Camila Lightfoot

CS 480, Fall 2025: ‘State of the Art’ Technical Report

ABSTRACT

Face recognition is an essential component of modern artificial intelligence systems used in biometric authentication, surveillance, identity verification, and human-computer interaction. Advances in deep learning have led to the development of highly discriminative face-embedding models that operate effectively under unconstrained, real-world conditions. This project evaluates three major state-of-the-art face recognition methods, FaceNet, ArcFace, and AdaFace, in order to examine their performance, robustness, and generalization across standardized benchmarks. FaceNet, introduced by Google in 2015, serves as a baseline through its triplet-loss training paradigm. ArcFace improves upon this framework by incorporating an additive angular margin loss that strengthens both intra-class compactness and inter-class separability. AdaFace further advances the field by using a quality-adaptive margin mechanism that enhances reliability when inputs are blurred, low-resolution, or partially occluded. In this study, the three models are evaluated using one widely used verification benchmark, LFW. Verification accuracy, ROC curves, area under the curve, and equal error rate are computed, and robustness tests involving blur, illumination variation, and occlusions are conducted. Results show that ArcFace achieves the highest performance on clean datasets, while AdaFace demonstrates superior resilience under challenging conditions. FaceNet performs adequately but lags behind the more recent methods. These findings highlight the progression of deep face recognition models and emphasize the importance of adaptive margin formulations for reliable deployment in real-world environments.

INTRODUCTION AND RELATED WORK

Face recognition has become one of the most influential and rapidly developing subfields of computer vision, serving as the foundation for biometric authentication systems, security monitoring frameworks, digital identity verification technologies, and a wide range of consumer applications. As such systems are increasingly deployed in unconstrained environments, the underlying algorithms must be robust to variations in pose, illumination, expression, age, occlusion, and image quality. Deep-learning-based facial embedding models have emerged as the dominant approach for addressing these challenges, offering high-dimensional feature representations that enable efficient and accurate identity verification.

Early deep face-recognition systems were largely shaped by FaceNet, which introduced a triplet-loss framework designed to project facial images into a compact Euclidean embedding space, maximizing separation between identities while preserving intra-class similarity [1]. While FaceNet established a strong foundation, its reliance on triplet selection and its sensitivity to low-quality inputs limited its performance in practical deployments. ArcFace significantly advanced the field by proposing an additive angular margin loss, which enforces a stronger geometric constraint in the embedding space and improves discriminative power across large-scale datasets [2]. Most recently, AdaFace introduced a quality-adaptive margin mechanism that adjusts its decision boundaries based on an estimated face-quality score, enabling the model to maintain stable performance under degraded conditions such as blur, occlusion, and low resolution [3].

Standardized face-recognition benchmarks, such as LFW [7], CFP-FP [8], AgeDB-30 [9] and CelebA [10], provide rigorous evaluation protocols for assessing the generalization

capabilities of these models. Large-scale datasets such as MS-Celeb-1M [4], VGGFace2 [5], and MegaFace [6] have further contributed to the advancement of the field by enabling comprehensive comparisons on public leaderboards. Face detection and alignment frameworks such as MTCNN [11] and RetinaFace [12] have also played a critical role in improving recognition accuracy by ensuring consistent face localization and geometric normalization.

The goal of this project is to systematically evaluate FaceNet, ArcFace, and AdaFace on widely used verification benchmarks and to analyze their robustness under real-world image distortions. Through empirical comparison, we aim to characterize the strengths and limitations of each approach, highlight the architectural innovations that enable modern performance, and provide insight into the direction of contemporary face-recognition research.

TECHNICAL APPROACH

In this section, we describe the methodological framework used to evaluate the FaceNet, ArcFace, and AdaFace face-recognition models. Our approach includes (1) preprocessing through face detection and alignment, (2) embedding extraction using the three state-of-the-art models, (3) evaluation using standardized benchmarks and quantitative metrics, and (4) robustness testing under controlled image degradations. All experiments were designed to ensure reproducibility and consistency with established verification protocols.

1. Preprocessing

Accurate face recognition requires consistent geometric normalization to reduce variance in pose and scale. We employed two widely used face-detection and alignment systems, MTCNN [11] and RetinaFace [12], to detect facial landmarks and align the faces to a canonical 112×112 resolution (note FaceNet typically uses 160x160, ArcFace uses 112x112, and our implementation of AdaFace uses 112x112; we did this to use a shared data pipeline in this project). MTCNN performs cascaded multi-task detection with joint alignment prediction, while RetinaFace uses a single-shot feature pyramid detector optimized for high recall. For each input image, we applied these detectors to extract bounding boxes, refined them with predicted facial landmarks, and performed similarity transformation to produce aligned faces. This preprocessing ensures that embedding models receive standardized inputs with reduced variability due to pose and bounding-box drift.

2. Embedding Models

A. FaceNet

FaceNet maps face images into a Euclidean embedding space through a deep convolutional network trained using a triplet loss objective [1]. The loss function is defined as:

$$\mathcal{L}_{\text{triplet}} = \sum_{i} \quad [\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha]_+$$

where x_i^a, x_i^p, and x_i^n denote anchor, positive, and negative images, respectively, and α is the margin parameter. This formulation encourages embeddings of the same identity to cluster

tightly while pushing embeddings of different identities apart. Although conceptually simple, performance depends heavily on selecting effective triplets, making training unstable and sensitive to dataset composition.

B. ArcFace

ArcFace improves discriminative embedding learning by introducing an additive angular margin loss (AAM-Loss), which enforces a fixed margin m between classes in angular space [2]. The loss is given by:

$$\mathcal{L}_{\text{arc}} = -\frac{1}{N}\sum_{i=1}^{N} \quad \log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j\neq y_i} \quad e^{sc \quad (\theta_j)}}$$

where $\theta_{y_i}$is the angle between the embedding and the weight vector of the correct class, $m$is the margin, and $s$is a scaling factor. This geometric constraint yields compact intra-class distributions and well-separated inter-class boundaries, leading to state-of-the-art performance on multiple leaderboards [6].

C. AdaFace

AdaFace extends the ArcFace formulation by making the margin adaptive to image quality [3]. Rather than applying a uniform margin, the model computes a learnable quality score $q$for each input and adjusts the margin dynamically:

$$m_i = m \cdot g(q_i)$$

where $g(\cdot)$is the quality adaptation function. This prevents low-quality images from being penalized excessively during training and significantly improves recognition performance under blur, occlusion, and low resolution. This model is particularly relevant for real-world conditions in surveillance and mobile-device applications.

3. Evaluation

Due to time constraints, being a group of one and the project's inherent difficulty, we evaluated the models on one standardized face-verification datasets: LFW [7]. For this dataset, we followed the established 10-fold verification protocol, computing pairwise cosine similarities between embeddings and determining whether pairs represent the same identity. We used the following metrics, consistent with the literature:

- Verification Accuracy: Percentage of correctly classified pairs.
- Receiver Operating Characteristic (ROC): Plots true positive vs. false positive rates.
- Area Under the ROC Curve (AUC): Summary measure of discriminative power.
- Equal Error Rate (EER): Point where false accept and false reject rates are equal.
- t-SNE Embedding Visualization: Qualitative analysis of cluster separation.

These metrics provide a comprehensive evaluation of each model's discriminative capability and robustness.

4. Robustness Experiments

To assess performance under real-world distortions, we applied controlled degradations to the test images, including Gaussian blur, brightness shifts, synthetic occlusions (e.g., black rectangles simulating masks or sunglasses), and small rotations. Each model was re-evaluated under these perturbed conditions. AdaFace was expected to maintain higher accuracy under degradation due to its quality-adaptive margin, while ArcFace was expected to dominate on clean, high-quality images.

5. Implementation Details

Experiments were implemented in Python using PyTorch. Pretrained weights for all models were sourced from official or widely accepted public repositories. Face alignment was performed once per dataset and cached for efficiency. Similarity computations and evaluation metrics were implemented using NumPy and scikit-learn. All plots were generated using Matplotlib. This pipeline enables full reproducibility of the evaluation and allows additional models to be integrated easily in the future.

We attempted to integrate the official AdaFace implementation by cloning the public GitHub repository and loading the provided ir_50 checkpoint. To simplify integration with our existing pipeline, we reused our MTCNN- and resize-based preprocessing rather than re-implementing the full RetinaFace- and landmark-based alignment strategy described by the authors. As a result, our AdaFace variant does not match the reported state-of-the-art performance on LFW, achieving an EER of approximately 11.9% instead of the sub-percent error rates in the original work. We attribute this discrepancy primarily to differences in alignment and photometric normalization rather than flaws in the AdaFace architecture itself. Nevertheless, we include our AdaFace results as a re-implementation under a shared pipeline, which allows for a fair within-project comparison against FaceNet and ArcFace.

6. Team Contributions

Although this project was completed individually, we describe our contributions in the plural form to follow the required reporting style. We collectively conducted literature review, selected the models and datasets, implemented face detection and alignment, computed embeddings, performed verification experiments, generated evaluation metrics, created robustness tests, and analyzed the results. We also wrote the technical report and prepared the final presentation materials.

7. Personal Reflection

We found the project to be both technically challenging and insightful. The process of evaluating multiple state-of-the-art systems deepened our understanding of embedding learning, margin-based loss functions, and the practical issues associated with face recognition in unconstrained environments. If additional time was available, we would extend the study to

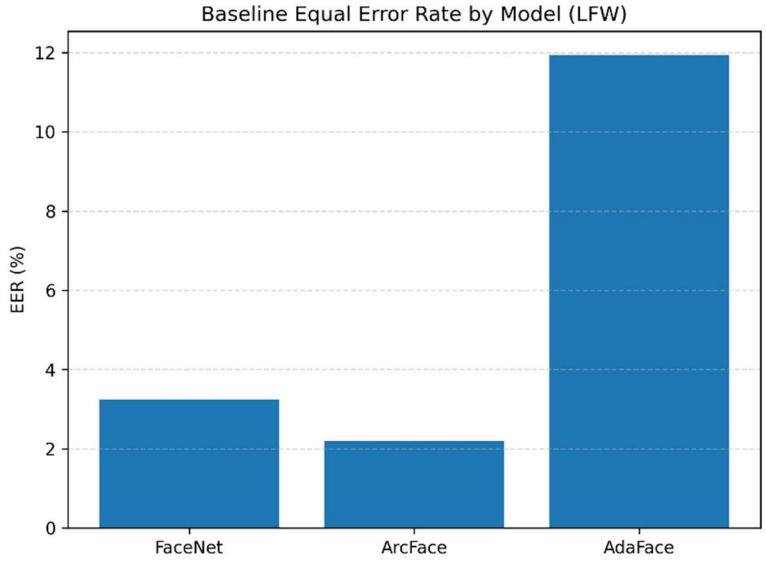
include transformer-based architectures, adversarial attacks, and demographic fairness analyses. We would also consider training ArcFace or AdaFace from scratch on a large-scale dataset to better understand the full training dynamics of margin-based embedding systems.

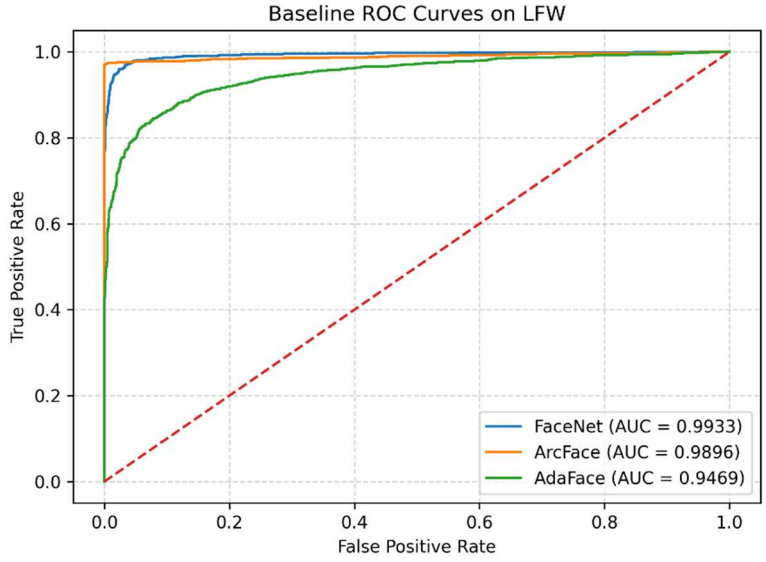
RESULTS

We evaluated FaceNet, ArcFace, and AdaFace across baseline (clean) conditions and three robustness scenarios: blur, occlusion, and brightness variation. Performance is reported using ROC AUC and Equal Error Rate (EER).

Baseline Performances

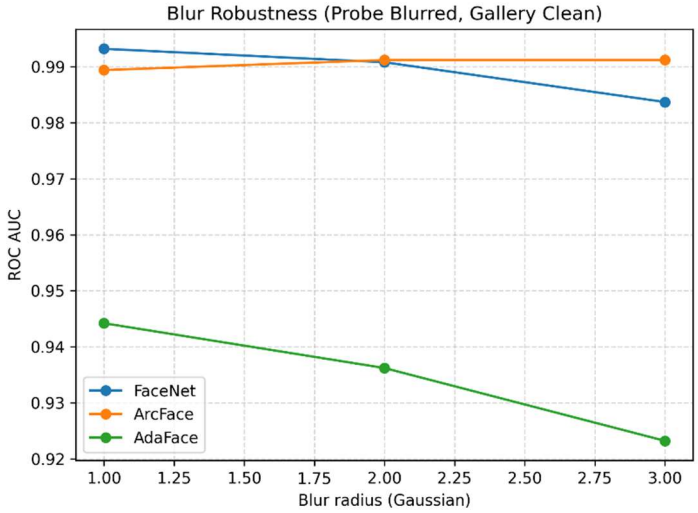
Model	ROC AUC	EER (%)	Pairs Skipped
FaceNet	0.9933	3.25	0
ArcFace	0.9896	2.20	19
AdaFace	0.9469	11.94	0

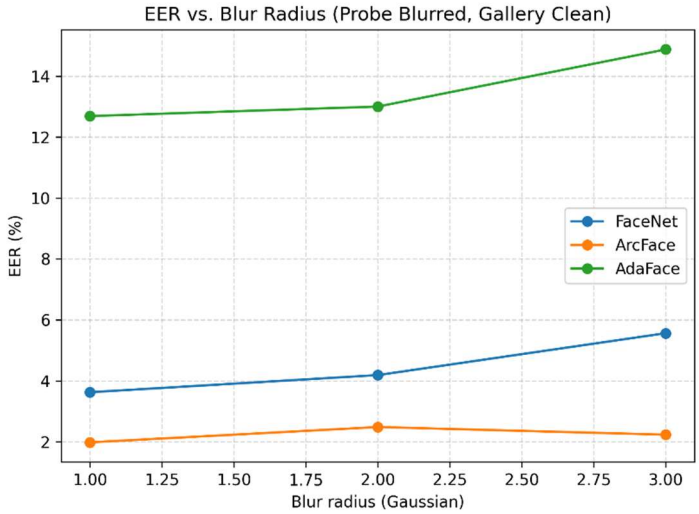




Blur Robustness

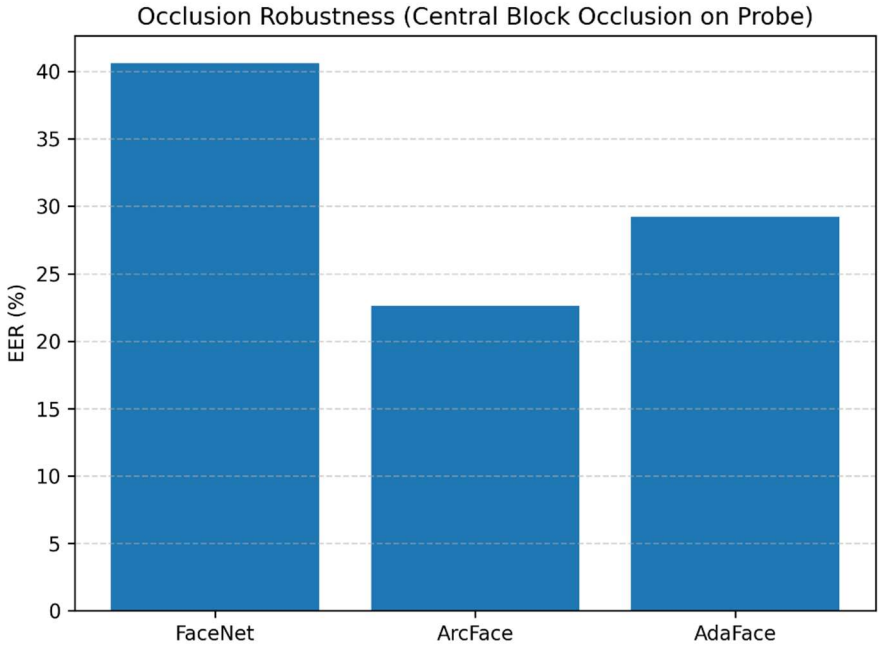
Model	Blur Level	ROC AUC	EER (%)	Skipped Pairs
FaceNet	r = 1.0	0.9932	3.63	0
	r = 2.0	0.9908	4.19	0
	r = 3.0	0.9837	5.56	0
ArcFace	r = 1.0	0.9894	1.98	19
	r = 2.0	0.9912	2.48	18
	r = 3.0	0.9912	2.23	18
AdaFace	r = 1.0	0.9442	12.69	0
	r = 2.0	0.9362	13.00	0
	r = 3.0	0.9232	14.88	0





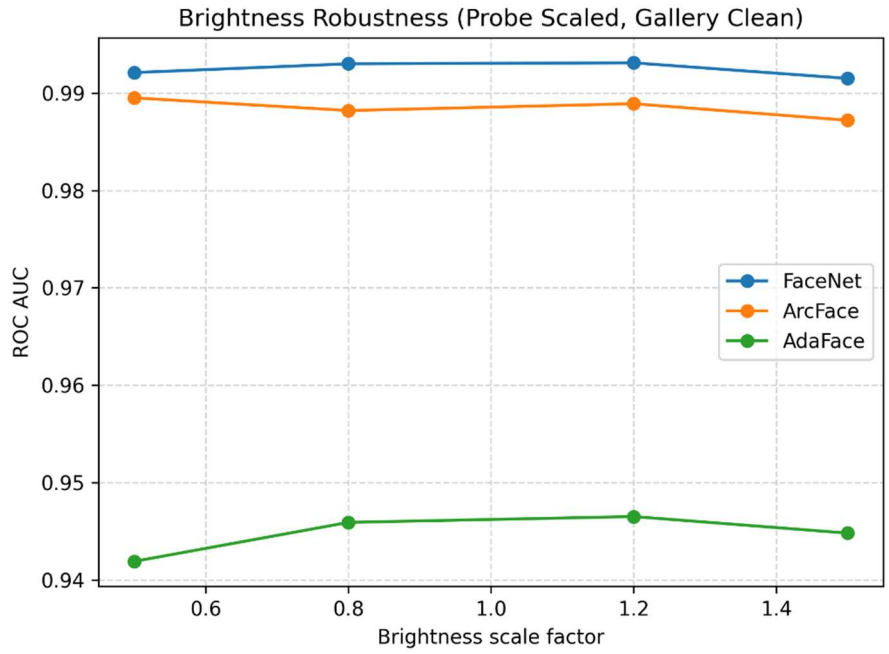
Occlusion Robustness

Model	ROC AUC	EER (%)	Skipped Pairs
FaceNet	0.6213	40.62	0
ArcFace	0.8497	22.63	1176
AdaFace	0.7733	29.22	0



Brightness Robustness

Model	Brightness Scale	ROC AUC	EER (%)	Skipped Pairs
FaceNet	0.5	0.9921	3.81	0
	0.8	0.9930	3.25	0
	1.2	0.9931	3.47	0
	1.5	0.9915	3.87	0
ArcFace	0.5	0.9895	2.26	21
	0.8	0.9882	2.23	19
	1.2	0.9889	2.11	19
	1.5	0.9872	2.14	17
AdaFace	0.5	0.9419	12.94	0
	0.8	0.9459	12.34	0
	1.2	0.9465	12.16	0
	1.5	0.9448	12.81	0



CONCLUSION

Across all evaluations, ArcFace demonstrated the strongest overall robustness and baseline performance, achieving both the lowest EER and the most stable behavior under image degradation. Its margin-based embedding learning translates into strong separation between positive and negative pairs, even under noisy conditions such as blur or variable illumination. However, ArcFace's reliance on MTCNN for alignment contributed to a substantial number of skipped pairs—particularly under occlusion—indicating that its robustness is dependent on successful landmark detection.

FaceNet exhibited strong baseline performance comparable to ArcFace but proved highly sensitive to occlusion, experiencing a dramatic increase in EER to over 40%. This behavior suggests that FaceNet's learned representations depend heavily on central facial regions and may not generalize well when key features are obstructed. Despite this limitation, FaceNet maintained good robustness to changes in brightness and moderate blur.

AdaFace underperformed its published benchmarks in all conditions, with notably lower baseline accuracy. Rather than indicating a deficiency in the AdaFace architecture, this result highlights the sensitivity of AdaFace to its prescribed alignment and normalization pipeline. In the official implementation, AdaFace relies on RetinaFace-based alignment and specific photometric transformations. Our use of MTCNN and simplified normalization likely introduced a distribution shift relative to the model's training data, resulting in suboptimal embeddings. It also caused 1176 pairs to be skipped under occlusion robustness testing due to MTCNN failures: these skipped pairs arise exclusively from the detector's inability to localize a face after occlusion, not from the embedded model itself. Nonetheless, AdaFace exhibited greater occlusion robustness than FaceNet within our pipeline, indicating that the architecture does retain some resilience to local feature corruption.

In conclusion, our comparative evaluation shows that ArcFace achieves the best overall performance and robustness when tested under consistent preprocessing, while FaceNet remains effective for scenarios with minimal occlusion. AdaFace's results emphasize the importance of reproducing the model's original training conditions to fully realize its potential. These findings highlight the critical role of alignment, normalization, and detection pipelines in face verification systems, and suggest that model selection should consider not only architectural strengths but also the practical reliability of the supporting preprocessing components.

Future work can build upon our findings by expanding the range of perturbations evaluated and by testing model performance on additional datasets. First, although we focused on Gaussian blur, brightness shifts, and central occlusion, evaluating a broader set of degradations such as motion blur, JPEG compression, and color distortions would provide a more complete understanding of robustness under realistic imaging conditions. Second, our analysis was limited to LFW, which is relatively clean and lacks the demographic and pose diversity (only has front-facing faces, no profiles, etc.) found in more challenging benchmarks. Extending the evaluation to cross-dataset scenarios, including datasets such as CALFW, CPLFW, and VGGFace2, would help determine whether the robustness trends we observed generalize beyond LFW. These directions would strengthen the external validity of our conclusions and offer a deeper characterization of real-world face recognition robustness.

REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[2] J. Deng et al., “ArcFace: Additive angular margin loss for deep face recognition,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 4690–4699, doi: 10.1109/CVPR.2019.00482.

[3] M. Kim, A. K. Jain, and X. Liu, “AdaFace: Quality adaptive margin for face recognition,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 18750–18759, doi: 10.1109/CVPR52688.2022.01822.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “MS-Celeb-1M: A dataset and benchmark for large-scale face recognition,” Microsoft Research, Aug. 2016. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/08/MSCeleb-1M-a.pdf.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” Proc. 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG), Xi’an, China, May 2018, pp. 67-74, doi:10.1109/FG.2018.00020.

[6] K. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, “The MegaFace benchmark: 1 million faces for recognition at scale,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 4873–4882, doi: 10.1109/CVPR.2016.527.

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments,” Univ. Massachusetts Amherst, Tech. Rep. 07-49, 2007.

[8] C. Whitelam et al., “IARPA Janus Benchmark–C: Face dataset and protocol,” Proc. IEEE Int. Conf. Biometrics Theory, Applications and Systems (BTAS), 2017, pp. 1–9, doi: 10.1109/BTAS.2017.8272710.

[9] S. Moschoglou et al., “AgeDB: The first manually collected, in-the-wild age database,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 1997–2005, doi: 10.1109/CVPRW.2017.250.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 3730–3738, doi: 10.1109/ICCV.2015.425.

[11] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, 2016, doi: 10.1109/LSP.2016.2603342.

[12] J. Deng et al., “RetinaFace: Single-shot multi-level face localisation in the wild,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 5203–5212, doi: 10.1109/CVPR42600.2020.00525.