview of the assessment framework for face recognition task

In the context of face recognition, the assessment framework performs the following steps to measure the impact of different influencing factors at the same time. The test dataset is firstly splitted to two groups, namely gallery faces and probe faces. Secondly, numerous

potential factors are applied to the input data independently. The face recognition system then extracts the deep features from the gallery faces and the distorted probe faces. The cosine similarity between the features are computed following a certain evaluation protocol. In the end, a task-oriented evaluation metric is adopted to measure the impact of each factor in a quantitative manner.

In line with the work in Deliverable 2.1, the proposed assessment framework now supports a wide range of influencing factors from both extrinsic environment and intrinsic processing operations. Figure. 2.2 depicts the motivation of the framework where a typical data acquisition and transmission pipeline in real world situations is demonstrated. Before being used for face recognition tasks, face image data often suffers from natural distortions during acquisition, such as noise, varying illumination, and low-resolution, followed by a set of pre-/post- processing operations, such as compression, denoising, and resizing. The assessment framework aims at analyzing the possible impact from any of the listed influencing factors as well as the combinations between them.

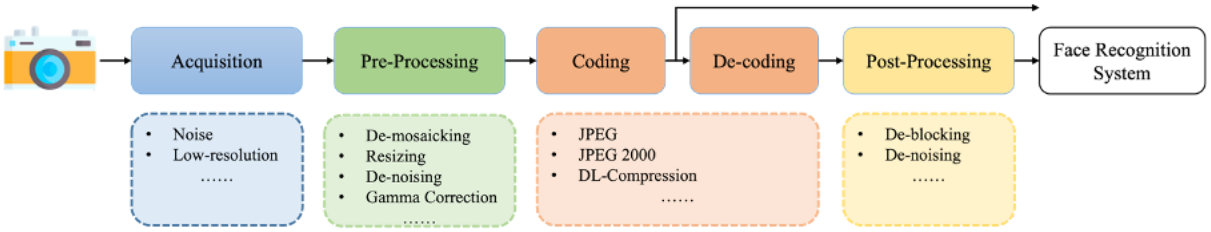


Figure. 2.2: Influencing factors in realistic situations.

The details of all operations used in evaluations are described below with the illustration of a typical example in Figure. 2.3. In general, the framework supports six categories of processing operations or corruption with more than ten minor types. Each type consists of different severity levels.

Compression: Lossy compression refers to the class of data encoding methods that remove unnecessary or less important information and only uses partial data to represent the content. These techniques are used to reduce data size for efficient storage and transmission content and are widely applied to image and video processing. JPEG compression is currently one of the most widely used compression algorithms for images and therefore, it is included in the proposed framework with multiple compression factors. As deep learning-based compression techniques are becoming increasingly popular, the technique developed by (Balle et al. 2018) is also included in this framework.

Smoothing: Image blurring – also known as smoothing – is a widely employed operation to reduce noise which simultaneously results in a reduction of details. Three frequently used filters with various kernel sizes are considered in our framework, including Gaussian, Median, and Average filters.

Noise: The acquisition of images can be easily affected by noise. This framework applies Additive White Gaussian noise (AWGN) with multiple levels of variance. To better reflect the

realistic situations, a synthetic Poissonian-Gaussian noise is also considered, the parameters of which are learned from real-world noisy images.

Enhancement: Image enhancement is generally a very frequently used technique of adjusting images for better display or further image analysis. The contrast and brightness of images are modified by separately applying linear adjustment and Gamma correction.

Resolution: Low-resolution data can significantly reduce the performance of modern deep learning-based detectors (Marciniak et al. 2013) (Li et al. 2012). This is often the case when the face recognition system is employed in an outdoor environment, where captured data could have limited resolution. In this framework, the low-resolution effect is synthesized by downsampling face images with bicubic interpolation

Combinations: It is even more common that the captured face data suffers from multiple distortion and processing operations in a short time. A mixture of two or three operations above is also considered, such as combining JPEG compression and Gaussian noise, making the test data better reflect more complex real-world scenarios.

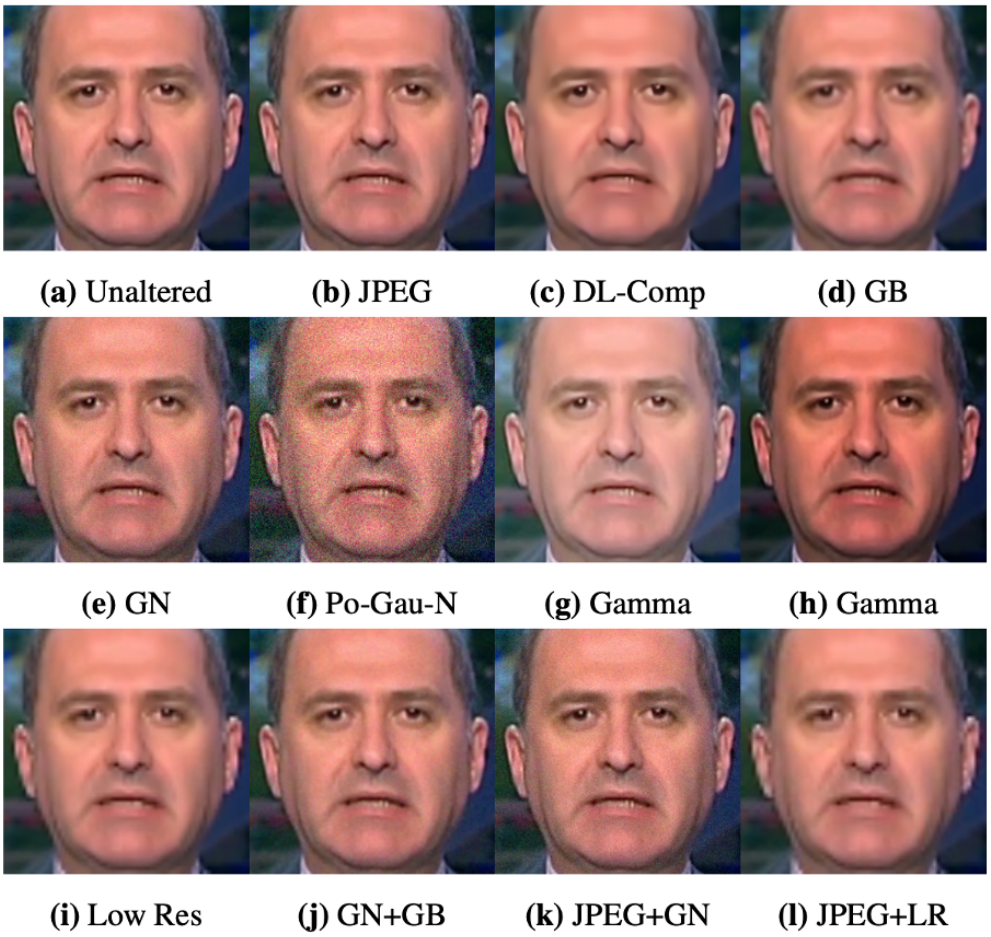


Figure. 2.3: Example of a typical face image picked from a common dataset (Rossler et al. 2019) after applying various distortions and operations. Some notations are explained as follows. DL-Comp: learning-based compression. GB: Gaussian blur. GN: Gaussian noise. Po-Gau-N: Poissonian Gaussian noise. Gamma: Gamma correction. +: mixture.

To use this assessment framework, The face recognition system should be trained on standard face datasets, such as MS1M (Guo et al. 2016) and CASIA-WebFace (Yi et al. 2014), which will be introduced in detail in Section 4. The processing and corruption operations are not applied on training data, but only the probe faces in test data. Furthermore, a number of different parameters for these operations aree adopted to better reflect their impact on face recognition systems.

Current face recognition systems are designed to be as high performing as possible on specific benchmarks. But this often results in sacrificing generalization ability to more realistic situations. The proposed assessment framework is capable of assessing the impact of different influencing factors that often appear in realistic conditions and meanwhile provides valuable insights on designing more robust techniques.

3. Reference Face Recognition Solutions

In general, a face recognition pipeline is built on three main solutions (tools), notably face detection, face alignment, and face recognition (a classifier). The face detection locates the faces in the image or the video frame, and then the face alignment module calibrates the faces and crops them into a predefined size. Finally, the face recognition module extracts discriminative features from the preprocessed faces and performs the recognition task. Face alignment module generally utilizes detected facial landmarks and performs spatial transformation techniques to calibrate faces to a normalized layout, which can be embedded to face detection pipeline. This section will introduce the face recognition solutions adopted by the XAIface project, notably RetinaFace for face detection and ArcFace and MagFace for face recognition.

The overall workflow of the reference face recognition pipeline is illustrated as Figure. 3.1. As depicted in the figure, face images are first fed into RetinaFace face detection solution where it gets bounding boxes and facial landmarks to perform preprocessing, such as cropping and face alignment. The processed face images are then fed into the Arcface face recognition system to further extract deep features for further evaluation. The extracted deep features are used to calculate similarities among the face images according to the adopted evaluation protocols to perform different face recognition tasks.

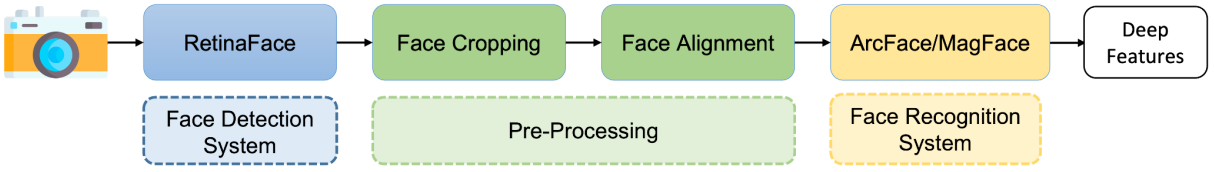


Figure 3.1: Overview of the face recognition pipeline

3.1. Reference Solution 1: RetinaFace for Face Detection

RetinaFace (Deng et al. 2020) is a robust single-stage face detector, which performs pixel-wise face localization on various scales of faces by taking advantage of joint extra-supervised and self-supervised multi-task learning. RetinaFace not only outperforms the state-of-the-art face detection algorithms on the WiderFace (Yang et al. 2016) “Hard” test set, but also proves to help improve the performance of face recognition algorithms. An additional advantage of RetinaFace is that it simultaneously detects faces and five facial landmarks.

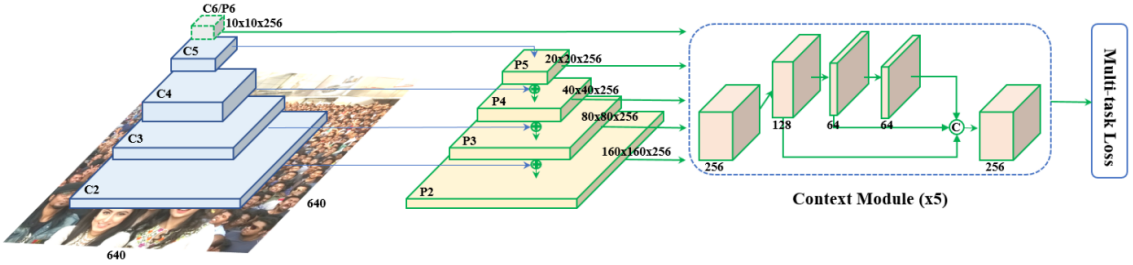


Figure 3.2: Architecture of RetinaFace face detection solution

In detail, RetinaFace adopts a single-shot multi-level face localization approach. The model consists of three main components, including feature pyramid network, context head module, and cascade multi-task loss. The feature pyramid network takes the input image and outputs multiple feature maps at 5 different scales. 4 out of the 5 feature maps are computed from the output of the corresponding pretrained ResNet using top-down and lateral connections, while the last feature map is calculated through a 3x3 convolution, the parameter of which is randomly initialized with the Xavier method (Xavier and Bengio 2010). To strengthen the non-rigid context modeling capacity and increase the receptive field, the deformation convolutional network (DCN) is used to replace all the 3x3 convolution layers in this architecture. Moreover, cascade regression along with multi-task loss are used to improve face localization. The RetinaFace adopts multi-task learning and aims at minimizing the following loss function:

$$L = L_{cls}(p_i + p_i^*) + \lambda_1 p_i^* L_{box}(t_i + t_i^*) + \lambda_2 p_i^* L_{pts}(l_i + l_i^*) + \lambda_3 p_i^* L_{pixel}.$$

The loss function is composed of the following 4 parts:

- **Face classification loss** L_{cls}**:** It is a softmax loss for binary classes to determine if there is a face not not.
- **Face bounding box regression loss** L_{box}**:** t_i and t_i^* represents the coordinates of the predicted bounding box and the ground-truth box associated with the positive anchor. The box regression targets is normalized following the strategy in (Girshick 2015)
- **Facial landmark regression loss** L_{pts}**:** l_i and l_i^* represents the predicted five facial landmarks and ground-truth. Similar to the bounding box, the landmark regression also employs the target normalization.
- **Dense regression loss of** L_{pixel}**:** It compares the pixel-wise difference of the 3D rendered face and the original 2D face

According to the published paper, RetinaFace with ResNet (He et al. 2016) as backbone, achieves a performance of 91.4% average precision on the WiderFace (Hard) test set and was able to run at 13 frames per second images of resolution 640x480. In addition, RetinaFace helps boost the performance of a face recognition system by providing more accurate face bounding boxes and landmarks. After replacing the MTCNN (Kaipeng et al.

2016) face detection algorithm by RetinaFace, the verification accuracy of ArcFace on CFP-FP (Sengupta et al. 2016) dataset has improved from 98.37% to 99.49%.

3.2. Reference Solution 2: ArcFace for Face Recognition

For the classification task of the adopted XAIface face recognition pipeline, we propose to use the ArcFace solution, which is published in Computer Vision and Pattern Recognition Conference (CVPR) (2019), entitled "*ArcFace: Additive Angular Margin Loss for Deep Face Recognition*". This face recognition algorithm uses ResNet as a backbone architecture and proposes a novel Additive Angular Margin Loss to obtain highly discriminative features for face recognition.

The ResNet model is one of the most famous deep neural networks used as backbone in multiple computer vision tasks. The fundamental breakthrough of this deep neural network is to simultaneously train extremely deep neural networks and deal with the vanishing gradient issue. The ResNet resolves the vanishing gradient problem using the shortcut connections technique which essentially skips the training of one or more layers, creating residual blocks. These residual blocks design the path for the gradient to follow back to the earlier layers. The ResNet model to be adopted by XAIface in the context of ArcFace may be ResNet50 or ResNet100, depending on the relative importance given to recognition performance and complexity in a context where explainability is the key target.

The ArcFace face recognition pipeline adopts a novel Additive Angular Margin Loss function, which is improved from the classical Softmax loss. Despite that the Softmax loss is one of the most widely used classification loss function, it does not explicitly optimize the feature embedding to enforce higher similarity for intra-class samples and diversity for inter-class samples, which generates a recognition performance gap in large intra-class face appearance variation (e. g. facial expressions, pose variation, age gap) and large-scale test datasets.

The following equation illustrates the Softmax loss function:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}$$

Where:

x_i: The facial feature of the i^{th} sample belonging to the y_i^{th} class

W_j: The j^{th} column of the weight matrix W

b_j: The j^{th} bias term

N: The batch size

n: The class number

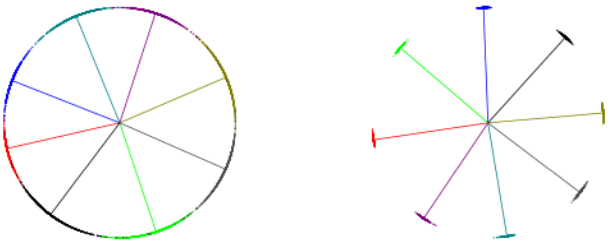
To generate the ArcFace loss from the Softmax loss, the following steps are pursed:

1. Fix the bias b_j to zero.
2. Transform the logit as where θ_j is the angle between the weight W_j and the feature x_i.
3. Normalize the weights and features to make the predictions only depend on the angle between them, i.e fix the individual weight $\left\|W_j\right\| = 1$ by L2 normalization and fix the embedding feature $\left\|x_i\right\|$ by L2 normalization and rescale to s.
4. Add an additive angular margin penalty m between x_i and W_{yi} to enhance the intra-class compactness and inter-class discrepancy.

The following equation presents the ArcFace loss function:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{yi}+m))}}{e^{s(cos(\theta_{yi}+m))}+\sum_{j=1;j\neq yi}^{n} e^{s(cos\theta_j)}}$$

The suggested ArcFace loss outperforms the Softmax loss by re-centering the distributed facial features around each feature center on the hypersphere (see Figure 3.2), and enforcing a more evident gap between classes.



(a) Softmax (b) ArcFace

Figure. 3.2: Example of the softmax and ArcFace loss on 8 identities with 2D features. All the face features are pushed to the arc space with a fixed radius. The geodesic distance between the two classes is more evident after applying the additive angular margin.

The ArcFace loss function has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere. The authors present an extensive experimental evaluation over all recent state-of-the-art face recognition methods and demonstrate excellent results on various face verification and identification benchmarks with multiple evaluation metrics.

According to data from the published paper, the ArcFace face recognition solution reports high verification performance on two most widely used face recognition benchmarks, i.e. Labeled Faces in the Wild (LFW) (Huang et al. 2007) and YouTube Faces (YTF) (Wolf et al. 2011). As shown in the following table, ArcFace trained on MS1MV2 dataset with ResNet100 backbone beats all other methods in the leaderboard.

Table 3.1: Verification performance (%) of different face recognition solutionson LFW and YTF datasets

Solutions	LFW	YTF
DeepID[1]	99.47	93.20
DeepFace[2]	97.35	91.40
VGGFace[3]	98.95	97.30
FaceNet[4]	99.63	95.10
Baidu[5]	99.13	-
Center Loss[6]	99.28	94.90
Range Loss[7]	99.52	93.70
Marginal Loss[8]	99.48	95.98
SphereFace[9]	99.42	95.00
SphereFace+[10]	99.47	-
CosFace[11]	99.73	97.60
MS1MV2, R100, **ArcFace**	**99.83**	**98.02**

The ArcFace solution is also compared with previous state-of-the-art face recognition solutions in terms of TAR (@FAR=1e-4) on IJB-B and IJB-C datasets. The following table shows that ArcFace can obviously boost the performance by at least 5%.

[1] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In NIPS, 2014.
[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In CVPR, 2014
[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In BMVC, 2015.
[4] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, 2015
[5] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv:1506.07310, 2015.
[6] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In ECCV, 2016.
[7] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao. Range loss for deep face recognition with long-tail. In ICCV, 2017.
[8] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In CVPR Workshop, 2017.
[9] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In CVPR, 2017
[10] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song. Learning towards minimum hyperspherical energy. In NIPS, 2018.
[11] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In CVPR, 2018.

Table 3.2: 1:1 verification TAR (@FAR=1e-4) (%) on the IJB-B and IJB-C dataset

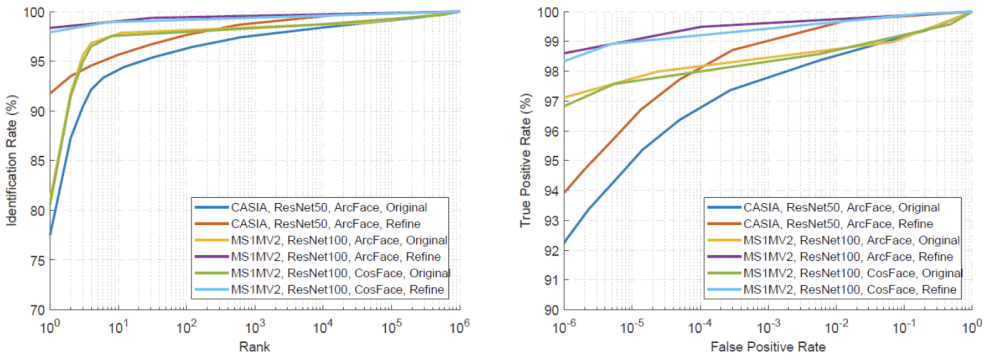
Method	IJB-B	IJB-C
ResNet50[12]	78.4	82.5
SENet50	80.0	84.0
ResNet50+SENet50	80.0	84.1
MN-v[13]	81.8	85.2
MN-vc	83.1	86.2
ResNet50+DCN(Kpts)[14]	85.0	86.7
ResNet50+DCN(Divs)	84.1	88.0
SENet50+DCN(Kpts)	84.6	87.4
SENet50+DCN(Divs)	84.9	88.5
VGG2, R50, **ArcFace**	89.8	92.1
MS1MV2, R100, **ArcFace**	**94.2**	**95.6**

The evaluation of ArcFace is also performed on both verification and identification scenarios using the MegaFace dataset. This dataset includes two protocols for large and small training sets. To perform a fair comparison, the ArcFace is trained on CAISA (resp. MS1MV2) using ResNet50 (resp. ResNet100) under small and large protocols respectively. It is evident from Table 3.3 that Arcface outperforms the other state-of-the-art face recognition solutions for both small and large protocols. In addition, as illustrated in Figure 3.3, ArcFace showed superiority over CosFace and forms an upper envelope of CosFace under both identification and verification scenarios. Moreover, it showed a higher verification and identification accuracy after refining the whole MegaFace dataset from the wrong labels.

Table 3.3: Face identification and verification evaluation of different face recognition solutions on MegaFace Challenge1 using FaceScrub as the probe set. “Id” refers to the rank-1 face identification accuracy with 1M distractors, and “Ver” refers to the face verification false reject rate (TAR) at 10^{-6} false accept rate (FAR). “R” refers to data refinement on both probe set and 1M distractors.

[12] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In FG, 2018
[13] W. Xie and A. Zisserman. Multicolumn networks for face recognition. In BMVC, 2018
[14] W. Xie, S. Li, and A. Zisserman. Comparator networks. In ECCV, 2018.

Methods	Id (%)	Ver (%)
Softmax	54.85	65.92
Contrastive Loss	65.21	78.86
Triplet	64.79	78.32
Center Loss	65.49	80.14
SphereFace	72.73	85.56
CosFace	77.11	89.88
AM-Softmax	72.47	84.44
SphereFace+	73.03	-
CASIA, R50, ArcFace	**77.50**	**92.34**
CASIA, R50, ArcFace, R	**91.75**	**93.69**
FaceNet	70.49	86.47
CosFace	82.72	96.65
MS1MV2, R100, ArcFace	**81.03**	**96.98**
MS1MV2, R100, CosFace	80.56	96.56
MS1MV2, R100, ArcFace, R	**98.35**	**98.48**
MS1MV2, R100, CosFace, R	97.91	97.91



(a) CMC (b) ROC

Figure 3.3: Cumulative match curve (CMC) and receiver operating characteristic curve (ROC) of different models on MegaFace. Results are evaluated on both original and refined MegaFace dataset (Aaron and Ira 2017).

3.3. Reference Solution 3: MagFace for Face Recognition

The second face recognition pipeline adopted by the XAIface project is MagFace (Meng et al. 2021). It is published in Computer Vision and Pattern Recognition Conference (CVPR) (2021), entitled "MagFace: A Universal Representation for Face Recognition and Quality Assessment". Beyond the previous face recognition solutions, MagFace focuses on the problem that the performance of face recognition systems degrades when facing

various-quality face images. To alleviate this issue, Meng *et al* proposed a novel loss function named MagFace, which guides the neural network to learn a more universal feature embedding to adaptively measure the quality of the given face images using the magnitude information. Meanwhile, an adaptive mechanism, which pulls easy samples to class centers while pushing hard samples away is introduced to learn well structured inner-class feature distributions. The MagFace demonstrates a state-of-the-art performance on different face recognition benchmarks, in particular improving the face recognition accuracy in the wild.

Prior work (Deng et al. 2019) often optimizes the model based on a cosine-similarity face recognition loss beyond a fixed margin m, which results in unstable inner-class structure in an unconstrained environment. The natural intuition of MagFace is that high-quality image samples x_i should concentrate in a small region around the cluster w with a high certainty level. By assuming the positive correlation between the image quality and feature magnitude, MagFace additionally proposes a new framework to encode quality factor by optimizing over the magnitude $a_i = ||f_i||$ of each feature vector f_i and meanwhile keeps the cosine-based loss function. Moreover, the magnitude-aware angular margin $m(a_i)$ is proposed and will be penalized during training when magnitude a_i is very large. On the contrary to constraint the freedom of high-quality samples and stably push them to class center, a monotonically decreasing convex function $g(a_i)$ with respect to feature magnitude a_i is designed and works as a regularization.

To sum up, the MagFace extends ArcFace by introducing a magnitude-aware margin and regularizer to enforce higher diversity for inter-class samples and more similarity for intra-class samples. It optimizes the following loss function:

$$L_{Mag} = \frac{1}{N}\sum_{i=1}^{N} L_i \text{ , where}$$

$$L_i = -\log \frac{e^{s(\cos(\theta_{yi}+m(a_i)))}}{e^{s(\cos(\theta_{yi}+m(a_i)))} + \sum_{j=1;j\neq yi}^{n} e^{s(\cos\theta_j)}} + \lambda_g g(a_i)$$

The first term in the loss function is similar to the ArcFace loss, except that the fixed angular margin m is replaced by a magnitude-aware angular margin $m(a_i)$. The hyperparameter λ_g is a trade-off between the classification and regularization.

The MagFace shows an outstanding performance on both easy and difficult face recognition benchmarks.

Table 3.4: Verification accuracy (%) on relatively easy benchmarks

Method	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
Softmax	99.70	98.20	97.72	95.65	92.02
SV-AM-Soft	99.50	95.10	95.68	94.38	89.48

max					
ShpereFace	99.67	96.84	97.05	95.58	91.27
CosFace	99.78	98.26	**98.17**	**96.18**	92.18
ArcFace	99.81	98.40	98.05	95.96	92.72
MagFace	**99.83**	**98.46**	**98.17**	96.15	**92.87**

Table 3.5: Verification accuracy (%) on relatively difficult benchmarks. "*" indicates the results quoted from the original paper.

Method	IJB-B(TAR@FAR)			IJB-C(TAR@FAR)		
	1e-6	1e-5	1e-4	1e-6	1e-5	1e-4
VGGFace2*	-	67.10	80.00	-	74.70	84.00
CenterFace*	-	-	-	-	78.10	85.30
CircleLoss*	-	-	-	-	89.60	93.95
ArcFace*	-	-	94.20	-	-	95.60
Softmax	**46.73**	75.17	90.06	64.07	83.69	92.40
SV-AM-Softmax	29.81	69.25	84.79	63.45	80.30	88.34
SphereFace	39.40	73.58	89.19	68.86	83.33	91.77
CosFace	40.41	89.25	94.01	87.96	92.68	95.56
ArcFace	38.68	88.50	94.09	85.65	92.69	95.74
MagFace	40.91	89.88	94.33	89.26	93.67	95.81
MagFace+	43.32	**90.36**	**94.51**	**90.24**	**94.08**	**95.97**

4. Datasets

In this section we report on the process that led to the selection of the databases to be used in the project. In addition, a brief description of the databases is given below. However, for a more detailed description of the databases, the reader is invited to read document D3.2 “Face image dataset”.

A number of databases were selected according to a list of criteria defined by the consortium (see Table 4.1). The objective of the selected criteria is on the one hand to ensure that the database has the necessary characteristics for the development of the techniques envisaged in XAIface, and on the other hand to ensure the reproducibility of the experiments.

Table 4.1: List of criteria for database selection.

Criterion - type	**Criterion - definition**	**Criterion - values**
Abstract	important features/information about the database	text
Availability	In case the database is no longer available, are there available models trained on that database?	text
Database composition	# classes	#
	# samples per class	#
	Total # of samples	#
	Male / female %	male: #%, female #%
	Ethnicity %	%, not provided
	Age / Age groups %	%, not provided
	# of acquisition sessions	#
	Time span between acquisition sessions	# days
	PIE* variations	yes / no (notes)
	DB size	# GB
	test/train/val split?	yes / no (notes)
Documentation and baseline evaluation	Documentation available and quality assessment	yes (rating 1-5) / no
	Used with ArcFace?	yes / no (notes, ref. to article(s))
	Used with MagFace?	yes / no (notes, ref. to article(s))
Sample features	Data type	e.g. images, videos
	Data format	e.g. TIFF, JPG, PNG, AVI
	Faces are aligned?	yes / no (alignment method)

	Faces are cropped?	yes / no (cropping method)
	Other processing?	yes / no (notes)
	Sample size	#x# pixels
Acquisition	Acquisition sensor	
	Acquisition modality	e.g. visible, thermal, NIR**, etc.
	Multimodal acquisition?	yes / no (list of acquisition modalities and sensors)
	Acquisition conditions	e.g. controlled, uncontrolled
Annotation	Annotation	list of annotated features
	Annotation method	e.g. manual, automatic
GDPR	GDPR compliance	notes
	Collection	yes / no
	Use	yes / no
License	Publicly available	yes / no (notes)
	Commercial usage and changes allowed?	yes / no (notes)
About	Link to DB	url
	Provider	name of provider
	Associated article	reference
Use in XAIface	Used for what task in XAIface?	Task number (notes)

*Pose Illumination Expression, **Near InfraRed

4.1. Short Description of Datasets

4.1.1. AgeDB

The AgeDB (Moschoglou et al. 2017) dataset is used in age-invariant face verification in the wild experiments since it is a manually collected database with a large range of ages for each subject. This property makes AgeDB highly beneficial when training models for age progression experiments. Every image is annotated with identity, age, and gender attributes. AgeDB-30, which is a subset of AgeDB, has been used for validation with MagFace and ArcFace, which are the selected FR pipelines in XAIface.

4.1.2. Labeled Faces in the Wild

Labeled Faces in the Wild (LFW) (Huang et al. 2007) is a public benchmark for face verification. It is a database of face photographs designed for studying the problem of unconstrained face recognition. The dataset contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the portrayed person. The faces were detected by the Viola-Jones face detector (Viola and Jones 2001). LFW is

often selected as a standard reference. However, for some identities only a small number of samples is provided and contains a relatively small proportion of women (according to authors). LFW is used for validation of ArcFace and MagFace.

4.1.3. Cross-Pose LFW

The Cross-Pose LFW (CPLFW) (Zheng and Deng 2018) is an improved version of the LFW face dataset, where more pose variations of the same persons were added while keeping the same identities as in the LFW dataset. The CPLFW dataset is used to achieve face verification. The evaluation of multiple DL face recognition models on CPLFW showed that the accuracy drops by about 15%-20% compared to LFW (see Table 2.2b).
CPLFW is used for validation of ArcFace and MagFace.

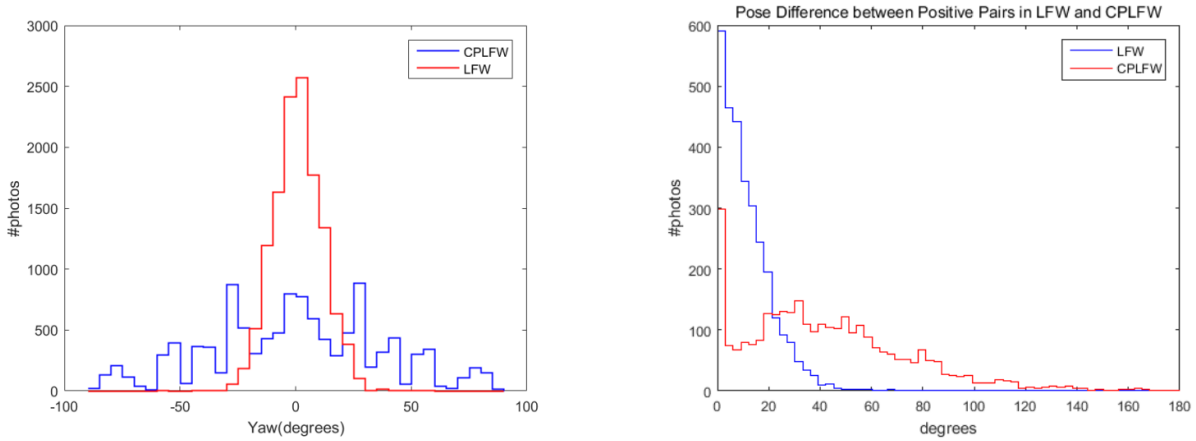


Figure 4.1: Pose variation comparison between LFW and CPLFW.

4.1.4. Cross-Age LFW

The Cross-Age LFW (CALFW) (Zheng et al. 2017) is an improved version of the LFW face dataset, where more face pairs with age gaps were added to add age variation and intra-class variance while keeping the same identities as in the LFW dataset. The CALFW dataset is used to achieve face verification. The evaluation of multiple DL face recognition models on CPLFW showed that the accuracy drops by about 10%-17% compared to LFW (see Table 2.2b). CALFW is used for validation of ArcFace and MagFace.

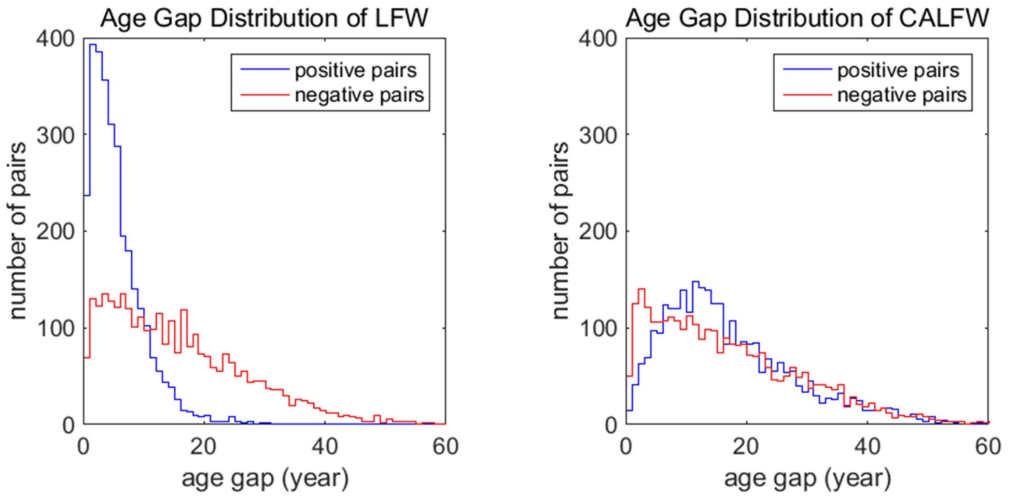


Figure 4.2: Age gap comparison between LFW and CALFW.

Table 4.2: Comparison of verification accuracy (%) on LFW and CPLFW using ArcFace.

Method	**LFW**	**CPLFW**	**CALFW**
ArcFace	99.82%	92.08%	95.87%

4.1.5. DiveFace

DiveFace (Morales et al. 2020) is a dataset designed for bias analysis. It is obtained by extracting balanced sets of face images, according to gender and ethnicity, from the MegaFace database. MegaFace contains images from Flickr. Apparently the MegaFace has been recently decommissioned.

DiveFace contains annotations equally distributed among six classes related to gender and ethnicity (male, female and three ethnic groups). Gender and ethnicity have been annotated following a semi-automatic process. There are 24K identities (4K for class). The average number of images per identity is 5.5 with a minimum number of 3 for a total number of images greater than 150K. Although DiveFace is no longer available, it has been selected as an example of a balanced dataset that might be useful to recreate for experiments in XAIface.

4.1.6. The IARPA Janus Benchmark-C (IJB-C)

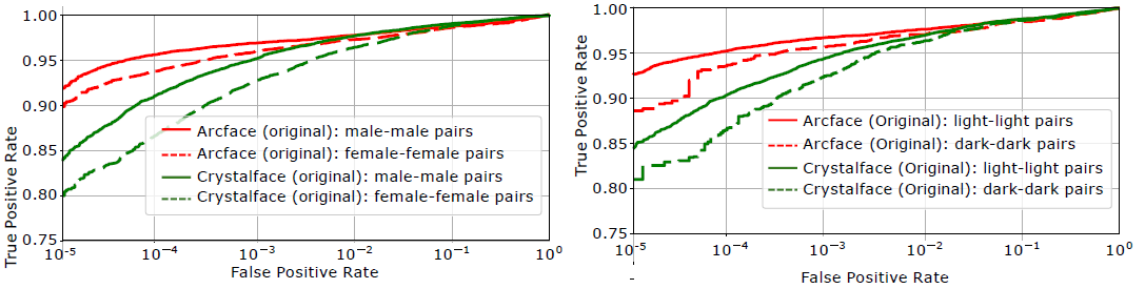
Despite the importance of rigorous testing data for evaluating face recognition algorithms, all major publicly available faces-in-the-wild datasets are constrained by the use of a commodity face detector, which limits, among other conditions, pose, occlusion, expression, and illumination variations. In 2015, the NIST IJB-A dataset, which consists of 500 subjects, was released to mitigate these constraints (Whitelam et al. 2017).

In 2017, IARPA Janus Benchmark-B (NIST IJB-B) database was released, a superset of IJB-A. IJB-B consists of 1,845 subjects with human-labeled ground truth face bounding boxes, eye/nose locations, and covariate metadata such as occlusion, facial hair, and skin

tone for 21,798 still images and 55,026 frames from 7,011 videos. IJB-B was also designed to have a more uniform geographic distribution of subjects across the globe than that of IJB-A.

IJB-C (Maze et al. 2018), released in 2018, adds 1,661 new subjects to the 1,870 subjects released in IJB-B, with increased emphasis on occlusion and diversity of subject occupation and geographic origin with the goal of improving the representation of the global population. Annotations on IJB-C imagery have been expanded to allow for further covariate analysis, including a spatial occlusion grid to standardize the analysis of occlusion. Due to these enhancements, the IJB-C dataset is significantly more challenging than other datasets in the public domain and will advance the state of the art in unconstrained face recognition. IJB-C has been used for evaluation of ArcFace and MagFace (Meng et al. 2021).

Note: It is well known that IJB-C shows gender- and skin tone-wise bias (Dhar et al. 2021).



Anyway, the authors took care to select a large variation of “geographic regions” and did not use “celebrity-only” media. Amazon Mechanical Turk has been used to get good metadata (occlusion, facial hair, gender, capture environment, skin tone, age, and face yaw) so that it should be possible to mitigate and analyze bias-issues.

4.1.7. CASIA-WebFace

The CASIA-WebFace (Yi et al. 2014) is the second largest public dataset available for face verification and recognition problems. This database is used for face verification and face identification tasks and any individual or group is allowed to use this database for educational or non-commercial use free of charge. The face images in the database are crawled from the Internet, more specifically from IMDB by the Institute of Automation, Chinese Academy of Sciences (CASIA). Image collection and identity annotation have been performed following a semi-automatic process. The dataset contains 494,414 face images of 10,575 real identities. This database has been used for training MagFace and the trained model is available online.

4.1.8. MS1MV2 (cleaned version of MS1M, provided by InsightFace)

The MS1MV2 is a refined version of the MS-Celeb-1M (Guo et al. 2016). This large-scale database is used for training face recognition systems and even though the official dataset is

no longer available, trained models are public to all internet users. The original images present in the database were collected from the Internet and the subjects collected were selected according to their popularity on the web. MS1MV2 consists of 5.8M images of 85K different identities. It has also been used to train MagFace and the trained model is publicly available.

4.1.9. FairFace

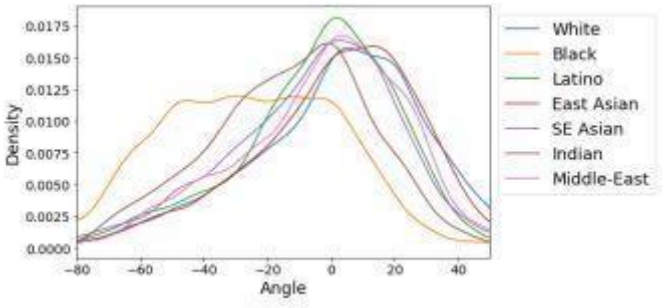


Figure 4.3: Individual Typology Angle (ITA), i.e., skin color, distribution of different races measured in our dataset.

FairFace (Kimmo and Jungseock 2019) is a dataset focused on race balance for bias estimation. In order to mitigate the race bias, the authors emphasize a balanced race composition in the dataset by defining 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino, and ensuring, as shown in Figure 4.3, equal representation. The images were collected from the YFCC-100M Flickr dataset (Thomee et al. 2016) and labeled with race, gender, and age groups thus making possible a bias estimation for all 3 categories. FairFace contains 108,501 images not currently available and just pretrained models on this dataset are still publicly available.

4.1.10. FairFaceRec

The FairFaceRec dataset is a superset of the IJB-C (Maze et al. 2018) dataset created for the ChaLearn challenge. The participants of this challenge were asked to develop fair face verification methods aiming for a reduced bias in terms of gender and skin color. The new superset consists of 13k images from 3k new subjects along with a reannotated version of IJB-C (140k images from 3.5k subjects), totaling ~153k facial images from ~6.1k unique identities. The new database was annotated for gender and skin color as well as for age group, eyeglasses, head pose, image source and face size. Although DiveFace is no longer available, it has been selected as an example of a balanced dataset useful in future experiments in XAIface.

4.1.11. WIDERFace

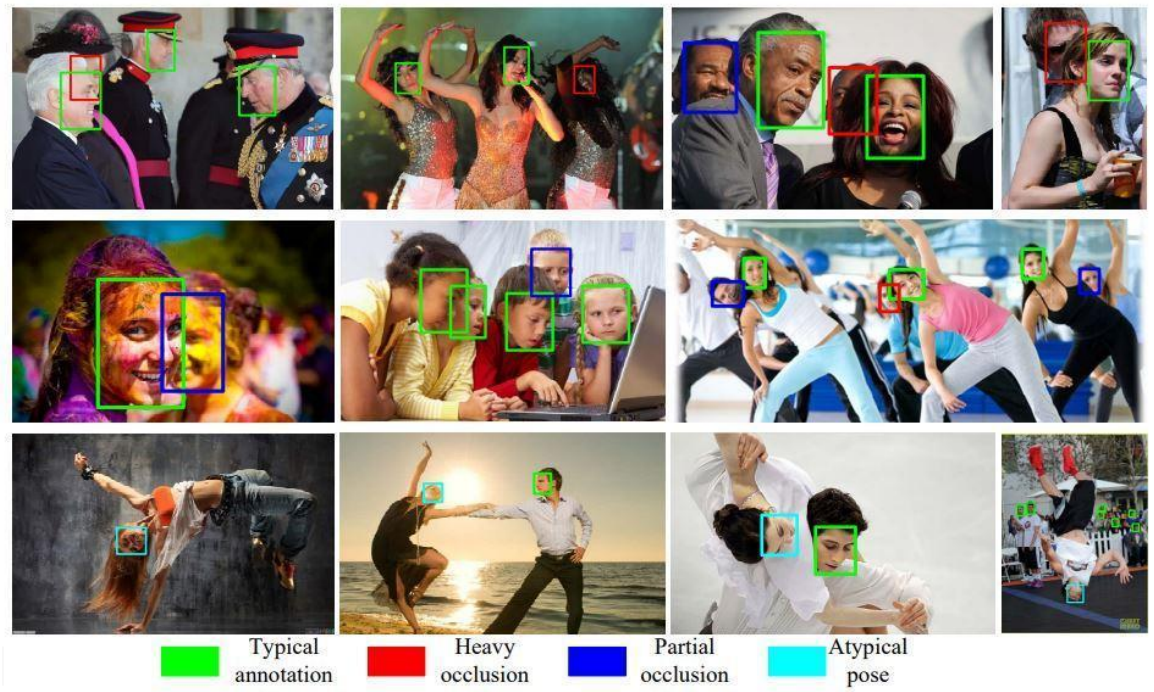


Figure 4.4: Examples of annotation in WIDER FACE dataset (Best view in color).

WIDERFace (Yang et al. 2016) is a database designed for face detection purposes. It contains rich annotations, including occlusions, poses, event categories, and face bounding boxes and it is composed of 32,203 images, labeling 393,703 faces with a high degree of variability in scale, pose and occlusion. The authors suggest a dataset division into training (40%), validation (10%) and testing (50%) sets. Although the database is not publicly available anymore, XAIface members have had access to it. Furthermore, several pre-trained models can be found on the Internet.

4.1.12. VIP_attribute_extended (extended by EURECOM)

The VIP_attribute is a dataset composed of facial images, annotated for gender, body height, weight and BMI which has been used to prove that facial images contain discriminatory information pertaining to those traits. The database is publicly available under request to the authors and consists of mainly frontal face images of celebrities (mainly actors, singers and athletes) collected from the web. It contains one image of each of the 1026 subjects enrolled in it, specifically 513 female and 513 male celebrities. The VIP_attribute_extended is an extension of the annotation of the VIP_attribute database performed by EURECOM. The original database was extended by adding for every subject annotations of their hairstyle, presence and type of facial hair and presence of glasses thus making possible further studies of those categories (Dantcheva et al. 2018).

5. Protocols

Protocols describe detailed reference implementations of the face-recognition methods themselves, the datasets used, and performance measures applied to our face-recognition pipelines in the project used, which ensures a project-wide and comparable consistency of evaluation results obtained. It is critical to define and agree on an almost limited set of reference protocols, as there exist a vast amount of evaluation protocols in literature and it is simply not possible to evaluate against all of them to show at least comparable or improved face-recognition performance when introducing novel developed explainability techniques within the project.

Reference protocols and benchmark definitions for evaluation are usually heavily dependent on related datasets and often published together in so-called “challenges”. This is especially important because there is a clear tendency in increasing database-scale observed during the past decades and a clearly defined selection of (randomized) subsets used e.g. for training and testing is needed to (comparable) make the evaluation process manageable.

The first step in XAIface’s face-recognition reference pipelines is the face-detection step robustly extracting potential face locations in an arbitrary image or video frame. Several face-detection datasets and related testing protocols have been proposed in the literature e.g. Face Detection Data Set and Benchmark (FDDB) (Jain and Learned-Miller 2010) or Annotated Faces in the Wild (AFW) (Zhu and Ramanan 2012). Nevertheless, due to the progress in the face detection research community, it has been necessary to increase the difficulty in pose, scale, facial expression, occlusion, and background clutter thus leading to more complex databases and protocols such as WIDER FACE (Yang et al. 2016). Its evaluation protocol follows an external/internal scenario where the face detector is trained either on any external data or on a provided training/validation partition. The test data partition is always separated. The detection metrics follow the definition of the bounding box evaluation metrics defined in the PASCAL VOC dataset (Everingham et al. 2010) which is basically the ratio of overlap ratios (intersection over union) exceeding 50% area coverage.

Regarding the training protocol, face recognition protocols can be mainly grouped according to subject-dependant and subject-independent protocols (Wang and Deng 2021). Subject-dependent protocols (e.g. used in the early FERET (Phillips et al. 1998) or even challenge 2 of MS-Celeb-1M (Guo et al. 2016)) predefine all testing identities in the training set and thus the recognition process is reduced to a simple classification problem. This is much easier to handle as subject-independent strategies, where the testing identities are different from the training data and thus the recognition model has to generalize the representation for all unknown faces - exhibiting heavy intra-subject variations. Anyway, subject-independent protocols are of higher practical relevance and thus most major face-recognition benchmarks such as LFW (Huang et al. 2007), IJB-x (Maze et al. 2018) or Megaface (Aaron and Ira 2017) follow this paradigm and are thus taken into consideration for our project.

Regarding the evaluation protocol, practical usage is of main interest for the face recognition testing procedure. One of the most important recognition protocols with high practical

relevance to access and control systems is face-verification. **Face-verification** is a 1:1 similarity checking protocol directly comparing two face-images or features either if they are from the same or different persons (= typical **access** scenario). Typical measures used for comparison are the receiver operating characteristics (ROC) or simplified, single number measures such as “mean accuracy” or the true acceptance rate (TAR) at a certain working point with a defined false accept rate (FAR).

Face-identification is a more difficult 1:N (one to many) evaluation protocol relevant to e.g. forensic or in-video searches which can take place in “**closed-set**” or “**open-set”** scenarios. The “**closed set**” scenario compares only faces inside a certain database (all person-identities are “known”), while the **"open-set"** identification scenario includes also “query” instances which are not encoded in the database (identities are ”outside” the database) and their robust detection is an additional challenge for the recognition system.

The most important performance measure for the closed-set face re-identification process is the so-called rank-N metric based on the percentage of correctly returned matches (retrieval rate) within the top N results. Note, that rank-1 metric as e.g. used in IJB-x protocol definition is identical to the exact match returned. For the open-set identification scenarios, the performance measures are more focussed on the applicability of high throughput applications such as video-browsing, and large scale face-search systems. According to Wang et.al. (Wang and Deng 2021) there are currently only a very few challenges to the task of open-set face-recognition such as e.g. IJB-x (Maze et al. 2018). Typical performance measures are false-negative and false-positive identification rates (FNR, FPR) - sometimes diagrammed also in a ROC-manner to allow for proper threshold selection or comparison according to a selected, common working-point.

The face-recognition pipeline(s) selected by XAIface consortium so far consists of a Retina-Face based face-detection and ArcFace / MagFace face-recognition modules. For the face detection protocol, we will follow the WIDER FACE (Yang et al. 2016) protocol using bounding box coverage as performance measure, as the database is one of the latest and most complex ones published in the last years. Unfortunately, the protocol of WIDER FACE does only contain annotations for training and validation. The latter can be used for verification tasks but for the testing protocol we will have to rely on several pre-trained models (publicly available) which can be used as “indirect” ground-truth replacement. Moreover, it will be possible to submit detection results to the organizers of the original WIDER FACE challenge from time to time and get independent results[15].

For the face recognition part of the XAIface pipelines we select two protocols from the five candidates already identified above namely IJB-x and LFW (Huang et al. 2007)-protocol. The main reason for this is that the FERET (Phillips et al. 1998) protocol and database are rather old (approx. 20 years) and thus outdated. Megaface (Aaron and Ira 2017) and MS-Celeb-1M (Guo et al. 2016) should not be used anymore due to the information on database-homepages and several research pages (e.g.

[15] This statement is based on JOANNEUM RESEARCH personal experience from other projects using WIDER FACE as a reference protocol.

https://paperswithcode.com/datasets). We briefly describe those two protocols in the following.

The most recent protocol of the latter two we will focus the **IARPA Janus Benchmark-C face challenge[16] (IJB-x)** - published in 2018 - which is actually one of the most recent and most comprehensive protocols and database definitions based on conducted literature research. Besides standard anchor verification and 1:N identification protocols, it also provides end-to-end evaluations, including detection + recognition modules in still images AND videos. This is also of main interest for the practical usage of such reference pipelines, and hence we will use it as the main evaluation anchor in the XAIface project.

The challenge description[17] defines 8 protocol-tests from which we select the following 2 main anchors namely a.) **Test 1: 1:1 Verification** and b) **Test 4: 1:N Mixed search** for our project..

Other anchors as e.g. **Test 11: Wild Probe Mixed** for end-to-end benchmarking or even **Test 6: Face Detection** (in the case the WIDER FACE anchor selected shows some shortcomings during evaluation) might be selected in later stages of the project and described in later versions of this document. Moreover, the IJB-C protocols provide two disjoint “galleries” used to support 1:N open-set identification scenarios.

Regarding the performance measures, we will use receiver operating characteristics ROC for the verification scenario and average Cumulative Match Characteristic (CMC) - a ROC like graphical visualization of retrieval rate (%) plotted against logarithmic rank-n for validation of XAIface-reference pipeline implementations and measuring the influence of developed explainability modules. In addition we will also use the Detection Error Trade-of (DET) curve, a similar graphical visualization like ROC plotting false rejection rates (FRR) versus false acceptance rates (FAR) using logarithmic scales.

Please note, that in contrast to traditional verification tasks IJB-C utilizes the concept of subject-specific modeling, in which a single template is generated for a subject based upon the available pieces of media – a paradigm shift from the traditional process of creating a template for every available piece of media (e.g., still images and frames) (Maze et al. 2018). We are confident that this concept of averaging over certain templates will not degrade explainability capabilities and eventually slightly modify the protocols in the case of problems.

The second protocol we selected is **LFW**. Despite being originally released in 2007, the LFW database and protocol for face verification[18] is still very useful due to the fact that a lot of algorithms use it as a reference for face verification (1:1 matching, pair matching). Moreover, up to now a lot of comparable methods are still published and thus it is still a valid benchmark. Currently, 4 different variants of the database exist (original, funneled (ICCV 2007), aligned LFW-a, and "deep funneled" images (NIPS 2012) where the latter two provide

[16] *https://www.nist.gov/programs-projects/face-challenges*
[17] *https://www.nist.gov/document/readmepdf-1*
[18] *http://vis-www.cs.umass.edu/lfw/*

the best results (Learned-Miller et al. 2016). In our project we decided to use the original version.

According to the protocol defined in (Huang et al. 2007) to minimize fitting to test data our reference protocol will use only group 2 of the two views provided for algorithm development (validation) and performance reporting respectively as we will not retrain the pipelines in our project. With respect to the 6 protocol and results paradigms mentioned in the reference above we will use the one with the least restrictions regarding the training as we do not plan to train on LFW database. In particular, we choose protocols and reference results from "unrestricted, labeled outside data"[19] giving the freedom to use additional or even different training data (annotations) and potential labelings. Hence the results reported by the selected protocols will be the "mean classification accuracy μ and standard error" as well as the ROC-curves (TPR/FPR) averaged over 10 folds (provided) of view 2. Please note that

$$\hat{\mu} = \sum_{i=1}^{10} p_i / 10$$

where p_i is the percentage of correct classifications on subset i of view 2. The standard error of the mean is defined by $S_E = \hat{\sigma} / \sqrt{10}$ where σ is the estimate of the standard deviation

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{10} \left(p_i - \hat{\mu}\right)^2 / 9}$$

. For more details we refer to (Huanag et al. 2008).

In the next section we describe the implementation of the protocols together with the desired target values. All the reference pipelines and protocols selected will be stored on a joint GIT repository and each partner of the project should install a local copy and ensure that the code works well and produces the same (desired) baseline results in order to have a fair comparison between partners and against literature or developed explainability approaches.

[19] *http://vis-www.cs.umass.edu/lfw/results.html#UnrestrictedLb*

6. Performance Metrics and Human Assessments

6.1. Objective Metrics

A face recognition system can operate in two modes depending on the application scenario, i.e., verification and identification. Identification mode can be further classified as either closed-set or open-set problem, depending on whether the probe set includes identities that are not in the database. Many objective performance metrics are designed to assess the performance of the learning-based face recognition system in different tasks, that are face verification, closed-set face identification, and open-set face identification.

6.1.1. Metrics for Face Verification

Face verification system computes one-to-one similarity between a captured probe image and the registered image in the database and determines whether they belong to the same identity. The outcome of a face verification system is binary - negative (i.e. no match) or positive (i.e. a match). We define P to be all positively predicted results and N to be all negatively predicted results. If a positive prediction is correct, it becomes a "True Positive" (TP), and otherwise designates it as "False Positive" (FP). Similarly, if a negative prediction is correct, it becomes a "True Negative" (TN), in contrast to a "False Negative" (FN) if it is not correct.

Face verification system is classically assessed by **False Match Rate (FMR)**, **False Non-Match Rate (FNMR)**, **Detection Error Tradeoff (DET) curve**, **mean verification accuracy (ACC)** and **receiver operating characteristic (ROC)**.

FNMR refers to the proportion of genuine attempts that are falsely declared not to match a template of the same object. Given a vector of N genuine scores, u, the false non-match rate is computed as the proportion below some threshold, T:

$$FNMR(T) = 1 - \frac{1}{N}\sum_{i=1}^{N} H(u_i - T),$$

where $H(x)$ is the unit step function, and $H(0)$ taken to be 1.

The FMR is the rate at which a biometric process mismatches biometric signals from two distinct individuals as coming from the same individual. So similarly, given a vector of N genuine scores, v, the false match rate is computed as the proportion above T:

$$FMR(T) = 1 - \frac{1}{N}\sum_{i=1}^{N} H(v_i - T).$$

The DET characteristic represents the tradeoff between the above two errors. It plots false non-match rate (FNMR) (Yaxis) vs. false match rate (FMR) (Xaxis) parametrically on threshold T, and often uses logarithmic scale.

Accuracy is simply defined as the percentage of the correct verification pairs and is based on the binary outcomes with "True Positive" (TP), "False Positive" (FP), "True Negative" (TP) and "False Negative" (FN).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

The ROC curve is created by plotting the true accept rate (TAR) against the false accept rate (FAR) at various threshold settings. Under biometric context, TAR measures the percentage of times that a system correctly verifies a true claim of identity, while FAR is defined as the fraction of impostor comparisons that exceeds the threshold incorrectly.

$$TAR = \frac{TP}{TP+FN}$$

$$FAR = \frac{FP}{FP+TN}$$

In general, the lower the cut-off threshold on a positive class, the more samples will be predicted as positive, i.e. higher true accept rate and false accept rate. There is a trade-off between a high true accept rate (TAR) and low error (FAR). To exactly evaluate the performance of a model based on its ROC curve, **Area Under ROC curve (AUC)** has been proposed. It takes value from 0 to 1, where a value of 0 means perfectly in accurate verification while value of 1 reflects the correct verification performance. In general, AUC=0.5 suggests that the model has no discrimination between two faces.

With the development of learning-based techniques, more accurate face recognition systems and evaluation metrics are required. Nowadays, customers of biometric applications consider more about the true accept rate (TAR) when the false accept rate (FAR) is a very low rate in most security verification scenarios. The new evaluation metric is denoted as **TAR@10^{-x} FAR**, where x is usually an integer ranging from 1 to 6. For instance, IJB dataset (Maze et al. 2018) evaluates TAR@10^{-3} FAR, Megaface (Meng et al. 2021) dataset focuses on TAR@10^{-6} FAR.

6.1.2. Metrics for Closed-set Face Identification

Face identification task determines a probe face image belonging to which registered identity in the gallery set. The probe face is compared with every subject in the gallery set. Thus, the identification task is often referred to as one-to-N face matching.

In the closed-set scenario, the identity of each probe face is already registered in the gallery set. **Rank-N** and **cumulative match characteristic (CMC)** are commonly used metrics in this scenario.

Specifically, the given identification protocol matches a probe face against a gallery set of enrolled face images and returns results in rank-order based on the similarity scores. The

rank-N metric reports the percentage of probe searches that are the true-match and are ranked at top-N of the ranking list. Whilst the CMC curve is created by plotting identification rate against given rank values. The identification rate refers to the fraction of correctly identified probe faces. Therefore, the CMC curve reports the percentage of true matching under a given rank.

6.1.3. Metrics for Open-set Face Identification

The open-set identification task refers to the scenario that the identity of a probe face is not necessarily contained in the gallery set, which is more challenging to a face recognition system. **The true positive identification rate (TPIR)** and **false positive identification rate (FPIR)** are the most used metrics for open-set face identification tasks, more specifically for the following two situations.

Firstly, the identity of the probe face image corresponds to a registered identity in the gallery set. This situation is often called mate searching. A succeed mate searching means TPIR which represents the proportion of successful trials of mate searching. The second situation is called non-mate searching, where the probe face does not correspond to any registered identity in the gallery database. The FPIR is used to measure the proportion of non-mate probes that are wrongly identified as enrolled identity in the gallery set.

6.2. Human Assessment

In facial recognition related Operational Technology, human intervention takes place either in order to validate an adapted model before pushing to production, or on individual recognition result level (Human-Algorithm Teaming). Both interpretations will be discussed briefly in this subsection.

6.2.1. Metrics for Open-set Face Identification

To satisfy quality requirements and standards for responsible AI, models can be evaluated by data science specialists prior to application as part of the deployment workflow. This is generally referred to as Human-in-the-Loop (HITL) Machine Learning, whereas a human actor feedbacks low confidence results in order to improve the model while it is still in training, or manually checks the model for flaws before it goes live in a productive system as part of the business process. These approaches help ensure the quality of the applied model and have been implemented by multiple integrated machine learning software products (Ettun et al. 2020) (Wolfewicz et al. 2022). The process can be applied to both supervised, as well as unsupervised learning.

6.2.2. Human Validation of Face Recognition results

Once a model is deployed and in use in a productive system, processes concerning the manual validation of algorithmic Face Recognition classification results can be put in place.

In this case, a human operator confirms or denies the classification proposed by the algorithm. Such approaches help with issues concerning accountability and acceptance, since a human is still in the loop and verifying the result. It has although been shown, that computers have indeed surpassed humans in precision with regards to binary Face verification tasks (Lu and Tang, 2015). Another important point to consider regarding human validation of individual Face Recognition results is the bias introduced into human decision making when teaming between human actors and algorithms take place (Howard et al. 2022). The decisions of test subjects in this context were equally influenced by labels stating the result was generated by an algorithm as they were influenced by labels stating it was a human's prior decision. The Stanford Institute for Human-Centered Artificial Intelligence recommends rigorous piloted A/B Testing in order to assess the impact of Face Recognition outputs on the decisions of its users to detect and counter biased decisions (Ho et al. 2020).

7. Benchmarking

Benchmarking is defined as the process of assessing/measuring the performance of products, services and systems against those known to be leaders or references in one or more aspects of their operations, under well-defined conditions, typically according to representative application scenarios. Benchmarking provides key insights to help understanding how a product, service or system compares with similar, alternative products, services and systems. As such, benchmarking can help to identify areas and tools for improvements—either incremental (continuous) improvements or dramatic (process re-engineering) improvements.

Naturally, in the context of XAIface, benchmarking refers to face recognition systems and the reference benchmarking systems are the selected XAIface face recognition pipelines, notably based on the ArcFace and MagFace face recognition models. These reference face recognition systems have been selected due to their face recognition performance as well as their adoption and popularity as reference solutions in the face recognition research community.

In the context of XAIface, benchmarking may not only refer to the comparative performance assessment of novel face recognition models but, more importantly, to the comparative performance assessment of novel face recognition systems, eventually extending the ArcFace and MagFace reference systems, also including explainability tools. Explainability is becoming increasingly important for systems that rely on deep learning models, being one of the key objectives of the XAIface project, notably to “create tools that will allow assessment and measurement of performance and explanation of decisions of AI-based FR systems”. Since face recognition explainability may come at some face recognition performance cost, benchmarking is essential to address another key objective of the XAIface project, namely to “optimize the trade-off between interpretability and performance”.

With appropriate benchmarking, the face recognition solutions extended with novel explainability tools, to be developed in XAIface, will be characterized in terms of face recognition performance in relation to largely adopted reference solutions, notably in terms of the explainability versus recognition performance trade-off. This will allow adjusting this trade-off for different application domains, notably depending on how critical is the decisions’ explainability for each of them.

In the context of XAIface, appropriate benchmarking involves defining the precise, complete, meaningful processes and conditions under which different face recognition systems, notably with different explainability capabilities, may be compared in terms of face recognition performance. The XAIface benchmarking processes and conditions involve four key dimensions:

1. **Face recognition system** – The first dimension refers to the face recognition system which performance is being assessed, notably in comparison with the XAIface face recognition reference systems, i.e. ArcFace and MagFace (using ResNet50 or ResNet100), see Section 3. The XAIface reference face recognition systems should

be the first to be assessed to establish the reference performance to which the other, e.g. explainability-extended systems, will be compared. As for the XAIface reference systems, the recognition system under assessment may include several tools beyond the core face recognition model, e.g. ResNet50 or ResNet100. This is already the case with the RetinaFace tool included in the reference systems to perform face detection before the face recognition itself. Ablation experiments may be easily performed by defining subsets of the complete systems for evaluation purposes, thus excluding specific tools, to assess their specific impact on the final performance.

2. **Face datasets** – The second dimension refers to the selected face datasets adopted for model training, validation and testing since they are critical for the definition of the overall face recognition performance, see Section 4. Performing benchmarking using different face datasets for different face recognition systems would introduce another variable in the performance comparison process beyond the face recognition system itself, thus making it more difficult to derive solid conclusions.

3. **Experimental protocols** – The experimental protocol dimension refers to the type of recognition task whose performance is being measured using the selected face datasets. The most common face recognition tasks are the so-called verification and identification, which lead to the definition of verification and identification experimental protocols. For example, the ISO/IEC 2382-37:2017 standard defines "Verification as the process of confirming a biometric claim through biometric comparison" and "Identification as the process of searching against a biometric enrolment database to find and return the biometric reference identifier(s) attributable to a single individual". These protocols have to be precisely defined for solid benchmarking, see Section 5, since it is common to see in the literature protocols with the same name (thus ideally addressing the same basic recognition task) but corresponding to slightly different approaches, e.g. on the way they process the probe and gallery/dataset faces, thus leading to values for the performance metrics being reported that cannot be fairly compared.

4. **Performance metrics** – Finally, the last benchmarking dimension defines the precise performance metrics to be considered for each experimental protocol, e.g. verification and identification, see Section 6.

Table 7.1 illustrates the benchmarking dimensions and provides examples of key instantiations for each dimension. For the first dimension, i.e. the face recognition system, the examples include the XAIface reference face recognition systems, ArcFace and MagFace, using the ResNet50 or ResNet100 models, and preceded by RetinaFace for face detection; moreover these reference systems may integrate one of more explainability tools (e.g. tools A and B) and, finally, other face recognition systems (e.g. system X) may be also assessed with or without explainability tools. For the second dimension, i.e. the face datasets, it is important to distinguish the training, validation and testing datasets; naturally, for directly comparable performance results, it is critical that the same training, validation and testing datasets are used. For the third dimension, i.e. experimental protocols, several protocols may be defined although the verification and identification protocols are clearly the most commonly used. Finally, for the fourth dimension, an appropriate set of performance

metrics has to be selected, notably considering the selected experimental protocol, e.g. verification or identification.

Table 7.1: Benchmarking dimensions with example instantiations.

Face Recognition System	Face Datasets (training and testing)	Experimental Protocols	Performance Metrics

• RetinaFace + ArcFace (ResNet 50)
• RetinaFace + ArcFace (ResNet 100)
• RetinaFace + MagFace (ResNet 50)
• RetinaFace + MagFace (ResNet 100)
• …
• RetinaFace + ArcFace (ResNet 100) with explainability tool A
• RetinaFace + ArcFace (ResNet 100) with explainability tools A and B
• …
• Face recognition system X with explainability tool A
• Face recognition system X with explainability tool B
• …
	• Training
	  ○ MS1MV2
	  ○ LFW
	  ○ IJB-B
	  ○ IJB-C
	  ○ …
	• Validation
	  ○ …
	• Testing
	  ○ MS1MV2
	  ○ LFW
	  ○ IJBB
	  ○ IJB-C
	  ○ …
		• Verification
		  ○ …
		• Identification
		  ○ Closed-set
		  ○ Open-set
		  ○ …
		• …
			• False Match Rate (FMR)
			• False Non-Match Rate (FNMR)
			• Failure-To-Enrol rate (FTE)
			• Failure-To-Acquire (FTA)
			• Detection Error Trade-off Curve (DET)
			• TAR (@FAR)
			• Verification accuracy (%)
			• Receiver Operating Characteristics (ROC)
			• Rank-1 identification
			• Rank-N identification
			• False Negative Identification Rate (FNIR)
			• True Positive Identification rate (TPIR)
			• Cumulative Match Characteristic (CMC)
			• …

In practice, performing the benchmarking of a specific face recognition system implies selecting a path along the four dimensions in Table 7.1. For example, different recognition

tasks, e.g. verification versus identification, imply selecting benchmarking paths which are different at least in the last two dimensions.

The comparison of performance results obtained for equivalent benchmarking paths with the exception of the system in the first dimension is fair and reasonable and thus allows obtaining solid conclusions regarding the direct comparison of the involved systems. Very often one of the systems under comparison, which is taken as anchor is one of the XAIface reference pipelines.

In summary, appropriate benchmarking is critical for the efficient development of novel face recognition tools and systems, notably explainability tools as those to be developed in the XAIface project, notably to identify which tools bring effective performance benefits.

8. Conclusions

While face recognition is a long-standing technology and has been widely used in various applications, the explainability of such technology has received relatively less attention and achieved little progress. It is a critical issue to fully understand and explain the decisions made by face recognition systems, which is still an open challenge for even the current state-of-the-art face recognition systems, in particular for the deep learning-based technologies. One of the key objectives of this project is to identify the influencing factors, measure and explain their impacts such that one can better understand the underlying mechanisms in a black-box face recognition system and increase the degree of trust.

To help identify the significant influencing factors and measure their impact, a rigorous and systematic evaluation approach is required. In this context, this deliverable contributes to the project by providing a new methodology for performance assessment, by summarizing a comprehensive benchmarking workflow and by providing thorough literature investigation for each key element of the benchmarking process.

In Section 2, a new performance evaluation methodology is firstly proposed, which serves as an assessment framework in XAIface context and quantitatively analyzes the impact of each influencing factor to a face recognition system. In this framework, a wide range of influencing factors from both extrinsic environments and intrinsic data processing operations are considered. Afterward, a rigorous benchmarking process is illustrated, which offers an impartial and comparative performance assessment for different influencing factors. The four indispensable elements, that are reference face recognition pipelines, face datasets, evaluation protocols, and performance metrics, are described respectively. More specifically, Section 3 gives detailed description on two state-of-the-art face recognition solutions, notably ArcFace and MagFace. A large number of facial datasets for both training and testing purposes are included in Section 4. Section 5 introduces the evaluation protocol used for both verification and identification tasks. The performance metrics are summarized in Section 6, which helps in quantifying the impact of each influencing factor. Finally, the four key elements are integrated and the overview of such a benchmarking process is illustrated in Section 7.

To conclude, a rigorous evaluation framework and benchmarking approach is critical to achieve the objective of the XAIface project, particularly the goal of identifying and quantifying the impact of potential influencing factors to a face recognition system. The proposed assessment framework and the benchmarking process in this deliverable have a close collaboration with D2.1, where a comprehensive list of parameters are investigated. Additionally, it provides an evaluation system and an overall guideline to precisely identify the influencing factors and quantify their impacts.

References

Aaron, Nech, and Kemelmacher-Shlizerman Ira. “Level Playing Field For Million Scale Face Recognition, CVPR.” 2017.

Balle, Johannes, et al. “Variational image compression with a scale hyperprior.” *arXiv preprint arXiv:1802.01436*, p. 2018.

Chen, S., et al. “Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices.” *in Chinese Conference on Biometric Recognition*, Spring, 2018, pp. 428-438.

Dantcheva, A., et al. “Show me your face and I will tell you your height, weight and body mass index.” *24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3555-3560.

Deng, J., et al. “Arcface: Additive angular margin loss for deep face recognition.” *in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690-4699.

Deng, J., et al. “Retinaface: Single-shot multi-level face localisation in the wild.” *InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203-5212.

Dhar, P., et al. “PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recognition.” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Dodge, Samuel F., and Lina Karam. “Understanding how image quality affects deep neural networks.” *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1-6.

Ettun, Yochay. “Machine Learning Pipelines with Human Validations | cnvrg.io.” *Cnvrg.io*, https://cnvrg.io/machine-learning-pipelines-human-validations/. 2020

Everingham, M., et al. “The pascal visual object classes (voc) challenge.” *International journal of computer vision*, 2010, pp. 303-338.

Girshick, R. “Fast r-cnn.” *Proceedings of the IEEE international conference on computer vision*, 2015.

Grm, K., et al. “Strengths and weaknesses of deep learning models for face recognition against image degradations.” *Iet Biometrics*, vol. 7, no. 1, 2018, pp. 81-89.

Guo, Y., et al. “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition.” *European conference on computer vision*, 2016.

He, K., et al. “Deep residual learning for image recognition.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Hendrycks, Dan, and Thomas Dietterich. “Benchmarking neural network robustness to common corruptions and perturbations.” *Proceedings of the International Conference on Learning Representations*, 2019.

Ho, Daniel E., et al. “Evaluating facial recognition technology: a protocol for performance assessment in new domains.” *Denv. L. Rev*, 2020.

Howard, John., et al. “Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making.” *Plos one*, 2022.

Huanag, G., et al. “Labeled Faces in the Wild: A Database Forstudying Face Recognition in Unconstrained Environments.” *In Workshop on Faces in’Real-Life’Images: Detection, Alignment, and Recognition*, 2008.

Huang, G.B., et al. “Labeled faces in the wild: A database for studying face recognition in unconstrained environments.” *Technical report*, 2007.

Jain, V., and E. Learned-Miller. “Fddb: A benchmark for face detection in unconstrained settings.” *UMass Amherst technical report*, vol. 2, no. 6, 2010.

Kaipeng, Z., et al. “Joint face detection and alignment using multitask cascaded convolutional networks.” *IEEE Signal Processing Letters*, vol. 23, no. 10, August 2016, pp. 1499-1503.

Kamann, Christoph, and Carsten Rother. “Benchmarking the robustness of semantic segmentation models.” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020.

Karahan, S., et al. “How image degradations affect deep cnn-based face recognition?” *in 2016 international conference of the biometrics special interest group (BIOSIG)*, 2016, pp. 1-5.

Kimmo, K., and J. Jungseock. “Fairface: Face attribute dataset for balanced race, gender, and age.” *arXiv preprint arXiv:1908.04913*, 2019.

Learned-Miller, E., et al. “Labeled faces in the wild: A survey.” *In Advances in face detection and facial image analysis*, 2016, pp. 189-248.

Li, Pei, et al. “On low-resolution face recognition in the wild: Comparisons and new techniques.” *IEEE Transactions on Information Forensics and Security*, vol. 14, 2019, pp. 2000-2012.

Lu, Chaochao, and Xiaoou Tang. “Surpassing human-level face verification performance on LFW with GaussianFace.” *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Marciniak, Tomasz, et al. “Influence of low resolution of images on reliability of face detection and recognition.” *Multimedia Tools and Applications*, vol. 74, 2013.

Maze, B., et al. “IARPA Janus Benchmark - C: Face Dataset and Protocol.” *2018 International Conference on Biometrics (ICB)*, 2018, pp. 158-165.

Mehdipour Ghazi, M., and H. Kemal Ekenel. “A comprehensive analysis of deep learning based representation for face recognition.” *in Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 34-41.

Meng, Q., et al. “Magface: A universal representation for face recognition and quality assessment.” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14225-14234.

Michaelis, Claudio, et al. “Benchmarking robustness in object detection: Autonomous driving when winter is coming.” 2019.

Morales, A., et al. “SensitiveNets: Learning Agnostic Representations with Application to Face Images.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Moschoglou, S., et al. “AgeDB: The First Manually Collected, In-the-Wild Age Database.” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1997-2005.

Phillips, P., et al. “The FERET database and evaluation procedure for face-recognition algorithms.” *Image and vision computing*, 1998, pp. 295-306.

Rossler, Andreas, et al. “Faceforensics++: Learning to detect manipulated facial images.” *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Sengupta, S., et al. “Frontal to profile face verification in the wild.” *2016 IEEE winter conference on applications of computer vision (WACV)*, 2016.

Sun, Y., et al. “Circle loss: A unified perspective of pair similarity optimization.” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398-6407.

Thomee, B., et al. “YFCC100M: The new data in multimedia research.” *Communications of the ACM*, vol. 59, no. 2, 2016, pp. 74-73.

Viola, P., and M. Jones. “Rapid object detection using a boosted cascade of simple features.” *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001.

Wang, M., and W. Deng. “Deep face recognition: A survey.” *Neurocomputing*, vol. 429, 2021, pp. 215-244.

Wang, X., et al. “Mis-classified vector guided softmax loss for face recognition.” *in Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12241-12248.

Whitelam, C., et al. “IARPA Janus Benchmark-B Face Dataset.” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 592-600.

Wolf, Lior, et al. “Face Recognition in Unconstrained Videos with Matched Background Similarity.” *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Wolfewicz. “Human-in-the-Loop in Machine Learning: What is it and How Does it Work?” *Levity.ai*, 22 August 2022, https://levity.ai/blog/human-in-the-loop.

Xavier, Glorot, and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks.” *Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings*, 2010.

Yang, S., et al. “Wider face: A face detection benchmark.” *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

Yi, D., et al. “Learning face representation from scratch.” *arXiv preprint arXiv:1411.7923*, 2014.

Zheng, T., et al. “Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments.” *arXiv preprint arXiv:1708.08197*, 2017.

Zheng, T., and W. Deng. “Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments.” *Beijing University of Posts and Telecommunications, Technical Report 18-01*, February, 2018.

Zhu, X., and D. Ramanan. “Face detection, pose estimation, and landmark localization in the wild.” *In 2012 IEEE conference on computer vision and pattern recognition*, 2012, pp. 2879-2886.