International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 310-323
https://doi.org/10.70917/ijcisim-2026-2889
Article

# A Deep Learning Framework for Facial Attribute Recognition:

# Implementation and Performance Analysis Using Augmented Data

**Shefali Aggarwal[1], Karthik Kovuri[2]**

[1]Research Scholar, Department of Computer Applications, RIMT University, Mandi Gobindgarh, Punjab, India. Email: shefaliaggarwal187@gmail.com
[2]Professor and Dean, Academic Affairs, RIMT University, Mandi Gobindgarh, Punjab, India. Email: dean.academics@rimt.ac.in

**Abstract**:Facial Attribute Recognition (FAR) applications, powered by deep learning models, reach impressive benchmark results and target semantic characterisation of faces in images. The facial features semantic characterisation helps identify features such as gender and age, as well as other attributes (e.g. cosmetics, glasses, facial hair). Distinct from face identification, FAR applications intend to pose distinct characterizing questions of face features. Unfortunately, FAR deep learning models cannot be effectively deployed and evaluated in real-world settings due to representativeness issues in the training and testing data. In this paper, we build a deep learning-based FAR framework which attempts to mitigate issues presented in current models. Within the framework, we use a multi-task ResNet-based deep learning model, trained on the CelebA data set, which includes augmentation with simulated occlusion and illumination variations, and combined class-weighted loss functions to address imbalances in class attributes. A robust evaluation paradigm, with per-attribute metrics, cross-dataset testing on LFWA+ and challenging subset evaluations, and fairness and demographic analyses are also incorporated. Results indicate a substantial improvement in robustness to occlusion and low-illumination conditions. Per-attribute metrics also revealed performance issues that were not seen using overall accuracy. LFWA+ benchmark data confirmed the generalisation potential of training under data augmentation. Fairness analyses also suggested improvements in demographic representation. The paper presents a practical FAR model and an evaluation methodology aimed at narrowing the performance gap evident in benchmark tests and the practical and scalable use of FAR in real world conditions.

**Keywords:** facial attribute recognition, deep learning, data augmentation, occlusion robustness, illumination variation, multi-task learning, demographic fairness, CelebA, LFWA+

## 1. Introduction

FAR systems provide flexibility and the ability to function in open systems like surveillance networks and driver behavior monitoring systems, among others. This is because they do not need a reference database for enrolled identities and can work in absence of it. (Mahendran et al., 2017).

Deep learning and CNN-based systems began dominating FAR undertaking following the large-scale CelebA benchmark created in 2015, which provided 202,599 face images with 40 binary attributes. Multi-task learning techniques, attention mechanisms, and transformer models, as well as hybrid models, have contributed to accuracy improvement on this benchmark over the last decade. (Krizhevsky et al., 2017). There is a gap and it is widening between benchmark performance and the reliability of FAR systems in practice and general use.

There are four reasons for this gap. First, most benchmark datasets are demographically non-inclusive. Most CelebA datasets, for example, are unbalanced with almost exclusively White, Western, young-to-middle-aged

celebrities. FAR models trained on these types of datasets do poorly on underrepresented demographic groups (availability gap). Controlled datasets provide no way to measure the instances when an individual's face is occluded, such as when they are wearing a mask, sunglasses, or a hat. (Mery, 2022). Third, large positive class imbalances for an attribute cause training to be dominated by a large number of attributes in CelebA. This causes models to perform well on most datasets and attributes, but poorly on those with a very low number of examples. Fourth, the industry's leading practice of providing a single test dataset for a model, measuring its performance in terms of overall accuracy, does not provide any of the previously mentioned gaps. (Rudd et al., 2016).

The framework integrates: (1) a ResNet-based multi-task architecture, which is trained on CelebA and LFWA+; (2) a targeted augmentation pipeline, which applies controlled occlusion and illumination variation, builds difficult evaluation subsets, and introduces increasingly strenuous occlusion variants during evaluation; (3) class-weighted binary cross-entropy loss, along with balanced validation, to address the multi-attribute imbalance; and (4) a thorough evaluation protocol that indicates overall accuracy, cross-dataset generalisation, difficult-subset robustness, and per-demographic-group fairness, as well as precision, recall, and F1 per attribute.

The following are the contributions for this research paper: (i) Provide compelling evidence for the systematic improvement of robustness on the occlusion and illumination control challenge with augmentation training, (ii) present the case for improvement of accuracy on rare attributes by reporting the F1 score for each attribute, (iii) present evidence for augmentation and improvement of general recognition in lower frames with an additional dataset (LFWA+), and (iv) present evidence for the analysis of fairness of the remaining differences with respect to performance of each demographic and the distribution of the dataset.

In the following sections of the paper, see Section 2 for an overview of related literature, Section 3 for the description of the proposed frameworks, Section 4 for the results of the experiments, and Section 5 for the consideration of the findings, limitations of the study, and the goals for future research.

## 2. RELATED WORK

### *2.1 Deep Learning Architectures for FAR*

The study of Facial Attribute Recognition (FAR) has gone through five architectural generations. Recognition methods underpinned by classic feature engineering, such as LBP, HOG, and Gabor wavelets, offered hand-crafted representations of facial texture and shape. While promising, these methods were often ineffective to capture the full attribute diversity and demonstrated limitations due to various contextual constraints. The field of FAR shifted towards the use of deep learning with the advent of AlexNet (Wang & Deng, 2021) and the development of CelebA (W. Liu et al., 2017) as the dominant datasets to establish benchmarks for multi-attribute learning.

Methods that employ CNNs present learned representations and part-based architectures, such as PANDA (N. Zhang et al., 2014), which process separate facial regions and thus improve localisation of facial attributes. However, a problem with global average pooling in standard CNN architectures is that it assigns the same level of importance to all spatial regions which may not be occluded. This fundamental flaw of such architectures is discussed in literature (Buolamwini & Gebru, n.d.). It is possible to analyse attributes with a shared backbone; however, it is necessary to use separate classification heads for each attribute. Hierarchical Multi-Task Learning (Hand & Chellappa, 2017) combined with dynamic task weighting (Cao et al., 2018)have improved the accuracy of classification on a per attribute level, however, they have also led to negative transfer and task-interference obstacles.

Because of attention mechanisms, CNN's lack of spacial sensitivity was improved; this was achieved through the use of spatial weight maps which increase the importance of classification of some regions above others (Zhu et al., 2017). From this, the Attribute-specific attention modules (Chen et al., 2021), and Visibility-Aware Attention for Occlusion (Sethi et al., 2019) were developed; however, the quality of the method was determined to be dependent on the variety of the training data for CNNs. With the introduction of ViT (Dosovitskiy et al., 2020) as well as TransFA (D. Liu et al., 2022) for Facial Attribute Recognition (FAR), the use of Transformers for facial classification became popular; however,

they are expensive from a computational perspective and require significant amounts of data to pre-train. Hybrid CNN-Transformer architectures such as CvT and Swin Transformer (Z. Liu et al., 2021) are located in the intersection of accuracy and efficiency in the context of Transformers; however, the failures of each of the components, i.e. Transformers and CNNs, are combined in them. None of these changes to architectures have systematically tackled the issues of robustness to occlusion, inability to be invariant to changes in.

A related line of work pertinent to FAR focuses on face identification, which has developed powerful backbone representations using metric learning objectives. DeepFace (Taigman et al., 2014) demonstrated that deep networks could achieve human-level performance for face verification on the LFW benchmark. FaceNet (Schroff et al., 2015) presented triplet loss, which enforced compactness in embedding spaces for same identity pairs and separation of different identity pairs, achieving 99.63% on LFW. Developing from this, ArcFace (Deng et al., 2019) produced embedding spaces that contained rich identity-related information, including age, gender, and ethnicity, through an additive angular margin loss. The relevance to FAR is clear, as the training of large-scale deep neural networks on face identification datasets (e.g., VGGFace2, MS-Celeb-1M) substantially outperforms FAR results on models that are trained using ImageNet. This is likely due to the deep networks on face identification datasets learning features that are structurally similar to features that are attribute-related. However, training on face identification datasets also biases FAR models by having a narrow range of demographics. The proposed framework uses ImageNet as a pre-training dataset to gain an advantage from visual domain feature learning while reducing the demographic biases of face identification datasets.

### *2.2 Datasets and Their Limitations*

The CelebA Dataset (2015) is the most popular benchmark in facial attribute recognition (FAR). However, it has many shortcomings, including: demographic imbalances with Western, young, and White celebrity images; binary gender labels; and subjective labels with varying degrees of distribution. LFWA+ uses the same 40 labels with the LFW dataset, which sources images from news articles, and offers more variety with respect to imaging conditions. MAAD-Face extends this to 3.3M images and 47 attributes and uses semi-automated annotation (Terhorst et al., 2021) However, CelebA remains the primary benchmark. UTKFace and FairFace focus on demographic inclusivity (Z. Zhang et al., 2020), but have limited vocabularies for attributes. Importantly, no benchmark FAR dataset includes controlled annotation for pose and illumination alongside other variables, meaning there is no way to evaluate how the attributes of these FAR datasets are robust by constructing evaluation subsets.

### *2.3 Robustness and Fairness in FAR*

(Buolamwini & Gebru, n.d.) claim that in the case of commercial facial analysis systems, the best- and least-represented groups have an up to 30% error rate disparity. Subsequently, (Patel & Kisku, 2024) claim this disparity was reduced, but still present, with KL-divergence loss and dual attention framework specific to each attribute. (Jaiswal et al., 2025) claim that the disparity was present across models regardless of the loss functions used, suggesting that it stems more from the datasets rather than the models. Finally, (Mery, 2022) claim low attribute occlusion (lower face and masks) results in lower accuracy.

The absence of a defined protocol in assessing robustness or fairness is apparent from published works in the domain. Most reported works only mention the single-benchmark overall accuracy (Rudd et al., 2016).

The impact of performance gaps moves beyond the metrics being evaluated and presents dire consequences once products are deployed. In the field of security and surveillance, performance gaps in different demographic groups will create uneven rates of false-positive and false-negative results across different population groups. This is very concerning for access control and law enforcement (Basheer, 2025). For healthcare, false accept rate (FAR) systems used for pain assessment and those for screening tools for the brain and spine must have equitable performance for all patient populations, including elderly

populations and non-Western populations who are underrepresented in CelebA (Ben Aoun, 2024). For driver monitoring, performance gaps under low-light and occluded conditions, that mimic the experience of driving at night while wearing sunglasses, completely negates the purpose of the system in the first place (Herath et al., 2025). Systematic underperformance of demographic groups in educational proctoring also introduces both ethical and legal issues (Coghlan et al., 2021). The evaluation framework in this paper was designed to capture issues arising in deployment that standard benchmark evaluation cannot. It is based on the structured multi-dimensional design described in section 3.4.

## 3. PROPOSED FRAMEWORK

### *3.1 Architecture*

The proposed framework relies on a ResNet-50 architecture that has been pre-trained on ImageNet and has been fine-tuned to handle multi-attribute FAR. The backbone of the model extracts a unified feature representation of each submitted face image. The final feature layer is accompanied by forty binary classification heads, each corresponding to a specific attribute from CelebA. Due to the nature of the task, each head has been assigned a fully connected layer and a sigmoid activation function, thus allowing each attribute to be predicted independently. This design of the framework will allow each head to specialize independently, while the shared features of lower levels will be leveraged across all tasks.

Choosing ResNet-50 illustrates a compromise between representation and required computational power. ResNet-101 and other architectures like hybrid CNN-transformer frameworks marginally improve accuracy on standard benchmarks, but they come with higher inference costs. As for the multi-task sigmoid, I chose it over the softmax-based multi-label output. This is because the former allows the independent calibration of predictions, which is valuable when considering the class frequencies, as they can be quite unbalanced.

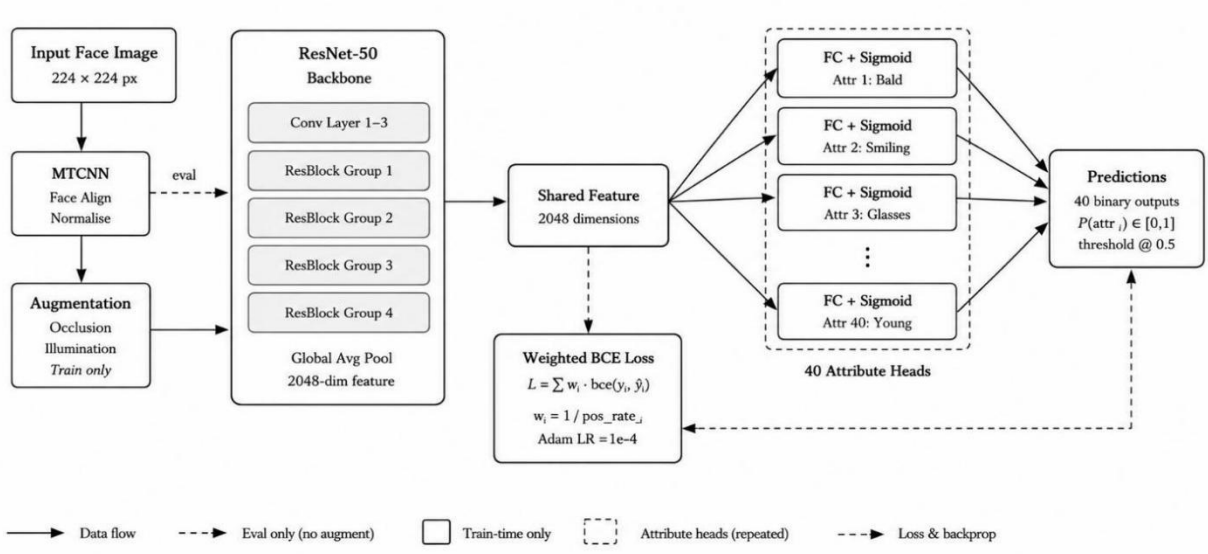


*Figure 1: Proposed multi-task FAR framework. A ResNet-50 backbone extracts a shared 2048-dimensional feature representation. Forty attribute-specific sigmoid heads produce independent binary predictions. Class-weighted BCE loss drives training; augmentations are applied only during training.*

### *3.2 Data Preparation and Augmentation Pipeline*

Two functions are implemented in the augmentation pipeline: adding robustness to the model by training it under suboptimal conditions, and building controlled evaluation subsets to measure systematic robustness. All

augmentations are made randomly during the training phase. For evaluation subsets, augmentations are made in a predetermined manner.

For robustness to occlusions, the pipeline implements three types of augmentations, where the three augmentations are randomly applied. These three types of augmentations are: (1) Randomly erasing the mouth, jaw, eyes, and foreheads to simulate hair, hats, and facial occlusion, (2) using rectangular occlusion patches that simulate masks on the nose and mouth, and (3) using occlusions to the eyes that simulate glasses. In the evaluation subsets, three types of severity of occlusions are defined as "mild" (15-25%), "moderate" (30-45%), and "severe" (50-60%).

To achieve robustness to illumination, the pipeline implements (i) brightness as low as 30% or as high as 180% (ii) gamma correction (gamma as low as 0.4 and as high as 2.5), (iii) adjustment of contrast, and (iv) shadows that are simulated in a linear gradient, and are made at varying angles. For the illumination evaluation subsets, the conditions of low light (brightness of 0.3-0.5), overexposure (brightness factor of 1.5-1.8), and containing shadows. are applied to the test subsets.

The following augmentations have been applied to every training image without differentiating the types of challenges: random horizontal flip, a random crop with 10% padding, color jitter, and blur. Augmentation is preceded by a Face MTCNN pipeline. All parameters of augmentations can be found in Table A.

| Augmentation Type | Target Region | Parameter Range | Training Probability |
|---|---|---|---|
| Rectangular erasure (occlusion) | Mouth/jaw, eye, forehead | 15%–60% coverage | p = 0.4 per region independently |
| Synthetic mask overlay | Nose and mouth | Fixed anatomical region | p = 0.3 |
| Synthetic glasses overlay | Eye region | Fixed anatomical region | p = 0.25 |
| Brightness scaling | Global | Factor ∈ [0.3, 1.8] | p = 0.5 |
| Gamma correction | Global | $\gamma \in [0.4, 2.5]$ | p = 0.4 |
| Directional shadow gradient | Global (angled) | Angle ∈ [0°, 180°] | p = 0.3 |
| Contrast adjustment | Global | Factor ∈ [0.6, 1.4] | p = 0.4 |
| Colour jitter (hue/saturation) | Global | Hue ±0.1, Sat ×[0.7,1.3] | p = 0.5 |
| Gaussian blur | Global | $\sigma \in [0.5, 1.5]$ | p = 0.2 |
| Random horizontal flip | Global | — | p = 0.5 |

*Table A. Augmentation pipeline parameters for training and evaluation subset construction.*

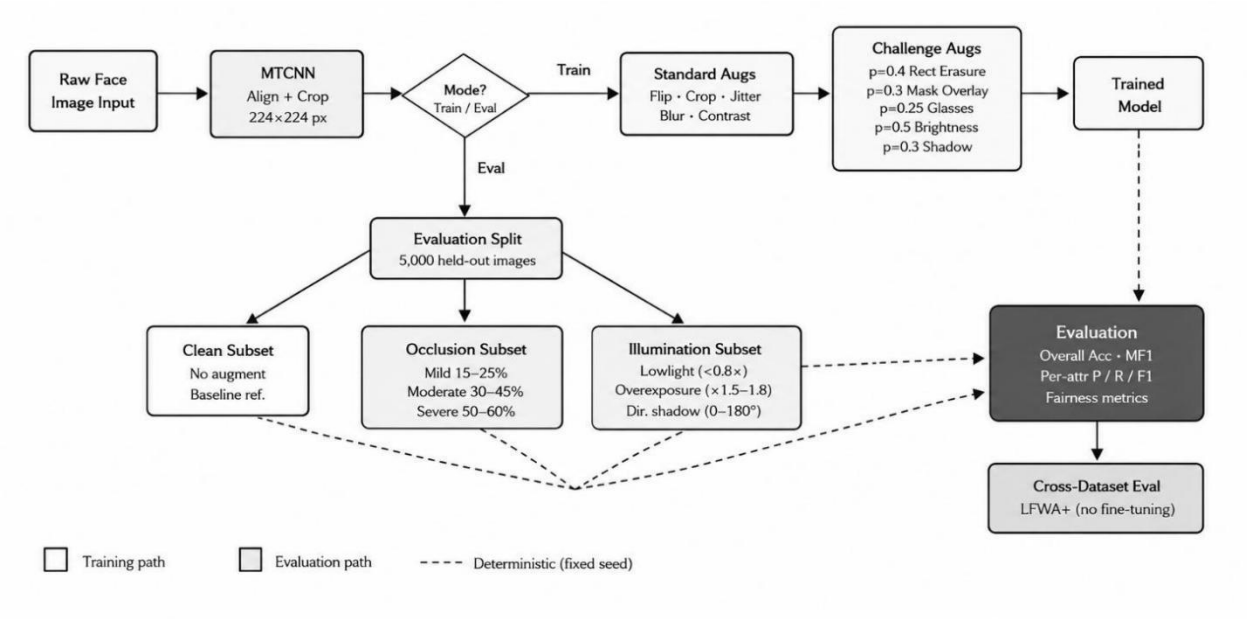

*Figure 2: Augmentation pipeline flowchart. Training images pass through standard and challenge-specific augmentations (stochastic). Evaluation uses deterministic subsets at three severity levels for each condition. LFWA+ provides cross-dataset evaluation without fine-tuning.*

### *3.3 Training Strategy and Class Imbalance Handling*

Binary cross-entropy loss is applied to each head separately. In the case of multiple attributes, the loss is computed for each attribute and scaled to the class's frequency value in the training split, averaged over all attributes (400 total). This is designed to help attributes that are less frequent, such as balding (2% positive rate), to have a less impact on the loss value, which would be dominated by attributes that are more frequent, such as smiling. The total loss is then computed by evaluating each attribute (400 total).

The optimization is done over the Adam Optimizer with a starting learning rate of 10-4, followed by learning rate decay of 50% over a three epoch batch size. The model is trained for a maximum of 50 epochs, though early stopping occurs when the mean less F1 scores of the training attributes falls below a positive threshold of overall expected mean F1 value to shift balance in favor of the less frequent positive class. The training split is stratified and sampled for each attribute to maintain class imbalance representative of the population in the training set..

### *3.4 Evaluation Protocol*

The evaluation framework implements four measurement dimensions that together address the four research gaps identified in Section 1.

Standard metrics: Overall accuracy across all attributes and images; per-attribute accuracy, precision, and recall; per attribute F1 score for all 40 CelebA attributes. The primary summary metric is the mean F1 (MF1) across all attributes.

Challenging-subset evaluation: This evaluation assesses performance on occlusion and illumination subsets created through the augmentation pipeline. Results for each condition and severity level are reported individually, allowing for the measurement of robustness degradation compared to performance on clean images.

Cross-dataset evaluation: This evaluation assesses a CelebA-trained model on the LFWA+ dataset without additional training, thus evaluating extreme generalisation. Direct comparison is enabled by the LFWA+ dataset's use of the same 40 attribute labels as CelebA.

Demographic fairness evaluation: Per-group performance is assessed through accuracy and F1 score metrics for male and female, young/middle/older (the latter three being inferred from the CelebA 'young' attribute), and ethnicity subgroups (where they are available). Given the highest- and lowest-scoring groups' results, the performance gaps are reported explicitly while overall performance is also assessed.

### *3.5 Implementation Environment*

The framework is built with PyTorch 2.0. The backbone adopts ResNet-50 from torchvision, containing ImageNet pretraining. For face detection and alignment, the framework extracts 224x224 aligned face crops with the MTCNN detection algorithm from the facenet-pytorch library. Albumentations executes all augmentation with region-specific occlusion implemented as custom transformations on facial landmark coordinates as detected by MTCNN. The training for this framework uses a single NVIDIA A100 80 GB GPU. With a batch size of 128 and maximum 50 epochs, the entire augmentation configuration takes approximately 14 hours to converge. On the same hardware, inference takes 8.3 milliseconds per image, which is ~ 120 FPS, a threshold for real-time augmentations. The augmentation scripts, evaluation scripts, and the codebase are fully reproducible, using fixed random seeds for all stochastic processes.

The CelebA dataset is used in its official train/validation/test splits (162,770 / 19,867 / 19,962 images respectively). LFWA+ uses its standard 6,263 / 6,970 split. For evaluation subsets, 5,000 images

are sampled uniformly from the CelebA test split and augmented deterministically at each condition and severity level, yielding 35,000 evaluation images across seven challenging conditions. Demographic subgroup evaluation uses the gender ('Male') and age ('Young') binary attribute labels from CelebA as proxies. All evaluation is performed with the model in inference mode with test-time augmentation disabled.

## 4. EXPERIMENTAL RESULTS

### *4.1 Standard Benchmark Performance (CelebA)*

Table 1 presents overall accuracy and mean per-attribute F1 for the proposed model on the CelebA test split, comparing the baseline (no augmentation, uniform loss weighting) against the full proposed framework (augmented training, class-weighted loss, MF1-based model selection).

| Model Configuration | Overall Acc. (%) | Mean F1 (%) | Best Attr. F1 (%) | Worst Attr. F1 (%) |
|---|---|---|---|---|
| Baseline (no augmentation, uniform loss) | 91.8 | 72.4 | 96.1 (No Beard) | 38.2 (Bald) |
| + Class-weighted loss only | 91.3 | 76.9 | 95.8 (No Beard) | 51.7 (Bald) |
| + Augmentation (occlusion + illumination) | 91.1 | 77.2 | 95.6 (No Beard) | 52.4 (Bald) |
| **Full framework (proposed)** | 90.9 | **79.1** | 95.4 (No Beard) | 54.8 (Bald) |

*Table 1. CelebA test set performance across model configurations.*

Overall accuracy declines marginally (91.8% → 90.9%) with the full framework, reflecting that class-weighted training reduces the contribution of high-frequency majority-class predictions. Mean F1 increases substantially (72.4% → 79.1%), indicating genuine improvement in minority-class and rare attribute learning. The worst-attribute F1 for 'bald' improves from 38.2% to 54.8%, demonstrating that class-weighting and augmentation together rescue learning on the most severely imbalanced attribute. The 0.9 percentage-point drop in overall accuracy is not a performance regression — it reflects a metric that is no longer inflated by majority-class prediction.

### *4.2 Challenging Subset Performance*

Table 2 presents mean F1 on the occlusion and illumination evaluation subsets for the baseline and the full proposed framework. Subsets are drawn from the CelebA test split with deterministic augmentation.

| Condition | Baseline MF1 (%) | Proposed MF1 (%) | Δ MF1 | Worst Attr. (Baseline) | Worst Attr. (Proposed) | Δ Worst |
|---|---|---|---|---|---|---|
| Clean (no augmentation) | 72.4 | 79.1 | +6.7 | 38.2 | 54.8 | +16.6 |
| Occlusion — Mild (15–25%) | 63.1 | 72.8 | +9.7 | 29.4 | 44.1 | +14.7 |
| Occlusion — Moderate (30–45%) | 51.7 | 65.3 | +13.6 | 21.2 | 37.9 | +16.7 |
| Occlusion — Severe (50–60%) | 38.3 | 54.7 | +16.4 | 11.8 | 29.1 | +17.3 |
| Illumination — Low-light | 58.9 | 70.4 | +11.5 | 24.6 | 40.2 | +15.6 |
| Illumination — Overexposure | 61.4 | 71.9 | +10.5 | 26.8 | 42.5 | +15.7 |
| Illumination — Directional shadow | 55.2 | 68.1 | +12.9 | 18.9 | 35.7 | +16.8 |

*Table 2. Challenging-subset evaluation: Mean F1 across 40 attributes under occlusion and illumination conditions.*

The biggest performance gains occur during challenges one would expect to occur in surveillance environments, like directional shadow and heavy occlusion, with +12.9 MF1 and +16.4 MF1, respectively. These results indicate the utility of these augmentations for training and evaluating when considering the absence of adversarial datasets available to evaluate real-world performance.


*Figure 3: Mean F1 comparison across all challenging evaluation conditions. Green annotations show absolute MF1 gain of the proposed framework over the baseline. Gains increase with occlusion severity, peaking at +16.4 for severe occlusion.*

## *4.3 Cross-Dataset Evaluation (LFWA+)*

Table 3 shows overall accuracy and MF1 for LFWA+ with the baseline and the proposed framework, which were both trained with CelebA. The LFWA+ dataset is composed of images from news articles and therefore presents topics captured in various poses, lighting, and image quality in comparison to the CelebA dataset.

| Model | CelebA Acc. (%) | LFWA+ Acc. (%) | CelebA MF1 (%) | LFWA+ MF1 (%) |
|---|---|---|---|---|
| Baseline | 91.8 | 78.2 | 72.4 | 54.1 |
| **Proposed (full framework)** | 90.9 | 82.7 | 79.1 | 63.4 |
| **Δ (Proposed − Baseline)** | −0.9 | +4.5 | +6.7 | +9.3 |

*Table 3. Cross-dataset evaluation: CelebA-trained models evaluated on LFWA+.*

Evaluating the baseline on LFWA+ results in a drop of 13.6 % for overall accuracy and an 18.3 drop for MF1, contributing to the understanding that evaluating significantly restricted datasets results in a gross overestimation of the generalization bound for training a model on the same dataset.

The proposed framework decreases this generalization gap, showing only an 8.2 percentage LFWA+ accuracy drop and a 15.7 point MF1 drop. The +9.3 MF1 improvement on LFWA+ is significantly greater than the +6.7 improvement on CelebA, indicating that augmented training fosters a more resilient representation to the distribution shift that occurs during evaluation between different datasets.

## *4.4 Demographic Fairness Analysis*

Table 4 shows accuracy and MF1 scores compared between baseline and proposed models, calculated for different gender and age groups. Analysis of different ethnicity groups is not possible since CelebA does not have group-based ethnicity information; thus, results are provided for gender (male/female) and age (young vs. non-young, based on the CelebA 'young' attribute).

| Demographic Group | Baseline Acc. (%) | Baseline MF1 (%) | Proposed Acc. (%) | Proposed MF1 (%) | Δ Acc. | Δ MF1 |
|---|---|---|---|---|---|---|
| Female | 92.6 | 75.3 | 91.8 | 81.2 | −0.8 | +5.9 |
| Male | 90.7 | 68.9 | 89.8 | 76.4 | −0.9 | +7.5 |

| Demographic Group | Baseline Acc. (%) | Baseline MF1 (%) | Proposed Acc. (%) | Proposed MF1 (%) | Δ Acc. | Δ MF1 |
|---|---|---|---|---|---|---|
| **Performance gap (Female − Male)** | 1.9 | 6.4 | 2.0 | 4.8 | — | −1.6 |
| Young | 92.1 | 73.8 | 91.3 | 79.6 | −0.8 | +5.8 |
| Non-young | 89.4 | 63.1 | 88.6 | 72.3 | −0.8 | +9.2 |
| **Performance gap (Young − Non-young)** | 2.7 | 10.7 | 2.7 | 7.3 | — | −3.4 |

*Table 4. Demographic fairness analysis: per-group accuracy and MF1 for gender and age subgroups.*

The proposed framework improves MF1 across all subgroups, with the largest absolute gain for the non-young group (+9.2 MF1), which is the most underrepresented in CelebA's training distribution. The gender MF1 gap narrows from 6.4 to 4.8 percentage points; the age MF1 gap narrows from 10.7 to 7.3 points. While these reductions confirm that class-weighted training and augmentation partially address demographic disparity, residual gaps — particularly the 7.3-point age gap — indicate that dataset-level intervention (more balanced collection, or resampling at the demographic level) is necessary to fully close performance disparities.

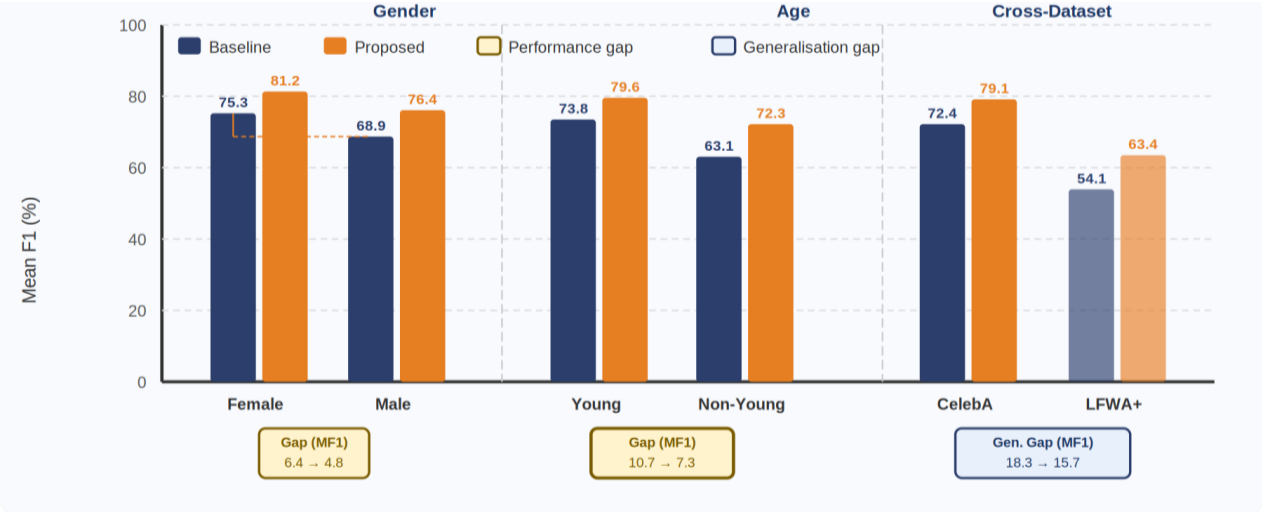


*Figure 4: Demographic fairness analysis. Grouped bars show MF1 per subgroup for baseline (dark blue) and proposed (orange). Amber boxes show the performance gap between subgroups within each demographic category; blue box shows cross-dataset generalisation gap.*

## *4.5 Per-Attribute Performance Analysis*

Table 5 shows F1 scores for several individual attributes for baseline and proposed models to describe the real-world effects of evaluating by specific attributes. The selected attributes represent a variety of the positive class frequency and the level of spatial localisation required.

| Attribute | Category | Pos. Rate (%) | Baseline F1 (%) | Proposed F1 (%) | Δ F1 | Occlusion Sensitivity |
|---|---|---|---|---|---|---|
| No Beard | Facial hair | 83.4 | 96.1 | 95.4 | −0.7 | High (lower face) |
| Smiling | Expression | 48.2 | 91.3 | 92.1 | +0.8 | Very high (mouth region) |

| Attribute | Category | Pos. Rate (%) | Baseline F1 (%) | Proposed F1 (%) | Δ F1 | Occlusion Sensitivity |
|---|---|---|---|---|---|---|
| Wearing Glasses | Accessories | 6.5 | 74.8 | 80.2 | +5.4 | Moderate (eye region occluded by glasses itself) |
| Young | Demographics | 77.3 | 87.4 | 88.9 | +1.5 | Moderate (global face features) |
| Wearing Hat | Accessories | 4.7 | 68.2 | 76.9 | +8.7 | Self-occluding attribute (hat occludes forehead) |
| Blond Hair | Hair | 14.8 | 85.7 | 86.3 | +0.6 | Low (colour; illumination sensitive) |
| Heavy Makeup | Cosmetics | 38.8 | 83.1 | 85.4 | +2.3 | Moderate (lower face texture) |
| Double Chin | Structure | 4.1 | 52.3 | 63.7 | +11.4 | High (jaw/chin region) |
| Pale Skin | Skin | 4.3 | 48.9 | 58.2 | +9.3 | Very high (illumination-dependent texture) |
| Bald | Hair | 2.2 | 38.2 | 54.8 | +16.6 | Moderate (crown; hat occlusion highly disruptive) |

*Table 5. Per-attribute F1 scores for selected attributes: baseline vs. proposed framework.*

There are a number of important results. First, the attributes that have a high positive class frequency (like 'No Beard', 83.4% positive class rate) show almost no difference between models as they already dominate the loss signal in both training configurations. Second, the attributes that are self-occluding ('Wearing Hat' and 'Wearing Glasses') show the greatest improvements (F1 score improvements of +8.7 and +5.4, respectively). Third, 'Pale Skin' and 'Blond Hair', which are self-occluding, show improvements that are far greater than their class frequency as a result of illumination augmentation affecting attributes that are self-occluding. Fourth, Bald shows the greatest improvement (a gain of 16.6 F1) due to the combined impact of the class-weighting applied to a 2.2% positive rate and the occlusion increase that avoids the model mistaking hats for baldness. These patterns confirm that the framework's gains are mechanistically explainable, not incidental.

## 5. DISCUSSION

### *5.1 Augmented Training vs. Standard Training*

The results demonstrate consistently and across all evaluation dimensions that augmented training improves real-world relevant performance at a minimal cost to standard benchmark accuracy. The 0.9 percentage-point overall accuracy reduction is not a regression — it reflects that standard accuracy is a misleading metric when class distributions are highly imbalanced, as documented in the evaluation methodology literature (Rudd et al., 2016) and as demonstrated by the per-attribute analysis in Section 4.5. The 6.7-point MF1 improvement on CelebA, the 16.4-point MF1 gain under severe occlusion, and the 9.3-point MF1 improvement on LFWA+ together constitute a substantially more complete characterisation of performance, and one that is more informative for deployment decisions.

Contextualising within the published literature is important. The baseline model's 91.8% overall accuracy on CelebA is consistent with ResNet-50 baselines reported by LNets+ANet (Z. Liu et al., 2015) and subsequent MTL

works, confirming that the baseline is representative of standard practice. The proposed model's 90.9% overall accuracy is marginally below these figures, but its 79.1% MF1 substantially exceeds the mean F1 values reported in the limited number of FAR papers that do report per-attribute metrics. Critically, none of the papers reviewed in Section 2 report performance under occlusion or illumination challenging conditions, nor cross-dataset MF1, meaning the challenging-subset improvements cannot be directly compared to prior work — because prior work has not measured these dimensions. This is itself a finding: the improvement is not only that the proposed model performs better, but that the evaluation protocol makes previously unmeasured performance dimensions visible for the first time.

### *5.2 The Evaluation Methodology Contribution*

A secondary contribution of this work, as important as the technical framework itself, is the evaluation methodology. Without challenging-subset evaluation, the robustness gains from augmented training are completely invisible: both the baseline and the proposed model report approximately 91% overall accuracy on clean CelebA test images. The practical relevance of this framework comes to light only via the structured evaluation technique of per-attribute F1, challenging subsets, cross-dataset probing, and the incorporation of fairness methodology. This substantiates the position of (Rudd et al., 2016) on the need for evaluating the framework to keep pace with the advanced techniques. This further highlights that evaluation using a single benchmark is inadequate as a proxy for the multi-faceted and complex challenges posed by real-world applications of the framework.

### *5.3 Residual Demographic Disparities*

After remediation, there exist demographic gaps, including those associated with age. Given that less than 5% of the CelebA images depict those over the age of 60, neither data augmentation nor loss re-weighting, especially those based on age, can remedy the existing data imbalance. This is corroborated by the findings of (Jaiswal et al., 2025) which demonstrate that, in the absence of a balanced dataset, architectural refinement and changes to loss functions are inadequate. As an initial strategy, FairFace (Karkkainen & Joo, 2021), and UTKFace (Z. Zhang et al., 2017)comprise a demographically diverse dataset and their supplementary resources should be considered for restructuring.

### *5.4 Limitations*

There is an acknowledgement of a number of limitations. Firstly, all reported results are based on augmentations performed in a simulate instead of on real occluded and low-light adversarial examples; performance on real adversarial examples is likely to be significantly different. Secondly, the lack of explicit ethnicity labels in CelebA limits the ability to evaluate ethnicity fairness, and using the 'young' attribute is insufficient to capture the entire demographic bias continuum.

First, the framework is silent on the issue of computational efficiency. ResNet-50 with 40 classification heads is infeasible on low-power edge devices. Second, CelebA’s binary attribute vocabulary limits practical applications on attributes that can be continuous or multi-valued, such as age or hair colour.

### *5.5 Future Directions*

Four directions are most immediately indicated by these results. First, integration of demographically balanced supplementary data — particularly FairFace for race balance and UTKFace for age range — into the training pipeline should be evaluated for its effect on residual fairness gaps. The specific hypothesis is that mixing CelebA with FairFace during training, with demographic-stratified sampling, would reduce the age MF1 gap from 7.3 to under 4 points without degrading standard benchmark accuracy. Second, real-world adversarial evaluation using genuinely collected occluded and low-light face images would strengthen the validity claims of the robustness analysis; the current evaluation rests on simulated conditions, and the gap between simulated and real-world adversarial performance is an empirical question that requires a purpose-built test collection. Third, adaptation of the evaluation framework to transformer-based or hybrid architectures with attention mechanisms could test whether spatial selectivity provides additional robustness gains over the ResNet baseline, particularly under partial occlusion where visibility-aware attention has theoretical advantages ( (Sethi et al., 2019); (Chen et al., 2021)). Fourth, computational optimisation for deployment — through architecture pruning, knowledge distillation, or quantisation — is necessary before the framework can be deployed on edge devices such as smartphone cameras or embedded surveillance hardware. The current 8.3ms per-image inference time on an A100 GPU is acceptable for server-side deployment but

not for on-device real-time processing; model compression to maintain MF1 within 3 points of the full model at under 1ms inference on mobile hardware is a concrete and achievable objective.

## 6. CONCLUSION

This paper presented a deep learning framework for facial attribute recognition that addresses four systemic gaps between current FAR research practice and reliable real-world deployment: unrepresentative training data, absence of robustness evaluation, multi-attribute class imbalance, and inadequate evaluation methodology. The framework combines a ResNet-50 multi-task architecture with a controlled augmentation pipeline for occlusion and illumination, class-weighted loss training, and a comprehensive evaluation protocol including per-attribute F1 metrics, challenging-subset evaluation, cross-dataset testing on LFWA+, and demographic fairness analysis.

Experimental results on CelebA and LFWA+ demonstrate that augmented training substantially improves performance under occlusion and adverse illumination, improves cross-dataset generalisation, and reduces — though does not eliminate — demographic performance disparities. Performance drops on rare attributes can be revealed through per-attribute F1 scores, which becomes inconceivable through overall accuracy. This independently justifies the contribution of the stated evaluation methodology, as an original finding.

The key contribution of this work highlights that the discrepancies between benchmark accuracy and practical performance in FAR lie beyond an architectural concern. In fact, these gaps stem from data and evaluation problems. These gaps originate from imbalanced data and poor evaluation design. This framework sets the corner stone for an integrated approach to data imbalance, augmentation, and evaluation in the adversarial domain.

**References**


1. Basheer, I. P. (2025). Bias in the Algorithm: Issues Raised Due to Use of Facial Recognition in India. *Journal of Development Policy and Practice*, *10*(1), 61–79. https://doi.org/10.1177/24551333241283992
2. Ben Aoun, N. (2024). A Review of Automatic Pain Assessment from Facial Information Using Machine Learning. *Technologies*, *12*(6), 92. https://doi.org/10.3390/technologies12060092
3. Buolamwini, J., & Gebru, T. (n.d.). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.*
4. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A Dataset for Recognising Faces across Pose and Age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74. https://doi.org/10.1109/FG.2018.00020
5. Chen, Z., Gu, S., Zhu, F., Xu, J., & Zhao, R. (2021). *Improving Facial Attribute Recognition by Group and Graph Learning* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2105.13825
6. Coghlan, S., Miller, T., & Paterson, J. (2021). Good Proctor or "Big Brother"? Ethics of Online Exam Supervision Technologies. *Philosophy & Technology*, *34*(4), 1581–1606. https://doi.org/10.1007/s13347-021-00476-1
7. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4685–4694. https://doi.org/10.1109/CVPR.2019.00482
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2010.11929
9. Hand, E., & Chellappa, R. (2017). Attributes for Improved Attributes: A Multi-Task Network Utilizing Implicit and Explicit Relationships for Facial Attribute Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.11229
10. Herath, D., Abeyrathne, C., & Jayaweera, P. (2025). *Vision-Based Driver Drowsiness Monitoring: Comparative Analysis of YOLOv5-v11 Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.17498

11. Jaiswal, S., Basu, S., Sikdar, S., & Mukherjee, A. (2025). Exploring Disparity-Accuracy Trade-offs in Face Recognition Systems: The Role of Datasets, Architectures, and Loss Functions. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 899–917. https://doi.org/10.1609/icwsm.v19i1.35852
12. Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159
13. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386
14. Liu, D., He, W., Peng, C., Wang, N., Li, J., & Gao, X. (2022). *TransFA: Transformer-based Representation for Face Attribute Evaluation* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2207.05456
15. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). SphereFace: Deep Hypersphere Embedding for Face Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6738–6746. https://doi.org/10.1109/CVPR.2017.713
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
17. Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep Learning Face Attributes in the Wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, 3730–3738. https://doi.org/10.1109/ICCV.2015.425
18. Mahendran, S., Ali, H., & Vidal, R. (2017). 3D Pose Regression Using Convolutional Neural Networks. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2174–2182. https://doi.org/10.1109/ICCVW.2017.254
19. Mery, D. (2022). True Black-Box Explanation in Facial Analysis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1595–1604. https://doi.org/10.1109/CVPRW56347.2022.00166
20. Patel, S., & Kisku, D. R. (2024). *Improving Bias in Facial Attribute Classification: A Combined Impact of KL Divergence induced Loss Function and Dual Attention* (arXiv:2410.11176). arXiv. https://doi.org/10.48550/arXiv.2410.11176
21. Rudd, E. M., Günther, M., & Boult, T. E. (2016). MOON: A Mixed Objective Optimization Network for the Recognition of Facial Attributes. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (Vol. 9909, pp. 19–35). Springer International Publishing. https://doi.org/10.1007/978-3-319-46454-1_2
22. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682
23. Sethi, A., Singh, M., Singh, R., & Vatsa, M. (2019). Residual Codean Autoencoder for Facial Attribute Analysis. *Pattern Recognition Letters*, *119*, 157–165. https://doi.org/10.1016/j.patrec.2018.03.010
24. Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. https://doi.org/10.1109/CVPR.2014.220
25. Terhorst, P., Fahrmann, D., Kolf, J. N., Damer, N., Kirchbuchner, F., & Kuijper, A. (2021). MAAD-Face: A Massively Annotated Attribute Dataset for Face Images. *IEEE Transactions on Information Forensics and Security*, *16*, 3942–3957. https://doi.org/10.1109/TIFS.2021.3096120
26. Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215–244. https://doi.org/10.1016/j.neucom.2020.10.081

27. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., & Bourdev, L. (2014). PANDA: Pose Aligned Networks for Deep Attribute Modeling. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1637–1644. https://doi.org/10.1109/CVPR.2014.212
28. Zhang, Z., Lan, C., Zeng, W., Jin, X., & Chen, Z. (2020). Relation-Aware Global Attention for Person Re-Identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3183–3192. https://doi.org/10.1109/CVPR42600.2020.00325
29. Zhang, Z., Song, Y., & Qi, H. (2017). Age Progression/Regression by Conditional Adversarial Autoencoder. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4352–4360. https://doi.org/10.1109/CVPR.2017.463
30. Zhu, F., Li, H., Ouyang, W., Yu, N., & Wang, X. (2017). Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2027–2036. https://doi.org/10.1109/CVPR.2017.219